

Detection of pleiotropic repeat expansions and their oligogenic contribution to amyotrophic lateral sclerosis



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

Volume 1

2023

Jennifer C. Hengeveld

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: _____

A handwritten signature in black ink, appearing to read 'Hengereid', is written over a horizontal line.

Date: 26/05/2023

List of abbreviations and gene symbols

3GS	Third-generation sequencing
AAV	Adeno-associated viral vectors
<i>AFF2</i>	FMR2- AF4/FMR2 family member 2
<i>AFF3</i>	AF4/FMR2 family member 3
ALS	Amyotrophic lateral sclerosis
ALS2	Alsin rho guanine nucleotide exchange factor ALS2
ANG	Angiogenin
<i>AR</i>	Androgen receptor
<i>ARX</i>	Aristaless related homeobox
ASO	Antisense oligonucleotide
<i>ATN1</i>	Atrophin 1
ATXN	Ataxin
BER	Base excision repair
Betaine	N,N,N-trimethylglycine
bp	Base pair
bps	Base pairs
<i>C11orf80</i>	Chromosome 11 open reading frame 80
<i>C9orf72</i>	Chromosome 9 open reading frame 72
CACNA1A	Calcium voltage-gated channel subunit alpha1 A
CCS	Circular consensus sequencing
CDS	Protein-coding exons
CE	Capillary electrophoresis
CI	Confidence intervals
<i>CNBP</i>	CCHC-type zinc finger nucleic acid binding protein
CNS	Central nervous system
<i>CSTB</i>	Cystatin B
<i>DIP2B</i>	Disco interacting protein 2 homolog B
DM1	Myotonic dystrophy type 1
DM2	Myotonic dystrophy type 2

<i>DMPK</i>	DM1 protein kinase
DMSO	Dimethyl sulfoxide
DPRs	Dipeptide repeat proteins
DRPLA	Dentatorubral-pallidoluysian atrophy
EPM1	Unverricht-Lundborg disease
F	Forward
FA	Friedreich ataxia
FALS	Familial ALS
FECD	Fuchs endothelial corneal dystrophy
FLA	Fragment length analysis
<i>FMRI</i>	FMRP translational regulator 1
FMRP	Fragile X mental retardation protein
<i>FRA10AC1</i>	FRA10A associated CGG repeat 1
<i>FRA11A</i>	Fragile Site A region of chromosome 11
<i>FRA12A</i>	Intellectual disability, associated with fragile site FRA12A
<i>FRA2A</i>	Intellectual disability, associated with fragile site FRA2A
FRA7A	Autism spectrum disorder, associated with fragile site FRA7A
FRAXA	Fragile X syndrome
FRAXE	Fragile XE syndrome
FRDA	Friedreich's ataxia
FTD	Frontotemporal dementia
<i>FXN</i>	Frataxin
FXPOI	Fragile X-associated primary ovarian insufficiency
FXS	Fragile X syndrome
FXTAS	Fragile X-associated tremor/ataxia syndrome
gDNA	Genomic DNA
<i>GIPC1</i>	GIPC PDZ domain containing family member 1
GOF	Gain-of-function
GWAS	Comprehensive genome-wide association studies
HD	Huntington's disease
HDL2	Huntington's disease-like 2
hg	Human genome

HRE	Hexanucleotide repeat expansion
<i>HTT</i>	Huntingtin
IGF	Insulin growth factor
INVDELs	Inversions flanked by indels
INVDUPs	Inverted tandem duplications
iPSCs	Induced pluripotent stem cells
<i>JPH3</i>	Junctophilin 3
LGMD2B	Limb-girdle muscular dystrophy type 2B
LMN	Lower motor neurons
LOF	Loss-of-function
<i>LRP12</i>	LDL receptor related protein 12)
MAPT	Microtubule-associated protein tau
MBNL	Muscleblind-like protein
miRNAs	Mitochondrial RNAs
mM	Millimolar
MMR	Mismatch repair
NC	Negative Control
NER	Nucleotide excision repair
ng	Nanograms
NGS	Next generation sequencing
NIID	Neuronal intranuclear inclusion disease
<i>NIPA1</i>	NIPA magnesium transporter 1
<i>NOP56</i>	Ribonucleoprotein
<i>NOTCH2NLC</i>	Notch 2 N-terminal like C
nts	Nucleotides
<i>NUTM2B-AS1</i>	Nutm2b antisense RNA 1
ONT	Oxford nanopore technologies
OPDM1	Oculopharyngodistal myopathy 1
OPDM2	Oculopharyngodistal myopathy 2
OPMD	Oculopharyngeal muscular dystrophy
OPML	Oculopharyngeal myopathy with leukoencephalopathy
OR	Odds ratio

<i>PABPN1</i>	Poly(A) binding protein nuclear 1
PBMCs	Peripheral blood mononuclear cells
<i>PGRN</i>	Progranulin
Pol II	Polymerase II
<i>PPP2R2B</i>	Protein phosphatase 2 regulatory subunit Bbeta
PSEN-2	Presenilin 2
RAN	Repeat Associated non-ATG
RBP	RNA-binding protein
REs	Repeat expansions
RISC	RNA-induced silencing complex
RNAi	RNA interference
rpPCR	Repeat-primed-PCR
SaCas9	Staphylococcus aureus-derived Cas9
SALS	Sporadic ALS
<i>SAMD12</i>	Sterile Alpha Motif Domain Containing 12
SBMA	Spinal and bulbar muscular atrophy
SCA1	Spinocerebellar ataxia 1
SCA10	Spinocerebellar ataxia type 10
SCA12	Spinocerebellar ataxia type 12
SCA17	Spinocerebellar ataxia type 17
SCA2	Spinocerebellar ataxia 2
SCA3	Spinocerebellar ataxia type 3
SCA31	Spinocerebellar ataxia type 31
SCA36	Spinocerebellar ataxia type 36
SCA6	Spinocerebellar ataxia type 6
SCA7	Spinocerebellar ataxia type 7
SCA8	Spinocerebellar ataxia type 8
scAAV9	Self-complementary adeno-associated virus serotype 9
<i>SEXT</i>	Senataxin
sgRNA	Single-guide RNA
shRNAs	Short hairpin RNAs
siRNAs	Short interfering RNAs

SMRT	Single molecule real-time
SNP	Single nucleotide polymorphism
<i>SOD1</i>	Superoxide dismutase type 1
SPG6	Hereditary Spastic Paraplegia type 6
<i>SQSTM1</i>	Sequestosome-1
STR	Short tandem repeat
SV	Structural variant
TADs	Topologically Associating Domains
<i>TARDBP</i>	TAR DNA binding protein
<i>TBP</i>	TATA-box binding protein
<i>TCF4</i>	Transcription factor 4
TDP-43	TAR DNA-binding protein 43
TE	Tris-EDTA
<i>TK2-BEAN1</i>	Thymidine kinase 2- brain expressed associated with NEDD4 1
T _m	melting temperature
TNR	Trinucleotide repeats
tracrRNA	Transactivation crRNA
<i>UBQLN2</i>	Ubiquitin-like protein 2
UMN	Upper motor neurons
UTR	Untranslated region
UV	Ultraviolet
<i>VAPB</i>	Vesicle-associated membrane protein
XLID	X-linked intellectual disability
xtrRNA	Expanded tandem repeat-containing RNA
<i>ZNF713</i>	Zinc finger protein 713

Acknowledgements

I would like to take this opportunity to express my heartfelt appreciation and gratitude to the following individuals and groups who have made significant contributions to the completion of my PhD thesis:

First and foremost, I extend my sincere thanks to the ALS and FTD patients, their caregivers, relatives, and the generous healthy volunteers. Their selfless participation and willingness to dedicate their time and energy have been crucial in advancing our understanding of these complex conditions. Without their invaluable contributions, this research would not have been possible.

I extend my deepest appreciation to my principal investigator, Dr. Russell, for his dedication, patience, and guidance throughout my research. His expertise, constructive feedback, and unwavering belief in my capabilities have shaped my academic growth and instilled in me a passion for scientific inquiry. I am honoured to have had the opportunity to work under his mentorship.

I would like to extend special thanks to the members of my research lab, particularly Mark, Ross, and Ciaran. Their collaboration, camaraderie, and unwavering support have been instrumental in my progress. I am grateful for the memorable walks we took around TCD grounds during stressful moments and for the shared breaks that provided solace during challenging times.

I would also like to express my gratitude to Bradly and Cassidy's lab for their collaboration and support. The opportunity to exchange ideas, share resources, and learn from their expertise has enriched my research experience.

This thesis is dedicated to my mother, who has always believed in my potential and encouraged me to pursue a career in the scientific field. Her unwavering support and endless sacrifices have been a driving force behind my accomplishments. This achievement is a testament to her love and unwavering belief in my abilities.

I am immensely grateful to my family and close friends, specially to my father, brother, sister, auntie Ana, Adriana, Sergio, Dore, Vane and Jon for their continuous support throughout the past five years. Their encouragement, understanding, and support have been a source of strength during the ups and downs of my academic journey.

Last but not least, I owe a special debt of gratitude to my partner, Saul, whose unwavering support and belief in my abilities have been instrumental in my perseverance during challenging times, particularly amidst the difficulties imposed by the COVID-19 pandemic. Saul's constant care, comforting presence, and willingness to go above and beyond, from making laugh to providing a listening ear for my complaints and troubles, have sustained me through the highs and lows of this journey. I am also grateful for Saul's invaluable assistance in creating and developing my sophisticated master Excel sheet for data analysis and interpretation.

To all those who have played a role, no matter how big or small, in shaping my academic and personal growth during this doctoral pursuit, I offer my heartfelt thanks. Your contributions have been invaluable, and I am truly grateful for your unwavering support and encouragement.

Table of contents

LIST OF ABBREVIATIONS AND GENE SYMBOLS.....	V
ACKNOWLEDGEMENTS.....	X
TABLE OF CONTENTS	XII
TABLE OF FIGURES.....	XVIII
TABLE OF TABLES.....	XXIV
SUMMARY.....	XXV
PUBLICATIONS FROM THIS THESIS.....	XXVII
1 CHAPTER 1.....	1
INTRODUCTION	1
AMYOTROPHIC LATERAL SCLEROSIS.....	1
CLASSIFICATION OF ALS AND CONTRIBUTING FACTORS	2
REPETITIVE SEQUENCES AND NEURODEGENERATIVE DISEASES.....	4
RE LENGTH AND SEVERITY OF THE DISEASE	9
INHERITANCE MECHANISMS OF RES	12
PATHOMECHANISMS IN REPEAT EXPANSION DISEASES	13
CONVERGENT DISEASE MECHANISMS IN ALS	27
SCOPE AND STRUCTURE OF THESIS	28
2 CHAPTER 2.....	30
DEVELOPMENT OF AN EFFICIENT, COST-EFFECTIVE, HIGHLY MULTIPLEXED GENOTYPING PANEL FOR NEURODEGENERATIVE REPEAT EXPANSIONS.....	30
2.1 INTRODUCTION	30
<i>Research Aims.....</i>	32
2.2 METHODS	33
<i>Genotyping Assays</i>	33
Polymerase Chain Reaction (PCR)	33
Optimisation of PCR	35
Oligonucleotide design	35
PCR specificity.....	37
Magnesium, dNTPs and DNA polymerase.....	37
Temperatures	38
Cycle number and length.....	39

PCR additives.....	40
Multiplex Polymerase Chain Reaction	40
Repeat Primed PCR (rpPCR)	41
Electrophoresis.....	42
Agarose gel electrophoresis	42
Capillary electrophoresis- Fragment Length Analysis (FLA).....	43
Sequencing.....	48
Study design.....	48
2.3 RESULTS	50
<i>PCR optimization</i>	50
Multiplex 1: HTT, JPH3, ATXN7 and ATXN10 genes.....	50
Optimization of HTT and JPH3 genes	50
Optimization of ATXN7 and ATXN10 genes	54
Multiplex PCR optimization.....	56
Multiplex 2: NOP56, PPP2R2B, TBP and TK2-BEAN1 genes.....	58
Optimization of NOP56, PPP2R2B and TBP genes	58
Optimization of TK2-BEAN1 gene	60
Multiplex PCR optimization.....	62
<i>Sequencing results</i>	65
ATXN7.....	65
ATXN10.....	67
HTT	68
JPH3.....	69
NOP56.....	71
PPP2R2B	73
TBP	74
TK2-BEAN1	76
<i>FLA results</i>	78
<i>Repeat count calibration</i>	81
<i>rpPCR Optimization</i>	89
2.4 DISCUSSION.....	93
2.5 SUMMARY.....	97
3 CHAPTER 3	98
POPULATION GENETIC ANALYSIS OF THE GENETIC PLEIOTROPY OF NEURODEGENERATIVE REPEAT EXPANSIONS IN ALS AND FTD	98
3.1 INTRODUCTION.....	98
<i>Heterogeneity of ALS sub-phenotypes</i>	99
<i>Common features between RE-associated diseases</i>	100
<i>Research aims</i>	101

3.2	METHODS	106
	<i>Sampling</i>	106
	<i>PCR Genotyping</i>	106
	<i>Statistical analysis and plotting</i>	107
3.3	RESULTS	110
	<i>Genotyping results (ALS)</i>	110
	Short Allele	110
	Longer repeat allele vs. ALS risk	110
	Previously reported ALS risk genes.....	110
	<i>ATXN1</i>	112
	<i>ATXN2</i>	114
	<i>C9orf72</i>	116
	<i>HTT</i>	118
	<i>NIPA1</i>	120
	Novel ALS risk genes.....	121
	<i>ARX</i>	122
	<i>FRA10AC1</i>	124
	<i>PABPN1</i>	126
	Probable ALS risk genes	127
	<i>AFF2</i>	128
	<i>AR</i>	130
	<i>ATXN3</i>	132
	<i>ATXN8</i>	134
	<i>CNBP</i>	136
	<i>FMR1</i>	137
	<i>FRA11A</i>	140
	<i>JPH3</i>	142
	<i>TK2-BEAN1</i>	143
	Possible ALS risk genes	145
	<i>ATN1</i>	146
	<i>ATXN7</i>	148
	<i>DMPK</i>	150
	<i>GIPC1</i>	152
	<i>NOP56</i>	154
	<i>NOTCH2NLC</i>	156
	<i>NUTM2B-AS1</i>	158
	<i>PPP2R2B</i>	160
	<i>TBP</i>	162
	Genes not associated with ALS risk	163
	<i>AFF3</i>	164

	ATXN10.....	166
	CACNA1A	168
	CSTB.....	170
	DIP2B	172
	FXN	174
	LRP12.....	176
	TCF4.....	178
	ZNF713.....	180
	<i>Genotyping results (FTD)</i>	<i>182</i>
3.4	DISCUSSION.....	185
	<i>Previously reported ALS risk genes.....</i>	<i>185</i>
	<i>Novel ALS risk genes</i>	<i>187</i>
	<i>Probable ALS risk</i>	<i>188</i>
	<i>Possible ALS risk</i>	<i>193</i>
	<i>Genes not associated with ALS risk.....</i>	<i>194</i>
	<i>FTD</i>	<i>194</i>
	<i>Site of onset, age at onset and disease duration in ALS</i>	<i>194</i>
3.5	SUMMARY.....	200
4	CHAPTER 4	202
	BURDEN OF NDRE GENES IN THE INCREASE OF ALS RISK IN THE IRISH POPULATION	202
4.1	INTRODUCTION.....	202
	<i>Research aims.....</i>	<i>204</i>
4.2	METHODOLOGY.....	205
	<i>PCR Genotyping</i>	<i>205</i>
	<i>C9orf72 expanded genotypes</i>	<i>205</i>
	<i>Data processing, statistical analysis, and plotting.....</i>	<i>205</i>
4.3	RESULTS	208
	<i>Burden of NDRE genes in ALS risk</i>	<i>208</i>
	<i>Gene co-occurrence.....</i>	<i>218</i>
	<i>Burden of 11 NDREs genes and ALS risk</i>	<i>218</i>
	<i>Age of onset and disease duration.....</i>	<i>219</i>
4.4	DISCUSSION.....	229
	<i>Burden of NDRE genes in ALS risk</i>	<i>229</i>
	<i>Gene-Gene Interactions in NDREs and their Relationship to ALS Risk in the Irish Population.....</i>	<i>230</i>
	<i>Age of onset and disease duration.....</i>	<i>230</i>
4.5	SUMMARY.....	233
5	CHAPTER 5	235

EXPLORING THE UTILITY OF TARGETED THIRD-GENERATION SEQUENCING AS A METHOD FOR REPEAT EXPANSION GENOTYPING.....	235
5.1 INTRODUCTION	235
<i>Research aims</i>	238
5.2 METHODOLOGY	239
<i>Study Participants</i>	239
<i>Cas9-mediated enrichment coupled to ONT sequencing</i>	239
<i>Data processing and analysis</i>	241
<i>Basecalling using Guppy</i>	241
<i>ONT sequence data analysis</i>	242
Basic Aligners	242
RE genotyping	243
RepeatHMM	243
Sniffels2	244
SquiggleKit	245
5.3 RESULTS	246
<i>RepeatHMM</i>	246
<i>Sniffles2 and SquiggleKit</i>	251
5.4 DISCUSSION	255
<i>RepeatHMM</i>	255
<i>Sniffles2 and SquiggleKit</i>	256
5.5 CONCLUSIONS AND FUTURE DIRECTIONS	258
6 CHAPTER 6.....	259
Discussion, Limitations & Future Direction	259
REFERENCES	269
APPENDIX A.....	329
SUPPLEMENTARY TABLES	329
2 CHAPTER 2.....	330
3 CHAPTER 3.....	363
4 CHAPTER 5.....	492
SUPPLEMENTARY FIGURES	500
2 CHAPTER 2.....	501
2.1 PANEL 3: <i>AFF3, CNBP, CSTB</i> AND <i>ZNF713</i> GENES	502
<i>Multiplex 3: CNBP and CSTB genes</i>	502

<i>Multiplex 4: AFF3 and ZNF713 genes</i>	505
<i>FLA results</i>	510
2.2 PANEL 4: <i>ATXN1, DMPK AND NIPA1</i> GENES.....	511
<i>Multiplex 5</i>	511
<i>FLA results</i>	517
2.3 PANEL 5: <i>DIP2B AND PABPN1</i> GENES	518
<i>Multiplex 6</i>	518
<i>FLA results</i>	523
2.4 PANEL 6: <i>AFF2, ARX AND FMR1</i> GENES	524
<i>Multiplex 7: AFF2 and ARX genes</i>	524
<i>Individual 8: FMR1 gene</i>	528
<i>FLA results</i>	530
2.5 PANEL 7: <i>AR, ATXN8, FRA11A AND NOTCH2NLC</i> GENES.....	531
<i>Multiplex 9: AR and ATXN8 genes</i>	531
<i>Multiplex 10: FRA11A and NOTCH2NLC genes</i>	535
<i>FLA results</i>	539
2.6 PANEL 8: <i>ATXN2, C9ORF72, GIPC1 AND NUTM2B-AS1</i> GENES	540
<i>Multiplex 11: ATXN2 and GIPC1 genes</i>	540
<i>Multiplex 12: C9orf72 and NUTM2B-AS1 genes</i>	543
<i>FLA results</i>	547
2.7 PANEL 9: <i>ATXN3, CACNA1A, FXN AND TCF4</i> GENES.....	548
<i>Multiplex 13: ATXN3 and TCF4 genes</i>	548
<i>Multiplex 14: CACNA1A and FXN genes</i>	552
<i>FLA results</i>	556
2.8 PANEL 10: <i>ATN1, FRA10AC1 AND LRP12</i> GENES	557
<i>Multiplex 15: ATN1 and FRA10AC1 genes</i>	557
<i>Individual 16: LRP12 gene</i>	560
<i>FLA results</i>	563
3 CHAPTER 3	564
APPENDIX B	580

Table of figures

FIGURE 1-1: SCHEMATIC REPRESENTATION OF REPEAT EXPANSIONS IN GENES ASSOCIATED WITH NEUROLOGIC DISEASES.	11
FIGURE 1-2: MOLECULAR MECHANISM DRIVING REPEAT EXPANSION PATHOGENESIS.	15
FIGURE 2-1: REPRESENTATION OF A PCR THERMAL CYCLING.	34
FIGURE 2-2 MULTIPLEX PCR OVERVIEW	41
FIGURE 2-3 GRAPHICAL REPRESENTATION FOR RPPCR ANALYSIS OF THE PRIMER BINDING.	42
FIGURE 2-4 AGAROSE GEL ELECTROPHORESIS WORKFLOW.	43
FIGURE 2-5 SCHEMATIC CONFIGURATION OF A CE INSTRUMENT FOR DNA SEPARATION.	44
FIGURE 2-6: ELECTROPHEROGRAM OF A MULTIPLEX PCR.....	46
FIGURE 2-7: <i>ATXN2</i> GENE ELECTROPHEROGRAM.	47
FIGURE 2-8 <i>FMR1</i> RPPCR ELECTROPHEROGRAM OBTAINED FROM THE SeqSTUDIO™.	47
FIGURE 2-9 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE PCR PRODUCTS FROM <i>HTT</i> AND <i>JPH3</i> PRIMERS USING TAQ DNA POLYMERASE HIGH FIDELITY (LIFE TECHNOLOGIES™).	51
FIGURE 2-10 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE PCR PRODUCTS FROM <i>HTT</i> AND <i>JPH3</i> PRIMERS.....	53
FIGURE 2-11 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE PCR PRODUCTS FROM <i>HTT</i> AND <i>JPH3</i> NEW SET OF PRIMERS.....	53
FIGURE 2-12 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE TEMPERATURE GRADIENT PCR PRODUCTS OF <i>ATXN7</i> AND <i>ATXN10</i>	55
FIGURE 2-13 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE PCR PRODUCTS OF <i>ATXN7</i> AND <i>ATXN10</i>	56
FIGURE 2-14 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE PCR PRODUCT OF <i>ATXN10</i> 2F/2R.	56
FIGURE 2-15 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE MULTIPLEX PCR PRODUCT OF <i>ATXN7</i> , <i>ATXN10</i> , <i>HTT</i> AND <i>JPH3</i>	57
FIGURE 2-16 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE MULTIPLEX PCR PRODUCTS FROM A GDNA GRADIENT OF <i>ATXN7</i> , <i>ATXN10</i> , <i>HTT</i> AND <i>JPH3</i>	58
FIGURE 2-17 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE INDIVIDUAL PCR PRODUCTS OF <i>NOP56</i> , <i>PPP2R2B</i> AND <i>TBP</i>	59
FIGURE 2-18 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE MULTIPLEX PCR PRODUCTS OF <i>NOP56</i> , <i>PPP2R2B</i> AND <i>TBP</i>	60
FIGURE 2-19 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE OPTIMISATION OF THREE DIFFERENT SETS OF PRIMERS FOR THE RE ON <i>TK2-BEAN1</i> GENE.	61

FIGURE 2-20 2-21 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE OPTIMIZED PCR PRODUCTS OF <i>TK2-BEAN1</i>	62
FIGURE 2-22 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE GRADIENT OF TEMPERATURE OF THE MULTIPLEX PCR OF <i>NOP56, PPP2R2B, TBP, TK2-BEAN1</i> AND <i>CACNA11A</i> GENES AND THEIR INDIVIDUAL PCRS. 63	
FIGURE 2-23 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE MULTIPLEX PCR PRODUCTS FROM A GDNA GRADIENT OF <i>NOP56, PPP2R2B, TBP, TK2-BEAN1</i> AND <i>CACNA11A</i>	64
FIGURE 2-24 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE PCR PRODUCTS OF THE OPTIMIZED MULTIPLEX PCR FOR OF <i>NOP56, PPP2R2B, TBP</i> AND <i>TK2-BEAN1</i>	64
FIGURE 2-25 CAPILLARY ELECTROPHEROGRAM OF THE MULTIPLEX 1: <i>HTT, JPH3, ATXN7</i> AND <i>ATXN10</i> GENES.	79
FIGURE 2-26 CAPILLARY ELECTROPHEROGRAM OF THE MULTIPLEX 2: <i>NOP56, PPP2R2B, TBP</i> AND <i>TK2-BEAN1</i> GENES.	80
FIGURE 2-27 SEQUENCING ELECTROPHEROGRAMS OF THREE DIFFERENT INDIVIDUALS FOR <i>HTT</i> CAG REPEAT COUNT CALIBRATION. ..	84
FIGURE 2-28 SEQUENCING ELECTROPHEROGRAMS OF THREE DIFFERENT INDIVIDUALS FOR <i>JPH3</i> CAG REPEAT COUNT CALIBRATION. .	85
FIGURE 2-29 SEQUENCING ELECTROPHEROGRAMS OF THREE DIFFERENT INDIVIDUALS FOR <i>ATXN10</i> ATTCT REPEAT COUNT CALIBRATION.	85
FIGURE 2-30 SEQUENCING ELECTROPHEROGRAMS OF THREE DIFFERENT INDIVIDUALS FOR <i>TK2-BEAN</i> TGGAA INSERTION/ TAAAA REPEAT COUNT CALIBRATION.	86
FIGURE 2-31 CAPILLARY ELECTROPHEROGRAM RESULTS OBTAINED FROM THE SeqStudio™ OF <i>DIP2B</i> rpPCR OPTIMISATION.	90
FIGURE 2-32 CAPILLARY ELECTROPHEROGRAM FROM THE OPTIMISED rpPCR FOR THE <i>DIP2B</i> GENE OBTAINED FROM THE SeqStudio™.	90
FIGURE 2-33 CAPILLARY ELECTROPHEROGRAM OF THE <i>DIP2B</i> rpPCR PRODUCTS OF TWO INDIVIDUALS WITH SUSPECTED CGG EXPANSIONS.	92
FIGURE 3-1 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE <i>ATXN1</i> GENE.	113
FIGURE 3-2 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE <i>ATXN2</i> GENE.	115
FIGURE 3-3 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE GGGGCC REPEAT EXPANSION IN THE <i>C9ORF72</i> GENE.	117
FIGURE 3-4 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE <i>HTT</i> GENE.	119
FIGURE 3-5 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGC REPEAT EXPANSION IN THE <i>NIPA1</i> GENE.	121
FIGURE 3-6 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGC REPEAT EXPANSION IN THE <i>ARX</i> GENE.	123
FIGURE 3-7 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGG REPEAT EXPANSION IN THE <i>FRA10AC1</i> GENE.	125
FIGURE 3-8 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE GCG REPEAT EXPANSION IN THE <i>PABPN1</i> GENE.	127

FIGURE 3-9 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CCG REPEAT EXPANSION IN THE <i>AFF2</i>	
GENE.	129
FIGURE 3-10 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE <i>AR</i>	
GENE.	131
FIGURE 3-11 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE	
<i>ATXN3</i> GENE.	133
FIGURE 3-12 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE	
<i>ATXN8</i> GENE.	135
FIGURE 3-13 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CCTG REPEAT EXPANSION IN THE	
<i>CNBP</i> GENE.	137
FIGURE 3-14 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGG REPEAT EXPANSION IN THE <i>FMR1</i>	
GENE.	139
FIGURE 3-15 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGG REPEAT EXPANSION IN THE	
<i>FRA11A (C11ORF80)</i> GENE.	141
FIGURE 3-16 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CTG REPEAT EXPANSION IN THE <i>JPH3</i>	
GENE.	143
FIGURE 3-17 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE TAAAA REPEAT EXPANSION IN THE	
<i>TK2-BEAN1</i> GENE.	145
FIGURE 3-18 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE	
<i>ATN1</i> GENE.	147
FIGURE 3-19 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE	
<i>ATXN7</i> GENE.	149
FIGURE 3-20 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CTG REPEAT EXPANSION IN THE <i>DMPK</i>	
GENE.	151
FIGURE 3-21 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE GGC REPEAT EXPANSION IN THE	
<i>GIPC1</i> GENE.	153
FIGURE 3-22 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE GGCCTG REPEAT EXPANSION IN THE	
<i>NOP56</i> GENE.	155
FIGURE 3-23 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE GGC REPEAT EXPANSION IN THE	
<i>NOTCH2NLC</i> GENE.	157
FIGURE 3-24 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGG REPEAT EXPANSION IN THE	
<i>NUTM2B-AS1</i> GENE.	159
FIGURE 3-25 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE	
<i>PPP2R2B</i> GENE.	161
FIGURE 3-26 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE <i>TBP</i>	
GENE.	163

FIGURE 3-27 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CCG REPEAT EXPANSION IN THE <i>AFF3</i> GENE.	165
FIGURE 3-28 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE ATTCT REPEAT EXPANSION IN THE <i>ATXN10</i> GENE.	167
FIGURE 3-29 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE POLYQ REPEAT EXPANSION IN THE <i>CACNA1A</i> GENE.	169
FIGURE 3-30 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CCCC GCCCGCG REPEAT EXPANSION IN THE <i>CSTB</i> GENE.	171
FIGURE 3-31 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGG REPEAT EXPANSION IN THE <i>DIP2B</i> GENE.	173
FIGURE 3-32 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE GAA REPEAT EXPANSION IN THE <i>FXN</i> GENE.	175
FIGURE 3-33 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGG REPEAT EXPANSION IN THE <i>LRP12</i> GENE.	177
FIGURE 3-34 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CTG REPEAT EXPANSION IN THE <i>TCF4</i> GENE.	179
FIGURE 3-35 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE CGG REPEAT EXPANSION IN THE <i>ZNF713</i> GENE.	181
FIGURE 3-36 THE COMPOSITION OF THE POPULATION OF FTD AND CONTROL SAMPLES FOR THE REPEAT EXPANSIONS IN THE <i>AR</i> , <i>ARX</i> , <i>ATXN1</i> AND <i>ATXN3</i> GENES.	183
FIGURE 3-37 THE COMPOSITION OF THE POPULATION OF FTD AND CONTROL SAMPLES FOR THE REPEAT EXPANSIONS IN THE <i>CNBP</i> , <i>C9ORF72</i> AND <i>NUTM2B-AS1</i> GENES.	184
FIGURE 4-1 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL.	210
FIGURE 4-2 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL, EXCLUDING <i>C9ORF72</i> GENE.	211
FIGURE 4-3 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL, EXCLUDING <i>C9ORF72</i> AND <i>ATXN2</i> GENES.	212
FIGURE 4-4 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL, EXCLUDING <i>C9ORF72</i> , <i>ATXN2</i> AND <i>NIPA1</i> GENES.	213
FIGURE 4-5 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL COMBINING 0-1 AND 2+.	214
FIGURE 4-6 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL COMBINING 0-1 AND 2+, EXCLUDING <i>C9ORF72</i> GENE.	215
FIGURE 4-7 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL COMBINING 0-1 AND 2+, EXCLUDING <i>C9ORF72</i> AND <i>ATXN2</i> GENES.	216
FIGURE 4-8 BURDEN OF ALL NDRES IN ALS RISK USING 99TH-PERCENTILE MODEL COMBINING 0-1 AND 2+, EXCLUDING <i>C9ORF72</i> , <i>ATXN2</i> AND <i>NIPA1</i> GENES.	217
FIGURE 4-9 GENE-GENE INTERACTION INCLUDING ALL NDRE GENES.	221
FIGURE 4-10 GENE-GENE INTERACTION INCLUDING ALL NDRE GENES, EXCLUDING <i>C9ORF72</i>	222
FIGURE 4-11 BURDEN FROM THE CLUSTER OF 11 GENES IN ALS RISK USING 99TH-PERCENTILE MODEL.	223

FIGURE 4-12 BURDEN OF REPEAT EXPANSIONS ON AGE OF ONSET.	225
FIGURE 4-13 BURDEN OF REPEAT EXPANSIONS ON DISEASE DURATION (EXCLUDING ONE OUTLIER).	226
FIGURE 4-14 BURDEN OF REPEAT EXPANSIONS ON AGE OF ONSET FROM THE CLUSTER OF 11 GENES.	227
FIGURE 4-15 BURDEN OF REPEAT EXPANSIONS ON DISEASE DURATION FROM THE CLUSTER OF 11 GENES.	228
FIGURE 5-1 Cas9-MEDIATES PCR-FREE ENRICHMENT PREPARATION FOR SEQUENCING.	241
FIGURE 5-2 IGV VISUALIZATION OF BWA AND NGMLR ALIGNMENTS FOR THE <i>AFF3</i> GENE IN CONTROL 1.	248
FIGURE 5-3 GRAPHIC REPRESENTATION OF THE GENOTYPE RESULTS FROM REPEATHMM COMPARED TO PCR RESULTS.	249
FIGURE 5-4 DISTRIBUTION OF REPEAT COUNTS WITHIN INDIVIDUAL READS SUPPORTING REPEATHMM GENOTYPES.	251
FIGURE 5-5 CAPILLARY ELECTROPHEROGRAM OF THE <i>C9orf72</i> RPPCR FOR SAMPLE 1675.	252
FIGURE 5-6 THE SEQUENCE OF THE ALTERNATE ALLELE FOR A 15,845 BASEPAIR HOMOZYGOUS INSERTION AT <i>C9orf72</i> CALLED BY SNIFFLES2.	253
FIGURE 5-7 RAW CURRENT SIGNAL FROM SAMPLE 1675 ON THE <i>C9orf72</i> REPEAT EXPANSION REGION.	254
SUPPLEMENTARY FIGURE S 2-1 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>CNBP</i> AND <i>CSTB</i> GENES.	502
SUPPLEMENTARY FIGURE S 2-2 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE MULTIPLEX PCR PRODUCTS FROM A GDNA GRADIENT OF <i>CNBP</i> AND <i>CSTB</i> GENES.	502
SUPPLEMENTARY FIGURE S 2-3 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>AFF3</i> AND <i>ZNF713</i> GENES.	505
SUPPLEMENTARY FIGURE S 2-4 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE MULTIPLEX PCR PRODUCTS FROM A GDNA GRADIENT OF <i>AFF3</i> AND <i>ZNF713</i> GENES.	506
SUPPLEMENTARY FIGURE S 2-5 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEXES 3 AND 4: <i>CNBP</i> , <i>CSTB</i> , <i>AFF3</i> AND <i>ZNF713</i> GENES.	510
SUPPLEMENTARY FIGURE S 2-6 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>ATXN1</i> , <i>NIPA1</i> AND <i>DMPK</i> GENES	511
SUPPLEMENTARY FIGURE S 2-7 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE MULTIPLEX PCR PRODUCTS FROM A GDNA GRADIENT OF <i>ATXN1</i> , <i>NIPA1</i> AND <i>DMPK</i> GENES.	512
SUPPLEMENTARY FIGURE S 2-8 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEX 5: <i>ATXN1</i> , <i>NIPA1</i> AND <i>DMPK</i> GENES.	517
SUPPLEMENTARY FIGURE S 2-9 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>DIP2B</i> AND <i>PABPN1</i> GENES.	518
SUPPLEMENTARY FIGURE S 2-10 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON 2.5% AGAROSE GEL OF THE MULTIPLEX PCR PRODUCTS FROM A GDNA GRADIENT OF <i>DIP2B</i> AND <i>PABPN1</i> GENES.	519
SUPPLEMENTARY FIGURE S 2-11 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEX 6: <i>DIP2B</i> AND <i>PABPN1</i> GENES.	523
SUPPLEMENTARY FIGURE S 2-12 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>AFF2</i> AND <i>ARX</i> GENES.	524

SUPPLEMENTARY FIGURE S 2-13 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL PCR PRODUCTS OF <i>FMR1</i> GENE.	528
SUPPLEMENTARY FIGURE S 2-14 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEX 7 AND INDIVIDUAL PCR 8: <i>AFF2</i> , <i>ARX</i> AND <i>FMR1</i> GENES.	530
SUPPLEMENTARY FIGURE S 2-15 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>AR</i> AND <i>ATXN8</i> GENES.	531
SUPPLEMENTARY FIGURE S 2-16 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>FRA11A</i> AND <i>NOTCH2NLC</i> GENES.	535
SUPPLEMENTARY FIGURE S 2-17 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEXES 9 AND 10: <i>AR</i> , <i>ATXN8</i> , <i>FRA11A</i> AND <i>NOTCH2NLC</i> GENES.	539
SUPPLEMENTARY FIGURE S 2-18 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>ATXN2</i> AND <i>GIPC1</i> GENES.	540
SUPPLEMENTARY FIGURE S 2-19 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>C9orf72</i> AND <i>NUTM2N-AS1</i>	543
SUPPLEMENTARY FIGURE S 2-20 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEXES 11 AND 12: <i>ATXN2</i> , <i>GIPC1</i> , <i>C9orf72</i> AND <i>NUTM2B-AS1</i> GENES.	547
SUPPLEMENTARY FIGURE S 2-21 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>ATXN3</i> AND <i>TCF4</i>	548
SUPPLEMENTARY FIGURE S 2-22 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>CACNA1A</i> AND <i>FXN</i>	552
SUPPLEMENTARY FIGURE S 2-23 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEXES 13 AND 14: <i>ATXN3</i> , <i>TCF4</i> , <i>CACNA1A</i> AND <i>FXN</i> GENES.	556
SUPPLEMENTARY FIGURE S 2-24 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL AND MULTIPLEX PCR PRODUCTS OF <i>ATN1</i> AND <i>FRA10AC1</i>	557
SUPPLEMENTARY FIGURE S 2-25 PHOTOGRAPHIC RECORDING OF THE ELECTROPHORETIC RUN ON AGAROSE GEL 2.5% OF THE INDIVIDUAL PCR PRODUCTS OF <i>LRP12</i> GENE.	560
SUPPLEMENTARY FIGURE S 2-26 CAPILLARY ELECTROPHEROGRAM OF MULTIPLEX 15 AND INDIVIDUAL PCR 16: <i>ATN1</i> , <i>FRA10AC1</i> AND <i>LRP12</i> GENES.	563
SUPPLEMENTARY FIGURE S 3-1 THE COMPOSITION OF THE POPULATION OF ALS AND CONTROL SAMPLES FOR THE SHORT ALLELES OF THE 35 NDREs GENES.	567
SUPPLEMENTARY FIGURE S 3-2 THE COMPOSITION OF THE POPULATION OF FTD AND CONTROL SAMPLES FOR THE SHORT ALLELES OF THE 35 NDREs GENES.	570
SUPPLEMENTARY FIGURE S 3-3 THE COMPOSITION OF THE POPULATION OF FTD AND CONTROL SAMPLES FOR THE LONG ALLELES OF THE 35 NDREs GENES.	579

Table of tables

TABLE 1-1 SUMMARY OF MOST COMMON GENES IN ALS.	6
TABLE 1-2 REPEAT EXPANSIONS AFFECTING THE NERVOUS SYSTEM.	9
TABLE 1-3: NEURODEGENERATION REPEAT EXPANSION (NDREs) GENES.	16
TABLE 1-4 NDRE DISEASES BY LOCATION AND PATHOMECHANISM.	22
TABLE 2-1 REPEAT COUNT CALIBRATION RESULTS FOR EACH GENE OBTAINED BY SEQUENCING.	83
TABLE 2-2 ALLELE-PLATE REPEAT COUNT CALIBRATION USING REPLICATED CONTROL 1 RESULTS.	88
TABLE 2-3 GENE/COST COMPARISON OF MULTIPLE NDREs PANELS.....	96
TABLE 2-4 BUDGET COMPARISON OF SEVERAL PROTOCOLS FOR THE NDREs PROJECT.	96
TABLE 3-1 SUMMARY OF THE MOST RELEVANT NDREs GENES DISORDERS FOR THIS PROJECT.	102
TABLE 3-2 GENERAL DEMOGRAPHICS.....	106
TABLE 4-1 THRESHOLD TABLE	207
TABLE 4-2 GENE COMBINATION FROM ALS PATIENTS WITH 3 LONG EXPANSIONS.....	224
SUPPLEMENTARY TABLE S 2-1 PCR COMPONENTS OF EACH 35 NDREs GENES.	331
SUPPLEMENTARY TABLE S 2-2 DETAILS OF THE OPTIMIZED PRIMERS FROM ALL THE NDREs GENES.....	336
SUPPLEMENTARY TABLE S 2-3 PCR COMPONENTS OF EACH FLA PANEL / MULTIPLEX PCR.	340
SUPPLEMENTARY TABLE S 2-4 COMPLETE REPEAT COUNT CALIBRATION RESULTS FOR EACH GENE OBTAINED BY SEQUENCING.	344
SUPPLEMENTARY TABLE S 2-5 COMPLETE ALLELE-PLATE REPEAT COUNT CALIBRATION USING REPLICATED CONTROL 1 RESULTS.	350
SUPPLEMENTARY TABLE S 2-6 RPPCR CONDITIONS	360
SUPPLEMENTARY TABLE S 3-1 ALS SHORT ALLELE STATISTICAL INFORMATION FROM ALL 35 NDREs GENES PLOTS.	364
SUPPLEMENTARY TABLE S 3-2 ALS LONG ALLELE STATISTICAL INFORMATION: POPULATION PROPORTIONS AND ORs.	380
SUPPLEMENTARY TABLE S 3-3 ALS LONG ALLELE STATISTICAL INFORMATION: SITE OF ONSET.	409
SUPPLEMENTARY TABLE S 3-4 ALS LONG ALLELE STATISTICAL INFORMATION: DISEASE DURATION AND AGE OF ONSET.....	438
SUPPLEMENTARY TABLE S 3-5 FTD LONG ALLELE STATISTICAL INFORMATION: POPULATION PROPORTIONS AND ORs.....	468
SUPPLEMENTARY TABLE S 4-1 CHOP CHOP	493
SUPPLEMENTARY TABLE S 4-2 IDT.....	496

Summary

This thesis aims to advance our understanding of the genetic mechanisms that contribute to amyotrophic lateral sclerosis (ALS). The focus is on investigating pleiotropy in genetic contributors to ALS by examining the existence of multiple repeat expansions linked to other neurodegenerative diseases. Through these efforts, we hope to improve patient diagnosis and treatment by enhancing our knowledge of ALS risk genes.

The first chapter of the thesis details the development of a polymerase chain reaction (PCR) panel for detecting intermediate repeat motifs in 35 neurodegenerative repeat expansion (NDREs) genes. This panel was created with the primary objective of simultaneously detecting multiple intermediate repeat expansions (RE) while preserving finite genomic DNA (gDNA) samples. This methodology is accurate, and results can be used in various studies, such as downstream bioinformatics, to enhance our understanding of neurodegenerative diseases.

The aim of the second chapter in this thesis is to investigate the hypothesis that rare repeat expansions contribute to a significant proportion of ALS cases. To achieve this, we utilized the PCR panel developed in chapter 1 to measure 35 NDRE genes in >1500 ALS patients and >1700 controls from the Irish ALS DNA bank, which allowed us to improve upon earlier, underpowered studies that only examined some of the NDREs. Our study provides evidence for the potential involvement of several genes, such as *C9orf72*, *FRA10AC1*, *ARX*, and *PABPN1*, in determining ALS risk, disease duration, and site of onset in patients. However, replication and further research is needed to fully understand the implications of these findings and their potential for developing targeted interventions for ALS patients with varying clinical phenotypes.

Following the genotyping of 35 NDRE genes in more than 1500 ALS patients and over 1700 controls from the Irish DNA bank, the aim of the third study in this thesis was to assess the individual-level burden of expanded repeats in NDRE genes in the Irish population with ALS. Our investigation focused on exploring the potential for additive oligogenic gene action and analysing the association between the number of repeat expansions and developing ALS, age of onset, and disease duration in the Irish population. The study revealed that the risk of developing

ALS increases with the number of intermediate or long repeats in certain genes, with up to 5 expansions associated with a higher risk (p -value: 1.07×10^{-9}). Our findings showed that the risk of developing ALS is influenced by the number of long repeat expansions, extending our understanding of the role of REs in ALS beyond the monogenic actions of specific genes such as *C9orf72*, *ATXN2*, or *NIPA1*. We also observed a statistically significant relationship between the number of repeat expansions and the age of onset (a decrease of 0.468 years) and disease duration (a decrease of 2.523 months), implying that individuals with a higher burden of expanded repeats may experience an earlier onset and a shorter disease duration.

Finally, the last chapter explores the utility of Nanopore sequencing technology in generating long-read sequences for genotyping repeat expansions. A Cas9-enrichment protocol that does not rely on PCR amplification is used to evaluate the performance of this alternative approach. The study demonstrates the opportunities provided by signal-level analysis of individual sequence reads, suggesting that Nanopore experiments may be better suited for discovery of repeat expansions rather than accurate genotyping by conventional alignment/variant calling pipelines, and that combining multiple approaches may be necessary for comprehensive characterization of repeat expansions.

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.

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., Siah, W. F., Chang, K. M., McKenna, M. C., Doherty, M. A., **Hengeveld, J. C.**, Vajda, A., Donaghy, C., Hutchinson, S., McLaughlin, R. L., S., Hardiman,

O., & Bede, P. (2021). Validation of diagnostic criteria for early-stage primary lateral sclerosis: Disease burden in “Probable PLS” (1527). *Journal of Neurology*, 96(15 Supplement), 1527.

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. (2022). Mapping cortical disease-burden at individual-level in frontotemporal dementia:

implications for clinical care and pharmacological trials. *Brain Imaging and Behavior*, 16 (3), 1196–1207.

Chapter 1

Introduction

Amyotrophic Lateral Sclerosis

Neurodegeneration is a process characterised by progressive and irreversible damage and loss of neurons and is a common final pathway in ageing and neurodegenerative diseases such as Huntington's disease, Parkinson's disease, Alzheimer's disease and amyotrophic lateral sclerosis (ALS) (Rowland and Shneider 2001). ALS is a late-onset fatal neurodegenerative disease of the human motor system, affecting the upper and lower motor neurons principally. The motor neurons are localized in the brain, brain stem and spinal cord, and they are the link of communication between the nervous system and all voluntary muscles of the body. Electrical messages are transmitted from the upper motor neurons (UMN, brain motor neurons) to the lower motor neurons (LMN, spinal cord motor neurons) and then to each muscle (Miller et al. 2003). The term amyotrophic usually refers to the atrophy of the muscle fibres due to the degeneration on their corresponding anterior horn cells, affecting the muscles. "Lateral Sclerosis" relates to the hardness of the anterior and lateral corticospinal tracts, where degeneration occurs, which are then replaced by gliosis (Rowland and Shneider 2001).

ALS covers a large variety of neurodegenerative syndrome spectrums. The presence of UMN and LMN features including the innervation of the multiple spinal cord regions and brain stem are the clinical hallmarks of ALS. ALS can present with either limb-onset disease (around 70%), or bulbar-onset (around 35%), which involve speech and swallowing difficulties and later development of limb features; or with an initial respiratory difficulty (around 5%). The most common symptoms are reduced exercise capacity and fatigue. Ultimately, the patient requires assistance for daily living activities. Most patients develop a respiratory compromise which leads to dyspnoea, orthopnoea and morning headaches (Leigh 2003). Malnutrition and weight

loss can be a consequence of dysphagia, which commonly is associated with poor prognosis (Ramirez et al. 2008). ALS is considered a very aggressive progressive disease; the survival rate is 20% within 5 to 10 years after the symptom onset, and 50% die within 30 months. Different studies have shown that for individuals of older age, the presentation of early respiratory muscle dysfunction and bulbar onset disease is associated with a low survival rate, while for a younger age, the independent predictors of prolonged survival rate are the presentation of limb-onset disease and extended diagnostic delay (Talbot 2009).

Since the 1990s, there has been a gain of interest in the clinical and experimental research of ALS. Main advances were the discovery of the involvement of glutamate in the neurotransmitter system and the identification of causative genes in familial ALS (FALS), leading to the identification and understanding of the survival rates in ALS by dependent factors, such as age of onset, disease progression, clinical presentation (phenotype), presence of respiratory failure and nutritional status (Fang et al. 2008). Worldwide, ALS affects 2 to 5 per 100,000 individuals (National Library of Medicine, 2016). Different studies reported an incidence of ALS in Europe of 2-3 per 100,000 (Chiò et al. 2013; Hardiman et al. 2017). The epidemiological characteristics of ALS patients vary by their race, region and nationality (Jun et al. 2018). A population-based study in 2007 showed that ALS is less common in a diverse population with Spanish genetic origin (Cronin, Hardiman, and Traynor 2007). A study based in the mortality disease rates showed that Cuba has 60% lower ALS deaths than European and North American populations, supporting the previous findings of lower frequency of ALS in Hispanic origins (Zaldivar et al. 2009). Other studies have shown that males have a higher frequency of developing ALS than women (3 per 100,000 and 2.4 per 100,000 people per year, respectively); while for familial ALS, women and men have the same incidence 2.4 per 100,000 people per year (Logroscino et al. 2010).

Classification of ALS and contributing factors

ALS is usually classified into “familial ALS” (FALS), in which another member of the proband’s family has or had ALS, and “sporadic ALS” (SALS), in which there is not a family history of the disease. Around 5 to 10% of ALS has a family history with a Mendelian pattern

of inheritance, and it varies depending on the gene that is involved. In most cases, the autosomal dominant pattern is seen, and less frequently are autosomal recessive patterns. Is very rare for ALS to be inherited in an X-linked dominant pattern (when the location of the disease gene is associated with the X chromosome) (National Library of Medicine, 2016).

Around 90-95% of individuals with ALS are described as sporadic (SALS). Usually, the age of onset seen in SALS patients is that it occurs in their late fifties or early sixties (Testa et al. 2004). The cause of SALS is mostly unknown, but it seems to be a combination of a genetic component and environmental factors. There is much information about the different environmental factors that could be related with the development of ALS, such as radiation and electromagnetic fields, chemical exposure and heavy metals (fertilisers, herbicides, insecticides, formaldehyde etc.) (Yu et al. 2014; Zarei et al. 2015). However, these data are based on different questionnaires that are given to patients and rely upon their memories, leading to recall bias. Hence, there is a lack of information on these environmental risk factors about their frequency and amount of exposure. Moreover, it is challenging to isolate a single environmental factor due to the late age at onset of ALS (Zarei et al. 2015). Research on family aggregation suggests that individuals with SALS have an overlap between common neurodegenerative disorders and ALS, suggesting there is a possibility that some genes can increase the overall risk of neurodegeneration between family members (Fallis and Hardiman 2009).

Although 5 to 10% of ALS cases are classified as FALS, this can vary by population and is also dependent on the level of detail in which family histories are taken. For example, in a comprehensive study of extended ALS pedigrees diagnosed between 1996 and 2016, the rate of FALS in Ireland was found to be around 20% (Ryan et al. 2018). Typically, the age of onset of FALS is around a decade earlier than sporadic cases. Also, in particular populations, such as certain parts of Japan and South West New Guinea, there is a geographical locus form for ALS where prevalence is 50-100 times higher in specific locations than in another part of the world (Kurtzke 1982; Kuzuhara et al. 2001; Kuzuhara and Kokubo 2005).

Since 2009, researchers have identified at least thirteen genes and loci that cause typical ALS clinical phenotypes. The most studied genes are included in Table 1-1. Other genes involved in ALS can carry repeat expansions in the genes, which can create a malfunctioned or cytotoxic

protein that could lead to the disease. *C9orf72* and *ATXN2* are two genes in which repeat expansions occur (Chen et al. 2016; Woollacott and Mead 2014).

Repetitive sequences and neurodegenerative diseases

A large portion of the human genome is composed of repetitive sequences (30%), and for most species, alterations on the length of these sequences in the DNA have created diversity during evolution. However, in humans, rapid alteration of these repeat loci has increased the diseases associated with them (Pearson et al. 2011). Repeat expansions (REs) underlie more than 40 diseases, most of them affecting the nervous system (In the realm of genetic variation, certain repetitive sequences within the human genome have captured the attention of researchers due to their intriguing propensity to undergo repeat expansions and contractions. These dynamic alterations, termed repeat expansions and repeat contractions, harbour the potential to bestow both health and disease outcomes. Understanding the molecular mechanisms underpinning these phenomena is pivotal to unravelling their implications in multiple neurodegenerative diseases.

Repeat expansions and contractions refer to the dynamic alterations in the length of specific DNA sequences characterized by repetitive motifs. These motifs, typically short sequences repeated in tandem, can be found across various regions of the genome. During DNA replication and repair processes, these sequences are prone to slippage events, leading to the amplification or reduction of the repetitive motifs. This phenomenon becomes particularly consequential when it occurs within or near protein-coding regions, as the altered repeat length can perturb the encoded protein's structure or function. Depending on the affected gene and the nature of the encoded protein, such events can contribute to both physiological diversity and pathological conditions.

The mechanisms underlying repeat expansions and contractions are inherently linked to DNA replication, repair, and recombination processes. For instance, one well-studied mechanism is the formation of hairpin-like structures during DNA replication, known as slipped-strand DNA structures. These structures can lead to the aberrant synthesis of DNA strands, resulting in the incorporation of additional repeats or the loss of existing ones. Another mechanism involves the

interaction of DNA-binding proteins that stabilize or destabilize repeat-containing regions, subsequently influencing the likelihood of repeat length alterations.

Several genes have become paradigms for studying repeat expansions and contractions. The gene responsible for Huntington's disease, HTT, features a trinucleotide repeat (CAG) within its coding region. As the repeat expands beyond a certain threshold, it triggers the onset of the disease. Similarly, within the context of myotonic dystrophy type 1 (DM1), the DMPK gene harbours a CTG repeat in its untranslated region, the expansion of which leads to the clinical manifestations of the disorder. Pioneering studies, such as those by Verkerk AJ and colleagues (1991) on Fragile X syndrome, have illuminated the propensity of certain repetitive sequences to dynamically change in length, sparking a cascade of research into the mechanisms governing these alterations (Verkerk et al 1991).

An illustrative instance of repeat contractions can be found in the recent 2023 paper titled 'Long-Read Nanopore Sequencing Identified D4Z4 Contractions in Patients with Facioscapulohumeral Muscular Dystrophy.' This study, among others, sheds light on the dynamic nature of repetitive sequences and their implications in genetic disorders. Within the context of facioscapulohumeral muscular dystrophy (FSHD), the contraction of D4Z4 repeats in the D4Z4 repeat array has been recognized as a pivotal factor in disease pathogenesis. Through the application of advanced long-read nanopore sequencing techniques, the authors meticulously unravelled the intricacies of these contractions, providing unprecedented insights into the mechanisms governing such alterations and use this technique for molecular diagnosis (Petra yeetong 2023).

Table 1-2). The first REs to be discovered were all trinucleotide repeats (TNR), which were suggested to be the most common feature of the type of disease in which they were observed. More recently, it has been observed that the size of the repeat motif ranges from trinucleotides to tetranucleotides, pentanucleotides, hexanucleotides and dodecanucleotides (Paulson 2018).

Table 1-1 Summary of most common genes in ALS.

GENE	PROTEIN PRODUCT	FUNCTION	% FALS	REFERENCE
<i>SOD1</i>	Superoxide dismutase 1	Binds to molecules of copper and zinc to break down toxic, charged oxygen molecules called superoxide radicals.	20%	(Kiernan et al. 2011; Rosen et al. 1993)
<i>TARDBP</i> (<i>TDP-43</i>)	TAR DNA binding protein	Gene expression and regulation (transcription, RNA splicing, transport and translation) and small RNAs.	5 - 10%	(Kiernan <i>et al.</i> , 2011)
<i>FUS</i>	Fusion in Sarcoma	Gene expression and regulation (transcription, RNA splicing, transport and translation) and small RNAs.	5%	(Greenway et al. 2006)
<i>ANG</i>	Angiogenin	Regulates transcription of the RNA	1%	(Greenway <i>et al.</i> , 2006)
<i>OPTN</i>	Optineurin	Involved in diverse cellular functions, including NF- κ B regulation, membrane trafficking, exocytosis, vesicle transport, reorganization of actin and microtubules, cell cycle control, and autophagy.	rare	(Belzil et al. 2011; Kiernan et al. 2011)

In the realm of genetic variation, certain repetitive sequences within the human genome have captured the attention of researchers due to their intriguing propensity to undergo repeat expansions and contractions. These dynamic alterations, termed repeat expansions and repeat contractions, harbour the potential to bestow both health and disease outcomes. Understanding the molecular mechanisms underpinning these phenomena is pivotal to unravelling their implications in multiple neurodegenerative diseases.

Repeat expansions and contractions refer to the dynamic alterations in the length of specific DNA sequences characterized by repetitive motifs. These motifs, typically short sequences repeated in tandem, can be found across various regions of the genome. During DNA replication

and repair processes, these sequences are prone to slippage events, leading to the amplification or reduction of the repetitive motifs. This phenomenon becomes particularly consequential when it occurs within or near protein-coding regions, as the altered repeat length can perturb the encoded protein's structure or function. Depending on the affected gene and the nature of the encoded protein, such events can contribute to both physiological diversity and pathological conditions.

The mechanisms underlying repeat expansions and contractions are inherently linked to DNA replication, repair, and recombination processes. For instance, one well-studied mechanism is the formation of hairpin-like structures during DNA replication, known as slipped-strand DNA structures. These structures can lead to the aberrant synthesis of DNA strands, resulting in the incorporation of additional repeats or the loss of existing ones. Another mechanism involves the interaction of DNA-binding proteins that stabilize or destabilize repeat-containing regions, subsequently influencing the likelihood of repeat length alterations.

Several genes have become paradigms for studying repeat expansions and contractions. The gene responsible for Huntington's disease, HTT, features a trinucleotide repeat (CAG) within its coding region. As the repeat expands beyond a certain threshold, it triggers the onset of the disease. Similarly, within the context of myotonic dystrophy type 1 (DM1), the DMPK gene harbours a CTG repeat in its untranslated region, the expansion of which leads to the clinical manifestations of the disorder. Pioneering studies, such as those by Verkerk AJ and colleagues (1991) on Fragile X syndrome, have illuminated the propensity of certain repetitive sequences to dynamically change in length, sparking a cascade of research into the mechanisms governing these alterations (Verkerk et al 1991).

An illustrative instance of repeat contractions can be found in the recent 2023 paper titled 'Long-Read Nanopore Sequencing Identified D4Z4 Contractions in Patients with Facioscapulohumeral Muscular Dystrophy.' This study, among others, sheds light on the dynamic nature of repetitive sequences and their implications in genetic disorders. Within the context of facioscapulohumeral muscular dystrophy (FSHD), the contraction of D4Z4 repeats in the D4Z4 repeat array has been recognized as a pivotal factor in disease pathogenesis. Through the application of advanced long-read nanopore sequencing techniques, the authors meticulously

unravelling the intricacies of these contractions, providing unprecedented insights into the mechanisms governing such alterations and use this technique for molecular diagnosis (Petra yeetong 2023).

Table 1-2 Repeat expansions affecting the nervous system.

(taken and modified from (Cardoso and Marques 2018; Paulson 2018; Rohilla and Gagnon 2017))

<i>Size repeat unit</i>	<i>Repeat motif</i>	<i>Neurologic disease</i>
<i>Trinucleotides</i>	CAG	>10 diseases: Huntington's Disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar atrophy (Kennedy disease), and seven spinocerebellar ataxias (SCA 1-3, 6-8, 12, 17).
	CGG	Fragile X tremor ataxia syndrome (FMR1), fragile X, fragile X mental retardation type 1 and type 2, mental retardation syndrome (FRA12A type).
	CTG	Huntington's disease-like2 (HDL2), fuchs corneal dystrophy (FECD), myotonic dystrophy type 1 (DM1), SCA8.
	GCC	FRAXE mental retardation.
	GAA	Fiederich ataxia, hereditary spastic paraplegia type 6.
	GCG	Oculopharyngeal muscular dystrophy (OPMD), X-linked mental retardation, Hereditary Spastic Paraplegia type 6 (SPG6).
<i>Tetranucleotides</i>	CCTG	Myotonic dystrophy type 2 (DM2).
<i>Pentanucleotides</i>	ATTCT	Spinocerebellar ataxia type 10 (SCA 10).
	ATTTC	Spinocerebellar ataxia 37 (SCA 37).
	TGGAA	Spinocerebellar ataxia type 31 (SCA31).
<i>Hexanucleotides</i>	GGCCTG	Spinocerebellar ataxia type 36 (SCA36).
	GGGGCC	<i>C9orf72</i> frontotemporal dementia (FTD)/ amyotrophic lateral sclerosis (ALS).
<i>Dodecanucleotides</i>	CCCCGCCCGCG	Myoclonic epilepsy (EPM1; Uverricht-Lundborg myoclonic epilepsy).

RE length and severity of the disease

The inheritance of REs is very dynamic when transmitted to the next generation as the sizes often change. Hence, the phenotypic range of disease may vary in affected individuals, and between members of the same family (Paulson 2018). The length of REs is important, and changes can be crucial. Coding sequences for REs become unstable around 29 to 35 units in length. However, for non- coding regions, a premutation (an expanded allele that is not sufficiently long to be pathogenic) starts between 55 to 200 units, and increase per generation

between 100-10,000 units. For both non-coding and coding alleles, as the length of the expansion increases and exceeds the threshold length, the likelihood of another unstable event in that site increases as well. Hence, the anticipation of the disease sometimes occurs from one generation to the next, becoming more severe and exhibiting an earlier age of onset (Kovtun and McMurray 2008; Mirkin 2007). Figure 1-1 shows a schematic view of the genes that cause neurological diseases from Table 1-3, underlying their repeat expansion motif, as well as their gene location (5' or 3' untranslated regions, protein-coding exons and introns).

Most diseases have a genotype-phenotype correlation between the severity of the disease and the repeat length. Patients tend to show earlier onset symptoms, and the disease gets more severe when the number of repeats gets longer. This characteristic is very well represented for the CAG (polyglutamine; polyQ) diseases. Around nine disorders have moderate CAG expansion, encoding an abnormal increase of glutamine in their respective disease protein. All nine CAG diseases have an inverse correlation between the age of onset and the length of the repeat: longer REs usually cause earlier and more profound symptoms and signs of the disease (Andrew et al. 1993; Figueroa, Verdugo, and Pezoa Nunez 2017; Rosenblatt et al. 2012). Many factors contribute to the age of onset of a RE disorder. For example, the length of the CAG RE is determinant for age of onset, accounting for 45-70% of this effect, and the remaining contribution is from other minor genetic modifiers and environmental factors that are under investigation (Paulson 2018).

Similarly, the relationship between age of onset/severity of the disease and length of the repeat expansion is very strong for several non-polyQ expanded repeats (i.e. FXS/FXTAS, Friederichs ataxia, DM1 and several SCAs). On the other hand, for *C9orf72*-mediated ALS/FTD, SCA8 and DM2, the correlation between the severity of the disease and RE is weak. The last three diseases are caused by complicated and large repeats, and the molecular features of the behaviour of the pathogenic repeat length are still under investigation (Paulson 2018).

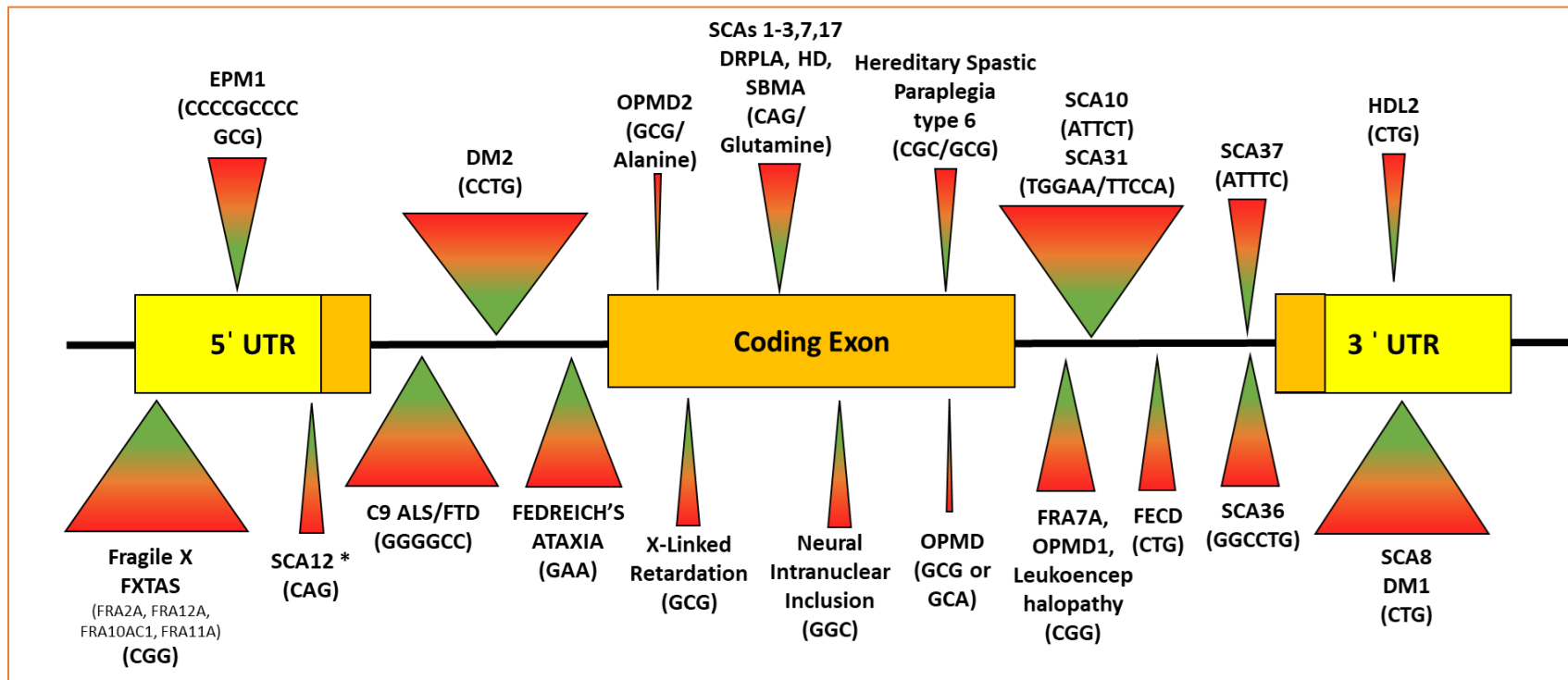


Figure 1-1: Schematic representation of repeat expansions in genes associated with neurologic diseases.

Each gene is schematically allocated into their location and linked to their RE and disease. The different size of the triangles represents approximately the variable sizes of the pathogenic RE associated with each disease. Green represents the normal length of RE, orange is intermediate RE and red is expanded. For SCA12*, the RE is in an intron, but has also promoter elements (taken and modified from Paulson, 2018).

Inheritance mechanisms of REs

Disease caused by REs have different modes of inheritance. The majority are inherited in an autosomal-dominant manner, a couple are autosomal recessive (i.e. DM1 and Friedreich's ataxia), SBMA and FXTAS are X-linked, and FXS and FRAXE mental retardation are X-linked recessive. The different molecular mechanisms leading to repeat expansion can be explained in the context of the different modes of inheritance (Paulson 2018).

There are different mechanisms associated with the natural expansion of tandem repeats. DNA replication in dividing cells is the major contributory factor of REs. During replication, the repetitive sequences may cause a problem in the fork, inverting or switching the template and adding an extra repeat (Kerrest et al. 2009; Mirkin 2007). The polymerase can cause expansions on either strand (leading or lagging) when moving, causing an uneven or imperfect Okazaki fragment and adding an extra repeat to complete the fragment ligation (Gadgil et al. 2017; Sloan, Gleizes, and Bohnsack 2016). RE events are related to DNA damage repair and non-canonical DNA structures (hairpins or quadruplexes) (Gordenin, Kunkel, and Resnick 1997; Spiro et al. 1999). There are various repair mechanisms: 1) replacement of DNA bases, including nucleotide excision repair (NER) and base excision repair (BER); 2) mismatch repair (MMR); which has been the primary driver of RE. MMR recognises small bulges and hairpins and expands the repeats. The enzyme MutS β identifies the odd structure, and to repair the damage it recruits the polymerase Pol β which, because of a distorted template, inserts extra repeats (Rohilla and Gagnon 2017).

The increase of expansions caused by MMR usually is on the order of one to three repeats at a time. Fast accumulation of uninterrupted REs through MMR or different mechanisms of DNA repair might be promoted by transcription on the repeat. For the repeat CGG on FXS, studies on mouse models have shown that transcription is essential for the expansion to occur (Zhao and Usdin 2015). Similarly, studies have confirmed that transcription at REs causes repeat instability, such as the formation of R-loops and DNA-RNA hybrids (Reddy et al. 2014). Other mechanisms of repeat instability are the oxidation of free DNA strands or misalignment on the reannealing strand, which leads to DNA damage repair signalling (Bak, Sakellariou, and Pena-Diaz 2014). Across the literature of the instability of repeats and their expansion is DNA metabolism, in which the DNA strand separates and reanneals at the RE sequence. These events

can drive the formation of different structures on the DNA, such as canonical structures, activating DNA damage responses that can lead to adding more repeats. These mechanisms of repair may lead to REs and cause multiple diseases (Bak et al. 2014; Iyer et al. 2015; Zhao and Usdin 2015).

With the discovery and recognition that expanded repeats are inherited dynamically, and as long as the RE is associated with increased instability, the phenomenon of clinical anticipation can be explained. The transmission of expanded repeats to the next generation tends to increase, coupled with early age of onset and severe disease. The extent of anticipation and the influence of the parent vary between diseases. Several CAG diseases are marked by a paternal effect. For example, a child born with a mutation in the genes responsible for HD, SCA2 and SCA7 inherited from the father can develop earlier symptoms than their fathers and, in some cases, even before the father develops the disease (Trang et al. 2015). On the other hand, for DM1, the mutation is always inherited from the mother (Ashizawa et al. 1992; Tsilfidis et al. 1992). However, not all RE diseases have significant anticipation or an identified parent-origin effect. For example, OPMD has a small RE size and does not show anticipation (Paulson 2018).

Pathomechanisms in repeat expansion diseases

Since the discovery of STRs in 1964 (KORNBERG et al. 1964), there has been significant growth in understanding how REs arise and contribute to more than 50 human disorders. The locations of disease-causing expansions are within protein-coding exons (CDS), promoters, 5' and 3' UTRs (untranslated regions), or introns. These repeats, with different expansion motifs, have an increasing impact on our understanding of cellular and molecular biology and neuroscience (Table 1-3). Research concerning these REs has provided insights into DNA-repair, chromatin biology RNA-protein condensates, non-canonical translation, and protein aggregation. Even though more than 1.5 million STRs have been identified in the genome (Fotsing et al. 2019), our knowledge of their non-pathogenic functions remains poorly understood (Balzano, Pelliccia, and Giunta 2021; Gymrek 2017; Hannan 2018).

STR expansions cause architectural changes in the structure of the DNA, and can produce a series of molecular processes through gain of function (GOF) or loss of function (LOF) mechanisms either at the DNA, RNA, or protein levels (Figure 1-2). These mechanisms may

cause cell toxicity, which is modulated by somatic instability of the repeat, in which a stochastic process leads to different lengths across the patient's tissues and cells over their lifetime (Malik et al. 2021; Mirkin 2007; La Spada and Taylor 2010).

Pathomechanisms of REs diseases can be classified into four main interconnected categories (Figure 1-2): 1. DNA-based mechanisms, such as LOF via repeat-induced transcriptional silencing and GOF via R-loop formation and activation of the DNA damage response. 2. GOF mediated by toxic RNA via gelation and RNA-binding protein (RBP) sequestration. 3. AUG-initiated repeat-harboring native protein GOF. 4. Repeat-associated non-AUG (RAN) translation of toxic repetitive peptides (Chintalaphani et al. 2021; Malik et al. 2021; Mirkin 2007; Rohilla and Gagnon 2017; La Spada and Taylor 2010). Table 1-4 shows detailed information on each NDRE gene, RE motif, associated neurodegenerative disease, and their associated pathogenic molecular mechanism (if known).

Understanding the biology of RNA in repeat expansion disorders includes the understanding of the abundance or concentration of RNA and its turnover and decay pathways. Latest research highlights the significance of characterising the abundance of cellular expanded tandem repeat-containing RNA (xtrRNA). Reports suggest that only one or some xtrRNA could be enough to generate focal aggregates. Therefore, knowing the concentration and type of xtrRNA in the cell is important for a better understanding of the role of it in the disease and interpretation of the data (Gudde et al. 2017; Taylor, Brown, and Cleveland 2016).

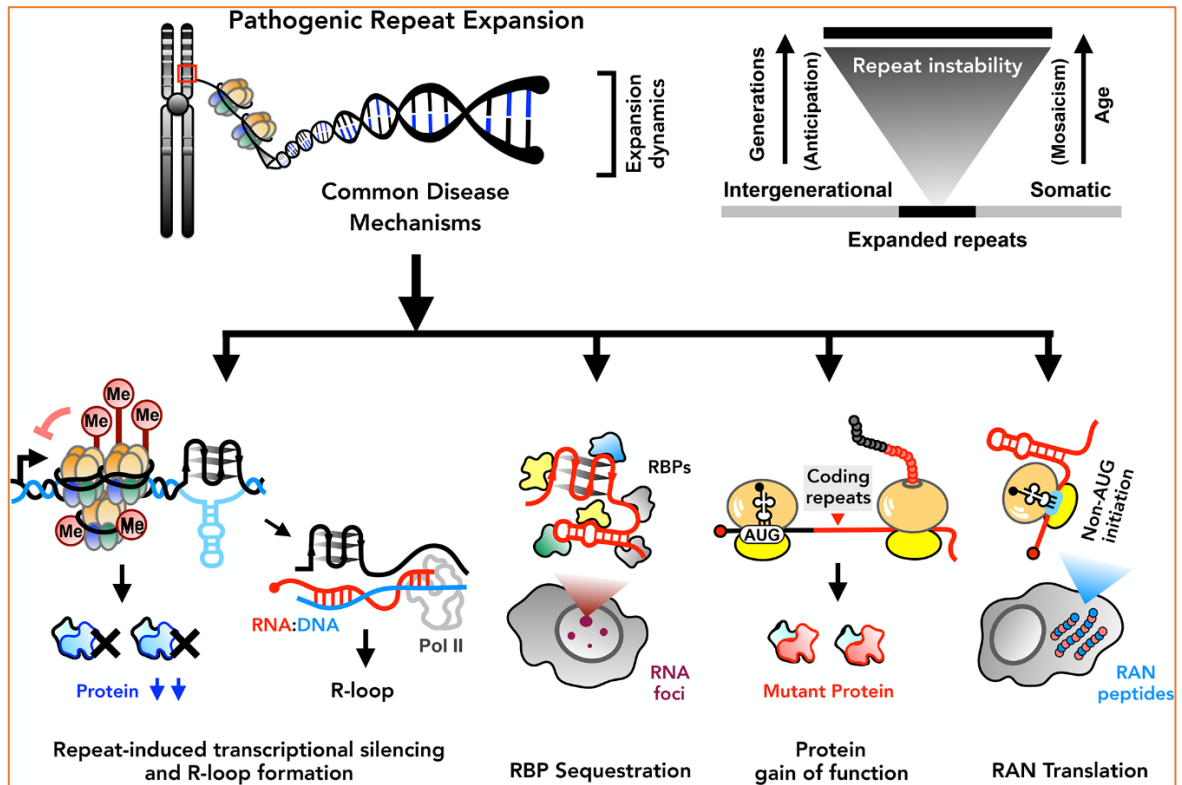


Figure 1-2: Molecular mechanism driving repeat expansion pathogenesis.

Hyper-methylation of promoter regions can result in transcriptional silencing, which can result in partial or complete loss of the native protein containing the repeats. Active transcription through the repeats, on the other hand, can cause the formation of R-loops (RNA:DNA hybrids that activate the DNA damage response pathway) and potentially exacerbate somatic repeat instability. Transcribed repeat RNAs fold into complex structures that interact with and sequester cellular RNA-binding proteins in an abnormal manner (RBPs). Expansions of trinucleotide repeats in protein-coding sequences result in mutant proteins that cause gain-of-function toxicity. Finally, in the absence of canonical AUG-mediated initiation, the coding and non-coding repeat RNAs are translated via repeat-associated non-AUG (RAN) translation, yielding toxic polymeric peptides. (Taken and modified from Malik *et al.*, 2021).

Table 1-3: Neurodegeneration repeat expansion (NDREs) genes.

Gene	Disorder name	Locus	Genomic Location	Tandem repeat motif (amino acid repeat)	Number of repeats:			Expanded repeats results in:				Reference
					Normal Alleles	Premutation/ Intermediate	Pathogenic length	Gene Silencing	xtrRNA transcription	xtrRNA proteins	xtrRNA foci	
AFF2/FMR3 (FMR2-AF4/FMR2 family member 2)	FRAXE mental retardation	Xq28	5'UTR	CCG	4-39	40-200	200-900	Yes*	No	No	No	Fazal <i>et al.</i> , 2020; Jorge <i>et al.</i> , 2013; Knight <i>et al.</i> , 1993; Rohilla & Gagnon, 2017
AFF3 (AF4/FMR2 family member 3)	FRA2A (Fragile site, 2 chromosome, A site)	2q11.1	Intron	CCG	8-17	Unknown	>300	No	Yes*/No	No	No	Metsu <i>et al.</i> , 2014a; Rohilla & Gagnon, 2017
AR (androgen receptor)	SBMA (spinal and bulbar muscular atrophy)	Xq11-q12	Coding	CAG	9-34	35-37	>38 -68	No	Yes	Yes*	L.D.	Ackerman <i>et al.</i> 2012; Fazal <i>et al.</i> , 2020; La Spada <i>et al.</i> 1991; Lieberman <i>et al.</i> , 2019; Rohilla & Gagnon, 2017
ARX (aristaless related homeobox)	X-linked mental retardation (XLMR)	Xp22.13	Coding	GCG GCN	7-10 7-16	Unknown Unknown	>20 >23	Yes	No	No	No	Fazal <i>et al.</i> , 2020; Jorge, Oliveira, Marques, & Santos, 2013; la Spada <i>et al.</i> , 2010; (Chintalaphani et al. 2021)
ATN1 (atrophin 1)	DRPLA (dentatorubral - pallidoluysian atrophy)	12p13.3 1	Coding	CAG	3-35	36-47	48-93	No	Yes	Yes*	L.D.	Fazal <i>et al.</i> , 2020; Koide <i>et al.</i> , 1994; Lieberman <i>et al.</i> , 2019; Nagafuchi <i>et al.</i> , 1994;

												Rohilla & Gagnon, 2017
<i>ATXN1</i> <i>(Ataxin 1)</i>	SCA1	6p23	Coding	CAG	6-35	36-38	39 -88	No	Yes	Yes*	L.D.	Fazal <i>et al.</i> , 2020; Lieberman <i>et al.</i> , 2019; Lin <i>et al.</i> , 2008, Orr <i>et al.</i> , 1993; Rohilla & Gagnon, 2017
<i>ATXN2</i> <i>(Ataxin 2)</i>	SCA2	12q24.1 2	Coding	CAG	15-24	27-33	34-77	No	Yes	Yes*	L.D.	Fazal <i>et al.</i> , 2020
<i>ATXN3</i> <i>(Ataxin3)</i>	SCA3	14q24-q31	Coding	CAG	13-36	37-60	61-84	No	Yes	Yes*	Yes	Bettencourt <i>et al.</i> , 2008; Fazal <i>et al.</i> , 2020; Kawaguchi <i>et al.</i> , 1994; Lieberman <i>et al.</i> , 2019; Rohilla & Gagnon, 2017
<i>ATXN7</i> <i>(Ataxin 7)</i>	SCA7	3p21-p12	Coding	CAG	7-18	18-37	37-460	No	Yes	Yes*	L.D.	David <i>et al.</i> , 1997; Fazal <i>et al.</i> , 2020; Katagiri <i>et al.</i> , 2015; Lieberman <i>et al.</i> , 2019; Rohilla & Gagnon, 2017
<i>ATXN8</i> <i>(ATXN8OS</i> <i>(Ataxin)</i>	SCA8	13q21	Both	CTG·CAG	16-34	51-79	>74	No	Yes	Yes*	Yes*	Fazal <i>et al.</i> , 2020; Ikeda <i>et al.</i> 2000; Koob <i>et al.</i> 1998; Koob <i>et al.</i> , 1999; Tanaka <i>et al.</i> 2011; Rohilla & Gagnon, 2017

ATXN10 (Atanxin 10)	SCA10	22q13.3	Noncoding	ATTCT (not translated)	10-32	N/A	>280-4500	No	Yes	?	Yes*	Fazal <i>et al.</i> , 2020; Matsuura <i>et al.</i> 2000; Wang <i>et al.</i> 2008
CACNA1A (calcium voltage-gated channel subunit alpha1 A)	SCA6	19p13	Coding	CAG (Q)	4-18	19-20	21-33	No	Yes	Yes*	L.D.	Fazal <i>et al.</i> , 2020; Katayama <i>et al.</i> 2000; Lieberman <i>et al.</i> , 2019; Zhuchenko <i>et al.</i> , 1997
CNBP (CCHC-type zinc finger nucleic acid binding protein)	DM2 (Myotonic dystrophy type 2)	3q21.3	Intron	CCTG	11-26	27-49	> 50	No	Yes*/No	?	Yes*	Fazal <i>et al.</i> , 2020; Liquori <i>et al.</i> 2001; Sun <i>et al.</i> 2011
CSTB (cystatin B)	EPM1 (Epilepsy progressive myoclonia)	21q22.3	Promoter	CCCCGC CCC GCG	4-11	Unknown	30- 125	Yes*	Yes/No	No	No	Fazal <i>et al.</i> , 2020; Laloti <i>et al.</i> 1997
C9orf72 (Chromosome 9 Open Reading Frame 72)	FTD/ALS	9p21.2	Intron	GGGGCC	2-25		>30	No	Yes	Yes*	Yes*	
FRA11A (C11orf80-Chromosome 11 Open Reading Frame 80)	Mental retardation, FRA11A (Fragile site, 11 chromosome, A site)	11q13.2	5'UTR	CGG	6-8	Unknown	500	Yes*	?	?	Yes*	Debacker <i>et al.</i> , 2007
DIP2B (Disco interacting protein 2 homolog B)	Mental retardation syndrome, FRA12A type (Fragile site, 12 chromosome, A site)	12q13.1 2	5'UTR	CGG	6-23	Unknown	≥ 350	No	?	?	?	Fazal <i>et al.</i> , 2020; Winneppenninck x <i>et al.</i> 2007

DMPK (<i>DM1 protein kinase</i>)	DM1 (Myotonic dystrophy type 1)	19q13.3	3'UTR	CTG	5-35	36-50	≥ 50	No	Yes*	Yes	Yes*	Fazal <i>et al.</i> , 2020; Mahadevan <i>et al.</i> , 1992; Singh <i>et al.</i> 2014, laferink, 2019; Kamsteeg, E. J. <i>et al.</i> , 2012; Lavedan, C. <i>et al.</i> , 1993; Warner, J. P. <i>et al.</i> 1996
FMR1 (<i>FMRP translational regulator 1</i>)	Fragile X-associated Tremor Ataxia Syndrome				5-53	Unknown	55-200	No	Yes	Yes*	Yes*	CChen <i>et al.</i> 2001; Fazal <i>et al.</i> , 2020; Heitz <i>et al.</i> , 1991; Kenneson <i>et al.</i> , 2001
	Fragile X mental retardation type 2	Xq27.3	5'UTR	CGG	5-53	Unknown	200-900					
	Fragile X tremors ataxia syndrome				5-53	54-230	> 200					
FRA10AC1 (<i>FRA10A associated CGG repeat 1</i>)	Mental Retardation, FRA10AC1 type	10q21.33	5'UTR	CGG	8-14	Unknown	> 200 CGG	No	No	No	Yes	Li, <i>et al.</i> , 2015
FXN (<i>FRATAXIN</i>)	Fredreich's ataxia (FRDA)	9q13-21.1	Intron	GAA	5-34	35-65	> 200	Yes*	Yes/No	No	No	Campuzano <i>et al.</i> , 1996; Fazal <i>et al.</i> , 2020; Rizik <i>et al.</i> 2014
GIPC1 (GIPC PDZ Domain Containing Family Member 1)	Oculopharyngo distal myopathy 2 (OPDM2)	19p13.12	Coding	GGC	12 - 32	Unknown	> 35					Deng <i>et al.</i> , 2020; Fazal <i>et al.</i> , 2020
HTT (<i>Huntingtin</i>)	HD (Huntington's disease)	4p16.3	Coding	CAG	6-35	Unknown	> 35	No	Yes	Yes*	L.D.	Krause <i>et al.</i> 2015; Fazal <i>et al.</i> , 2020; Lieberman <i>et al.</i> , 2019; MacDonald <i>et al.</i> , 1993

JPH3 (Junctophilin 3)	HDL2 (Huntington's disease-like2)	16q24.2	Coding	CTG	7-28	Unknown	66-78	No	Yes	?	Yes*	Holmes <i>et al.</i> , 2001; ; Fazal <i>et al.</i> , 2020; Krause <i>et al.</i> 2015
LRP12 (LDL receptor related protein 12)	OPDM1 (Oculopharyngodistal myopathy)	8q22.3	5'UTR	CGG	13-45	Unknown	unknown					;Fazal <i>et al.</i> , 2020; Ishiura <i>et al.</i> 2019
NIPA1 (NIPA magnesium transporter 1)	SPG6 (Hereditary Spastic Paraplegia type 6)	15q11.2	Coding	CGC/GCG	7-8	Unknown	> 8	Yes	Yes	Yes	Yes	Blauw <i>et al.</i> , 2012; Fazal <i>et al.</i> , 2020
NOP56 (ribonucleoprotein)	SCA36	20p13	Intron	GGCCTG	3-8 // 3-14	Unknown	> 650	No	Yes	?	Yes*	Fazal <i>et al.</i> , 2020; Kobayashi <i>et al.</i> , 2011; Lee <i>et al.</i> , 2016; Obayashi <i>et al.</i> , 2015.
NOTCH2NLC [NBPF19 (Notch 2 N-terminal like C)]	NIID (Neural Intranuclear Inclusion disease)	1q21.2	5'UTR	GGC	< 40	Unknown	66 -517					Sone <i>et al.</i> 2019, Ishiura <i>et al.</i> 2019
NUTM2B-AS1 / LOC642361 NUTM2B Antisense RNA 1)	Leukoencephalopathy	10q22.3	Noncoding	CGG/CCG	3-16	Unknown						Ishiura <i>et al.</i> 2019
PABPN1 (poly(A) binding protein nuclear 1)	OPMD (Oculopharyngeal muscular dystrophy)	14q11	Coding	GCG	6	Unknown	8-18	No	Yes	Yes*	No	Brais <i>et al.</i> , 1997; Richard <i>et al.</i> , 2017; Fazal <i>et al.</i> , 2020
PPP2R2B (protein phosphatase 2 regulatory subunit Bbeta)	SCA12/ Alzheimer	5q31-33	5'UTR	CAG	7-28	Unknown	66-78	No	Yes*	No	No	Brusco <i>et al.</i> , 2004; Fazal <i>et al.</i> , 2020; lone <i>et al.</i> , 2016.
TBP (TATA-box binding protein)	SCA17	6q27.18	Coding	CAG	25-42	Unknown	47-63	No	Yes	Yes*	L.D.	Fazal <i>et al.</i> , 2020; Lieberman <i>et al.</i> , 2019; Lone <i>et al.</i> , 2015.

<i>TCF4</i> (transcription factor 4)	FECD (Fuchs endothelial corneal dystrophy)	18q21.2	Intron	CTG	10-40	Unknown	> 50	No	?	No	Yes*	wieben 2012,2017; Fazal <i>et al.</i> , 2020
<i>TK2-BEAN1</i> (thymidine kinase 2- brain expressed associated with NEDD4 1)	SCA31	16q21	Intron	TGGAA	0	Unknown	>45	No	Yes	Yes*	Yes*	Fazal <i>et al.</i> , 2020; Sato <i>et al.</i> , 2009.
				TAAAA	8-21	Unknown	2.5-3.8 kb					
<i>ZNF713</i> (zinc finger protein 713)	Autistic Spectrum, FRA7A type (Fragile site, 7 chromosome, A site)	7p11.2	Intron	CGG	5-22	Unknown	> 85	No	Yes*/No	No	No	Fazal <i>et al.</i> , 2020; Metsu <i>et al.</i> , 2014

L.D.: Length-dependent; *: Indicates the most likely repeat-associated disease mechanism(s) for that disorder; ?: not known.

Table 1-4 NDRE diseases by location and pathomechanism.

Gene	Repeat motif	Disease	Protein	Location	Pathomechanism				
					Somatic instability	LOF	RBP sequestration	AUG-initiated protein GOF	RAN translation
<i>AFF2</i>	CCG	FRAXE	ALF transcription elongation factor 2	5' UTR	*	*			
<i>AFF3</i>	CGG	FRA2A	ALF transcription elongation factor 3	5' UTR	*	*			
<i>AR</i>	CAG	SBMA	Androgen receptor	CDS	*	*		*	
<i>ARX</i>	GCG	XLID	Aristaless related homeobox	CDS	*	*			
<i>ATN1</i>	CAG	DRPLA	Atrophin-1	CDS	*		*	*	
<i>ATXN1</i>	CAG	SCA1	Ataxin -1	CDS	*	*		*	
<i>ATXN10</i>	ATTCT	SCA10	Ataxin-10	Intron	*		*		
<i>ATXN2</i>	CAG	SCA2	Ataxin-2	CDS		*	*	*	
<i>ATXN3</i>	CAG	SCA3	Ataxin-3	CDS	*		*	*	
<i>ATXN7</i>	CAG	SCA7	Ataxin-7	CDS	*			*	

<i>ATXN8OS/ATXN8</i>	CTA/CAG	SCA8	ATXN8 opposite strand lncRNA	3' UTR/CD S			*	*	*
<i>C9orf72</i>	GGGGCC	FTD / ALS	C9orf72- SMCR8 complex subunit	Intron	*	*	*		*
<i>CACNA1A</i>	CAG	SCA6	Cave2.1	CDS				*	
<i>CNBP</i>	CCTG	DM2	CCHC-type zinc finger nucleic acid binding protein	Intron	*		*		*
<i>CSTB</i>	CCCCGC CCCGCG	EPM1	Cystatin B	Promoter		*			
<i>DIP2B</i>	CGG	FRA12A	Disco interacting protein 2 homolog B	5' UTR					
<i>DMPK</i>	CTG	DM1	DM1 protein kinase	3' UTR	*		*		*
<i>FMRI</i>	CGG	FXS	Fragile X	5' UTR	*		*		
<i>FMRI</i>	CGG	FXPOI	messenger ribonucleopr otein 1	5' UTR	*				*
<i>FMRI</i>	CGG	FXTAS		5' UTR	*		*		*

<i>FRA10AC1</i>	CGG	Mental Retardation, FRA10AC1 type	FRA10A associated CGG repeat 1	5' UTR					
<i>FRA11A</i>	CGG	Mental retardation, FRA11A (Fragile site,11 chromosome, A site)	Fragile site, folic acid type	5' UTR		*			
<i>FXN</i>	GAA	FA	Frataxin	Intron	*	*			
<i>GIPC1</i>	CGG	OPDM2	GIPC1 protein	5' UTR					
<i>HTT</i>	CAG	HD	Huntingtin	CDS	*		*	*	*
<i>JPH3</i>	CTG	HDL2	Junctophilin 3	3' UTR		*	*		
<i>LRP12</i>	CGG	OPDM1	LDL receptor related protein 12	5' UTR	*				
<i>NIPA1</i>	CGC/GCG	SPG6	NIPA magnesium transporter 1	CDS					
<i>NOP56</i>	GGCCTG	SCA36	NOP56 ribonucleoprotein	Intron			*		*

<i>NOTCH2NLC</i>	CGG	NIID	Notch 2 N-terminal like C	5' UTR			*		*
<i>NUTM2B-AS1</i>	CGG/CCG	OPML	NUT family member 2B antisense RNA 1	lncRNA	*				
<i>PABPN1</i>	CGC	OPMD	Poly(A) binding protein nuclear 1	CDS		*		*	
<i>PPP2R2B</i>	CAG	SCA12	Protein phosphatase 2 regulatory subunit beta	5' UTR				*	
<i>TBP</i>	CAG	SCA17	TATA-binding protein	CDS	*	*		*	
<i>TCF4</i>	CTG	FECD	Transcription factor 4	Intron	*		*		*
<i>TK2-BEAN1</i>	TGGAA/T TCCA	SCA31	Thymidine Kinase2/Bra n expressed associated with NEDD4 1	Intron			*		*

<i>ZNF713</i>	CGG	FRA7A	Zinc finger protein 713	Intron	*	*				
---------------	-----	-------	----------------------------	--------	---	---	--	--	--	--

(Taken and modified from Malik *et al.*, 2021)

Convergent disease mechanisms in ALS

Neurodegenerative diseases have, usually, overlapping symptoms that can be misread and confused within syndromes. Since the identification of TAR DNA-binding protein 43 (TDP-43) positive ubiquitinated cytoplasmic inclusions as a hallmark feature of ALS, studies have shown that more than half of patients with FTD, and almost all ALS cases exhibit this cellular feature in affected neurons, which has increased the interest in the overlap between these two neurodegenerative syndromes (Kiernan et al. 2011; Phukan et al. 2012). In a population-based study in Ireland, studies showed that 34.1% of ALS patients fulfilled criteria for cognitive impairment, while 13.8% of ALS cases meet the Neary criteria for FTD. Also, 46.9% of these patients had no cognitive impairment, and 14% had cognitive impairment without executive dysfunction (Phukan et al. 2012). Genes involved in familial clusters with cases of ALS or FTD or ALS-FTD are still unknown, but in 2011 researchers found a common locus on chromosome 9 corresponding to *C9orf72* (Renton et al. 2011).

The hexanucleotide repeat expansion (HRE), G₄C₂, has been identified as a common causative mutation for both ALS and FTD, adding them to the increasing class of noncoding repeat expansion disorders, including several spinocerebellar ataxias (SCA2, SCA8, SCA10), fragile-X-associated tremor/ataxia syndrome (FXTAS), and myotonic dystrophies (DM1 and DM2). The protein of *C9orf72* encodes 12 exons, three transcription variants, and two possible protein isoforms; and it is found in the cytoplasm of neurons and presynaptic terminals (Renton et al. 2011).

Renton *et al.* found that expansions more than 30 repeats in the *C9orf72* gene were present in 21.1% of SALS and 46.4% of FALS in a population in Finland (Renton et al. 2011). The number of HRE in *C9orf72* is critical for the determination of causation/susceptibility of ALS and FTD since patients with >30 HRE appears to confer risk either by RNA dysregulation or RAN translation which causes the motor neurons to accumulate dipeptide repeats. The pathogenic RE in *C9orf72* between 20-22 repeats appears to be related to FTD, 20-30 repeats have also shown to have a similar pattern of cognitive and behavioural impairment as those with more than 30 repeats. Therefore, knowing the exact number of repeats is important for ALS and FTD (Chen et al. 2016).

Another ALS risk gene with pathogenic RE is *ATXN2*, known for its role in SCA2. In 2010, researchers reported that malfunctions in the protein Ataxin-2 influence the TDP-43 toxicity seen in ALS. In this gene, the RE occurs in the protein-coding region of *ATXN2* and harbouring more than 34 CAG repeats causes SCA2, whereas between 27-33 repeats (intermediate length) act as a risk factor for ALS in 4.7% of patients from North America. Other studies have confirmed the association of this polyQ expansion in *ATXN2* with increment of ALS risk in other ethnic populations, such as European, Chinese and French Canadians (Chen *et al.*, 2011; Corrado *et al.*, 2011; Daoud *et al.*, 2011; Lee *et al.*, 2011; Ross *et al.*, 2011; Van Damme *et al.*, 2011). Such pleiotropy has been observed for many other ALS genes and exemplifies the existence of convergent disease mechanisms across neurodegenerative diseases (Kenna *et al.* 2013; McLaughlin *et al.* 2017). Table 1-3 shows a detailed number of genes that have been associated with neurodegenerative disorders, termed NDRE genes.

Scope and structure of thesis

The goal of this study is to uncover the pleiotropic effects of a broad range of neurodegenerative repeat expansions (REs) on the development of ALS, going beyond the known and understood factors. It is hypothesized that a GC-rich RE on any motor neuron-expressed transcript can act as a pleiotropic factor by affecting multiple cellular processes and contributing to the formation and persistence of pathological stress granules. By characterizing the pleiotropic effects of these REs, this research will shed light on the complex aetiology of ALS and pave the way for new therapeutic interventions. This research is presented over four chapters:

- Chapter 2: Development of an efficient, cost-effective, highly multiplexed genotyping panel for neurodegenerative repeat expansions.
 - This chapter aims to develop a multiplex genotyping panel for 35 NDREs genes by using PCR, gel electrophoresis, fragment length analysis (FLA), repeat-primed-PCR (rpPCR) and sequencing.
- Chapter 3: Population genetic analysis of the genetic pleiotropy of neurodegenerative repeat expansions in ALS and FTD.
 - This chapter aims to discover the association between all 35 NDREs genes and ALS risk.
- Chapter 4: Burden of NDRE genes in the increase of ALS risk.

- This chapter investigates the pleiotropy of the studied repeat expansion genes and ALS risk.
- Chapter 5: Exploring the utility of targeted third-generation sequencing as a method for repeat expansion genotyping.
 - The objective of this chapter is to investigate the effectiveness of Nanopore Technology in producing long-read sequences for genotyping repeat expansions, specifically through the implementation of Cas9-enrichment protocol.

Chapter 2

Development of an efficient, cost-effective, highly multiplexed genotyping panel for neurodegenerative repeat expansions.

2.1 Introduction

Short tandem repeats (STRs), also known as simple sequence repeats or microsatellites, are short, repeated DNA sequence motifs, typically of 1-6 bps in length, that constitute about 3 % of the entire human genome. Their distribution among chromosomes, however, is not uniform, they appear less frequently in subtelomeric areas (Koreth et al. 1996). The majority of STRs are located in noncoding areas, with just around 8% detected in coding regions (Ellegren 2000). Furthermore, their concentrations vary significantly between chromosomes. In humans, chromosome 19 contains the most significant number of STRs (Subramanian, Mishra, and Singh 2003). In the human genome, one STR occurs every 2,000 bp on average (Lander et al. 2001). Unique DNA sequences in a genome have a low mutation rate (about 1^{-9} mutations per nt each generation). In contrast, STR sequences have several orders of magnitude greater, ranging from 10^{-6} to 10^{-2} nt per locus per generation (Ellegren 2000). However poorly defined, a RE occurs when an STR is extended beyond the normal length observed in the healthy population (Fan and Chu 2007).

Despite their diversity, REs diseases share many genetic features. For example, all REs arise from a healthy population of polymorphic repeats, and the degree of the polymorphism around

the normal range varies between the diseases. The mutable normal repeats or “normal repeat range” have a higher tendency to expand during transmission, moving into the pathogenic range. This means that unaffected families can become affected if a de novo mutation occurs (Paulson 2018). Other general features that are common in diseases associated with REs include: 1) REs are unstable (dynamic), they often change size when transmitted to the following generation. 2) Longer repeats often cause the disease to be more severe and with earlier age of onset. 3) Clinical anticipation is common in successive generations, where the disease becomes more severe. 4) The overall phenotypes are variable, which can be associated with variation in repeat size.

TNRs are often found inside genes, and their number seems to be associated with the severity of diseases. For example, the CGG repeat, which encodes runs of arginine, is located on the 5' end of the FXTAS. The average number of repeats in an STR region is usually 29, but when repeats exceed 52 times, this can cause the STR region to become unstable and lead to rapid expansion. The CGG repeat number in the asymptomatic carrier ranges from 60 to 200. However, the patient with obvious symptoms has over 230 repeats of CGG (Liu et al. 2022). To better understand human neurodegenerative diseases, it is essential to understand the mutation process of STRs. The disorders caused by repeat expansion demonstrate the several ways that tandem repeats can alter gene structure and function: Exonic repeats can have relatively large effects on protein synthesis and function even with minor changes in repeat number, whereas the vast majority of the genome is made up of noncoding sequences, creating a large window of opportunity for repeat-mediated effects that do not involve changes in protein sequence to occur (Usdin 2008).

To date, at least 50 illnesses related to expansion loci have been documented, with half of them identified only in the previous ten years due to methodological challenges. These limitations continue to apply to the current commonly used molecular diagnostic approaches (exome or gene panels), resulting in missed diagnoses that are harmful to affected persons and their families, particularly for illnesses that are extremely rare and/or clinically indistinguishable. The majority of these illnesses have been uncovered through family-driven approaches, and many more are anticipated to be discovered. (Depienne and Mandel 2021). The discovery of tandem repeat expansions in human disorders is therefore still in its early stages, and many more ailments, whether monogenic or complex, are likely to be discovered. Furthermore, we have

seen a significant widening of phenotypes associated with known repeat expansions in recent years, as evidenced by the recent finding of *HTT* expansions in people with FTD and ALS (Dewan et al. 2021). As a result, repeat expansions should be looked for in unresolved illnesses of alleged genetic origin, even in symptoms that do not explicitly elicit a known expansion.

Due to the importance of REs in neurodegenerative diseases, effective strategies for their efficient genotyping are central to research into the aetiology of, and relationships between, these conditions. A central goal of the work detailed in this thesis is therefore to develop an efficient strategy to genotype the known neurological and neurodegenerative REs listed in Table 1-3 so that this strategy can be applied to better understand the role of REs in ALS and FTD. Although whole-genome sequencing using short-read technologies such as Illumina sequencing have become a staple in clinical genetic research, RE genotyping with this type of data is insufficiently precise due to issues associated with processing and alignment of repetitive sequence data. Instead we must rely on polymerase chain reaction (PCR) and capillary electrophoresis to obtain the basepair-level precision necessary for accurate RE genotypes. This is the main objective of this chapter.

Research Aims

This chapter details the work carried out to develop an efficient, cost-effective, highly multiplexed genotyping panel for 35 neurodegenerative repeat expansions using PCR and capillary electrophoresis. As ALS is highly aggressive with low survival rates after diagnosis, DNA samples from patients are a finite resource and should therefore be used as sparingly as possible. Therefore, the project was designed to optimise DNA usage, time, and money by designing a series of PCR protocols with multiple genes (multiplex PCR). This involved first developing optimised PCR protocols for each gene, then arranging genes into multiplex PCR setups. Multiplex PCRs then had to be re-optimised; this involved testing different polymerase protocols, gradients of temperatures and concentrations of reagents, such as magnesium, DMSO, GC-buffer enhancer, primers, and DNA concentration. The final fully optimised panel was used to investigate 35 NDRE genes in a cohort of 1500 ALS patients and 1700 neurologically normal controls from the Irish ALS DNA bank in Chapter 3.

2.2 Methods

Genotyping Assays

To identify the size of each RE for the NDRE genes in Table 1-3, PCR protocols for each gene were extensively tested and optimised. After determining the best PCR protocol for each gene, several primers with similar conditions were grouped into multiplex PCRs. The number of repeats in PCR amplicons was then determined by capillary electrophoresis.

Polymerase Chain Reaction (PCR)

The polymerase chain reaction (PCR) is a process for generating large number of copies of a precise DNA fragment using essential elements of a natural DNA synthesis and replication process. A PCR only requires a few of the components of the replication machinery. The enzymes and proteins necessary in replication to unwind the DNA are replaced by a process called denaturation, which consists of a heating step to break the hydrogen bonds between the bases of the DNA double strand (McPherson and Møller 2006).

Following the template denaturation, two short primer sequences bind to their complementary sequence at the DNA template according to standard base-pairing rules. These primers define the DNA template to be copied and amplified. Deoxynucleotides are then added to the 3'-OH group of both primers by a DNA polymerase, creating a new duplex DNA molecule. At the next heat step, all the double-stranded molecules, now heteroduplexes (original template DNA strand and newly synthesised DNA strand), are denatured. Primers will once again bind to their complementary sequences, and they are ready for the next round of synthesis. In the third cycle, the first exact length of double-stranded PCR product is generated. In the following cycles, the number of copies of the target DNA sequence increases exponentially. Theoretically, it should double its number by each PCR cycle, after 20 cycles a reaction should give around 10^6 new strands. Like other processes, PCR is not 100% efficient, and usually, it needs more rounds of PCR (often 25 to 40 cycles) depending on the concentration and purity of the DNA template and precise conditions that the product requires (Figure 2-1) (McPherson and Møller 2006).

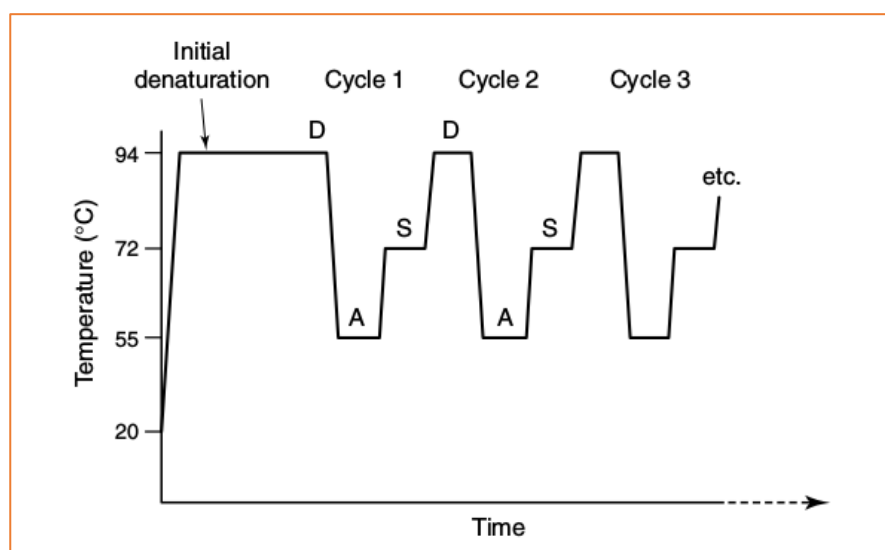


Figure 2-1: Representation of a PCR thermal cycling.

The reaction is heated from room temperature to an initial denaturation phase (D) of around 5 min at 94°C (it varies between enzymes) to ensure the original template strands are now single-stranded. Then follows a series of repeated cycling steps through temperatures for denaturation (D), annealing of primers to template (A) and DNA synthesis from the primer (S) (Taken from (McPherson and Møller 2006).

PCR is controlled using different temperatures for the three steps of the reaction: denaturation, annealing and extension. High temperatures are used to denature the strands of DNA, usually 94-95°C. Then, the temperature is lowered so the primers can anneal (bind) to their complementary sequence strand by base pairing. This temperature varies depending on the oligo and often is referred to as melting temperature (T_m) and is the temperature at which half of the primers anneal to the target sequence. T_m is calculated based on the percentage of CG and AT of the oligos, the more G/C, the higher the T_m would be since their greater hydrogen bonding (McPherson and Møller 2006). There are many formulas to calculate T_m , here is the simplest method for primers up to 20 nucleotides (nts) and another from 15-70 nts:

Up to 20 nts:

$$T_m = ((\text{Number of G} + \text{C}) \times 4^\circ\text{C} + (\text{Number of A} + \text{T}) \times 2^\circ\text{C})$$

Between 15-70 nts:

$T_m = 81.5 + 16.6(\log_{10}(I)) + 0.41 (\% \text{ G} + \text{C}) - (600/N)$, where I is the concentration of monovalent cations and N is the length of the oligonucleotide.

This temperature is crucial to ensure high specificity in the reaction; commonly the greater the T_m , the more specific the reaction will be. Finally, for efficient DNA synthesis, the polymerase needs an optimum temperature. All these three steps are cycled several times to amplify the target DNA sequence (about 25-40 depending on the application) (Figure 2-1).

Optimisation of PCR

PCR amplification is not always a success, and it may be necessary to optimise the conditions. Detailed below are the many factors that come into consideration during PCR optimisation; this project involved iterating over a large number of combinations of temperatures, reagents, enhancers and oligonucleotide primers to find optimal conditions for each gene of interest.

Oligonucleotide design

The selection of oligonucleotides (oligos) is one of the most critical components of PCR. Functional primers are critical for the overall success of the PCR experiment; without them there will be no amplicon. Although the selection of one pair of oligos may seem not so complicated, designing primer for multiplex PCR becomes more challenging (Lowe et al. 1990; Makarova et al. 1992; Pallansch et al. 1990; Rychlik and Rhoads 1989).

Designing oligos is a difficult task since there are many parameters for selecting an optimal primer set. The development of computer software has been a huge improvement in the selection of optimal PCR primers throughout different factors that affect the oligo's function. The balance between specificity and efficiency of amplification are the two basic parameters used when designing primers in PCR. Specificity relates to the frequency with which a mispriming occurs; poor specificity means that the PCR products would have amplicons derived from off-target genomic regions. Efficiency is how well the primer set increases the PCR product with each PCR cycle. There are several parameters to consider for an optimal primer set and to avoid the most common issues of oligo design (Bustin, Mueller, and Nolan 2020; Dieffenbach, Lowe, and Dveksler 1993; Gnffais, André, and Thibon 1991; Lowe et al. 1990; Montpetit et al. 1992; Rodríguez et al. 2015; Rychlik 1993; Rychlik and Rhoads 1989).

Parameters for designing PCR Primers (Bustin et al. 2020; Dieffenbach et al. 1993; Rodríguez et al. 2015):

1. The specificity depends on the primer length and annealing temperature. The shorter the primers are, the more efficient annealing of the targets is. The optimal length for PCR primer is between 18 to 30 bases long.

2. The G and C bases have a stronger hydrogen bonding, which helps with the stability of the primer. Having too many repeating C or G bases can cause primer-dimer formation. The optimal GC content of the primers is between 40 and 60% and ending the 3' primer with a G or C promotes binding (known as GC clamp).

3. T_m is dependent on the primer length. The optimal T_m should be between 65 and 75°C, keeping the difference between the pair lower than 5°C. Therefore, keeping the oligo's length on the short end will improve the T_m . The type of bases on the primers also affects the T_m ; A and T have lower melting temperature than G and C.

4. Secondary structures can be formed on primers. To avoid this, there should be a good balance of G/C and A/T regions throughout the domains.

5. Repetition of 4 or more of the same base, or a dinucleotide repeat (e.g., TGGGG or TATATATA), should be avoided.

6. Check for intra-primer homology or inter-primer homology. Primer homology refers to when there are more than three bases that complement each other within the primer. Inter-primer homology refers to complementary sequences between the forward and reverse primer. Both circumstances will lead to self-dimers or primer-dimers instead of the oligos annealing to the desired DNA sequence.

Coherent and well-inspected primer design is the most critical part of any project to quantify nucleic acids. Each new PCR design must be validated *in silico* to dismiss primer secondary structures (hairpins and primer-dimers) and nonspecific amplicon yields (Bustin et al. 2020; Rodríguez et al. 2015; Rychlik 1993). For this project, the software used for primer design was primer3 (available on <https://primer3.ut.ee>). After, the primers were checked *in silico* by two software: 1) UCSC in-silico PCR website (<https://genome.ucsc.edu/cgi-bin/hgPcr>): this checks specificity of primers on the whole genome and the efficiency of the oligos. 2) Snap gene viewer (Version 6.0.5) for visualizing the PCR product and length. Oligos were synthesised by the company Metabion (<https://www.metabion.com>).

PCR specificity

Not all primers need the same reaction conditions so they can work. In some cases, the 'standard' conditions work efficiently for some oligos with a large amount of product. At the same time, when using other primers with the same DNA template and same reaction conditions, it will give either no product or an intricate pattern of multiple products (McPherson and Møller 2006).

When the PCR is not optimal, some of the parameters of the reaction should be changed to increase or decrease (depending on the situation) the stringency of the PCR conditions appropriately. When no product is detected it is probable that the stringency is too high, and if there are multiple bands it needs to be increased (McPherson and Møller 2006).

The parameters that influence the specificity of the reaction are the T_m , cycling regime and the buffer composition. High specificity in PCR is favoured by:

- Optimal concentrations of magnesium, primers, deoxyribonucleotides triphosphate (dNTPs), other ions and DNA polymerase.
- High annealing temperatures, efficient denaturation and fast ramping rates.
- Length and number of cycles.
- PCR additives.

Magnesium, dNTPs and DNA polymerase

Magnesium ions (Mg^{2+}) are the most critical components of the PCR; their concentration affects the specificity and efficiency of the reaction. Taq polymerase depends on free Mg^{2+} ions, and the highest activity is shown when the concentration of these free ions are between 1.2-1.3 mM. Free Mg^{2+} depends on or is affected by the concentration of dNTPs. There is an equimolar binding between dNTPs and Mg^{2+} . For example, 200 μM of each dNTPs are equivalent to 800 μM in total, and if we add $MgCl_2$ with a concentration of 1.5 μM ; the result free Mg^{2+} is 700 μM which is less than the optimal concentration for the Taq polymerase, meaning that altering the concentration of $MgCl_2$ is modifying the concentrations of dNTPs in the reaction. If the concentration of Mg^{2+} is too low this leads to reduced yields, while excess can reduce the fidelity of Taq DNA polymerase and amplification of non-specific PCR products. Hence, altering the

concentration of MgCl_2 and dNTPs can lead to one primer/template pair performing significantly different from another reaction under identical conditions (McPherson and Møller 2006).

Each dNTP is made up of a nitrogenous base, deoxyribose sugar and a phosphate group. There are four different nucleotides: dATP, dTTP, dGTP and dCTP. PCR protocols suggest that 200 μM of each dNTP is enough to synthesise around 10 μg of PCR product. The fidelity of the Taq DNA polymerase is affected by the concentration of dNTPs in the reaction. If its concentration is low, the efficiency of the reaction is affected, while if the concentration is high the fidelity of the process will adversely be affected by driving the Taq DNA polymerase to misincorporate bases at a greater rate than normal (McPherson and Møller 2006).

Different modified nucleotides can be incorporated during PCR into the amplicons for many purposes, including the resolution of secondary structures. These can be done by adding 7 deaza-GTP, which is a modified nucleotide that, when added to the sequence, reduces the formation of secondary structures in GC-rich regions of the DNA, improving the PCR or sequencing (McPherson and Møller 2006).

When there is no PCR product or faint bands, it could mean that the concentration of Taq DNA polymerase is too low. The way to investigate this is by doing a gradient of concentrations of the polymerase and see which concentration fits the best for that reaction. It is important to note that thermostable proofreading polymerases can have a lower processivity than Taq DNA polymerases, meaning that the yield is affected, so more enzyme may be needed for successful amplification. Also, thermostable polymerases are inactive at high temperatures leading to a low or reduced level of PCR product. Hence it is essential to use a short period for denaturation at temperatures higher than 90°C (McPherson and Møller 2006).

Temperatures

- Denaturation

To make template DNA sequence available for PCR, double-stranded DNA must first be denatured to single-stranded DNA. An initial phase achieves this by usually five minutes

denaturation around 94°C. If this step is inefficient, the DNA will not be single-stranded, and this will restrict primers to bind into their complementary sequence and the DNA extension. For templates with a higher GC percentage, it might be necessary to increase the temperature perhaps to 96°C and the amount of time will depend on the polymerase to be used. At the beginning of each cycle, there is a denaturation step that will denature the future PCR products for the subsequent reaction. Times and temperatures vary among the DNA polymerases, so it is essential to read and try their recommended temperatures for PCR cycles (McPherson and Møller 2006).

- Annealing

PCR relies on the specificity with which a primer anneals only to its target sequence, so optimisation of this step is vital. The T_m is the critical temperature for a successful PCR; this lets the primer bind correctly to its complementary sequence, or if the temperature is not correct it will mismatch to untargeted sequences. Generally, the higher the T_m , the more specific the annealing of the primer to its target sequence. The lower the temperature, more mismatches of the primer to the template leading to amplification of untargeted sequences (McPherson and Møller 2006).

Cycle number and length

Generally, the number of cycles of a PCR must be kept to the minimum required to generate enough product for further analysis. This will reduce the number of errors and the accumulation of nonspecific products. A standard protocol has around 20 cycles; if it does not generate enough product, adding more DNA template should increase the product or if not, the number of cycles may be increased by five at 25, 30, 35 and 40 cycles until reaching the appropriate amount of PCR product (McPherson and Møller 2006).

The specificity of the PCR can also be influenced by the time it takes to change between temperatures during PCR cycling. Usually, when the ramp rate is faster, the specificity is higher, and the reaction is completed faster.

PCR additives

Many different polymerase kits contain buffers and additives to optimise the conditions of a PCR. Various "enhancer" mixtures can improve the efficiency or specificity of the PCR. The majority are chemicals that may increase the effectiveness of annealing temperature and DNA binding proteins. These additives can be added to the PCR reaction to reduce mismatch and enhance primer annealing. Some additives can also disrupt base pairing improving difficult templates such as GC-rich sequences and may improve the specificity of primer annealing (McPherson and Møller 2006). Some of the most common additives used to improve PCR reaction are:

- Dimethyl sulfoxide (DMSO): disrupts base pairing and is usually added up to 5-10% (v/v)
- Betaine (N,N,N-trimethylglycine): equalises the contributions of AT and GC base pairs towards duplex stability.

Multiplex Polymerase Chain Reaction

The design of two or more set of oligos for amplification of different targets in the same reaction mixture is called multiplex PCR (Chamberlain et al. 1988; McPherson and Pincus 2021; Shen 2019). Using this technique, it is possible to co-amplify more than one target in a single tube from the same DNA or specimen by adding fluorescent labels to the primers. Multiplex PCRs are more complicated to optimize (McPherson and Pincus 2021). For a successful multiplex PCR, the most important dependent variable is the design of the primers; they must have similar annealing temperatures. Oligos must be noncomplementary between each other to prevent primer-dimers and to avoid inefficient PCR reactions (Bustin et al. 2020). The products of the various targets must have distinctive sizes. This is necessary to detect them and separate them by gel electrophoresis or capillary electrophoresis (Chamberlain et al. 1988; McPherson and Pincus 2021; Shen 2019) (

Figure 2-2). The use of multiplex PCR can reduce the cost, time, and amount of DNA per specimen. This technique is more complicated to develop and usually less sensitive than single-primer-set PCR (Shen 2019).

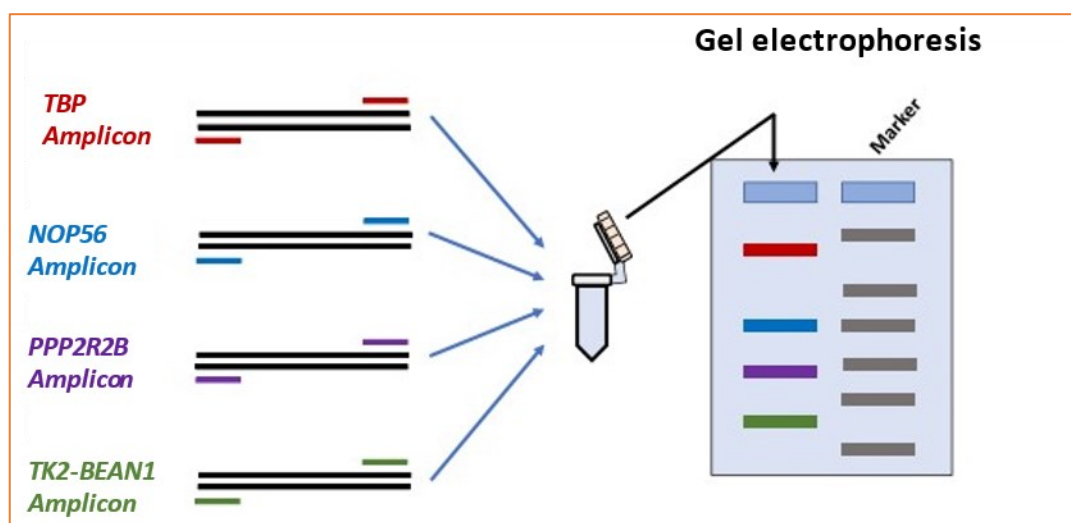


Figure 2-2 Multiplex PCR overview

Representative figure of a multiplex PCR electrophoresis. Each gene primer is represented in a distinctive colour. After the PCR is complete, products are visualized in an agarose gel by gel electrophoresis. Each amplicon has a specific length represented by their colour (Taken and modified from Shen, 2019).

Repeat Primed PCR (rpPCR)

Repeat primed PCR (rpPCR) is a type of PCR in which the reverse primer binds at different points of the repeat expansion to produce multiple amplicons of incrementally greater size (Ciotti et al. 2004; Musambil et al. 2014; Radvansky, Ficek, and Kadasi 2011). The lower concentration of the reverse primer in the reaction means that it will be fully consumed during early PCR cycles, after which a third primer (the anchor primer), which is complementary to the 5' end of the reverse primer (Figure 2-3), is preferentially used as the reverse primer (Ciotti et al. 2004; Musambil et al. 2014; Renton et al. 2011; Warner et al. 1996). Specificity is dictated by a fluorescently labelled forward primer, with the fluorophore usually conjugated to the 5' end. On completion of the rpPCR, an aliquot of the amplicons is sized by capillary electrophoresis; unless otherwise stated, for the majority of the PCR products in this study, capillary electrophoresis was carried out by Eurofins Medigenomix GmbH (Ebersberg, Germany).

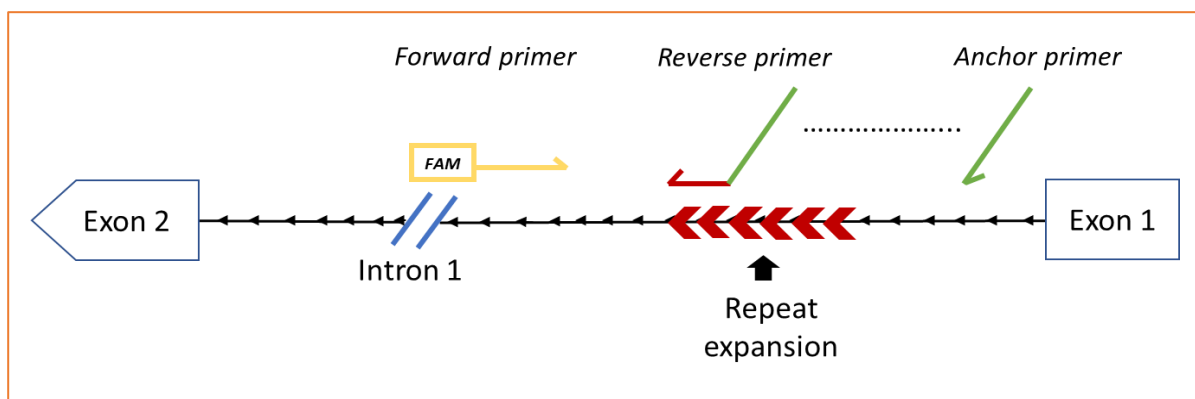


Figure 2-3 Graphical representation for rpPCR analysis of the primer binding.

This illustrates the use of rpPCR in genotyping the *C9orf72* hexanucleotide repeat expansion in intron 1 of the gene using a three-primer system. (Taken and modified from Renton *et al.*, 2011).

Electrophoresis

Agarose gel electrophoresis

Agarose gel electrophoresis is the most widely-adopted method for separating different sizes of DNA ranging from 100 bp to 25 kb (Lee et al. 2012; Sambrook et al. 2001). The agarose comes from seaweed genera *Gelidium* and *Gracilaria* and contains repeated subunits of agarobiose (L- and D-galactose) (Kirkpatrick 1990; Lee et al. 2012). During the process of gelation, the agarose polymers associate non-covalently forming a network of bundles with pore sizes. The sizes of the pores decrease with the increase of agarose in the mix. DNA molecules have a uniform mass/charge ratio. This characteristic allows these molecules to be separated by size within an agarose gel following a simple rule: the travel distance of a DNA molecule is inversely proportional to the log of its molecular weight. The migration rate of the DNA fragments through the agarose gel is determined by agarose concentration, DNA conformation, DNA size, voltage applied, type of agarose, ethidium bromide or any other dye, and the electrophoresis buffer (Helling, Goodman, and Boyer 1974; Lee et al. 2012; Smith, Aldridge, and Callis 1989). Once the DNA fragments have been separated on the gel and suitably stained, they can be visualized under ultraviolet (UV) light (Figure 2-4).

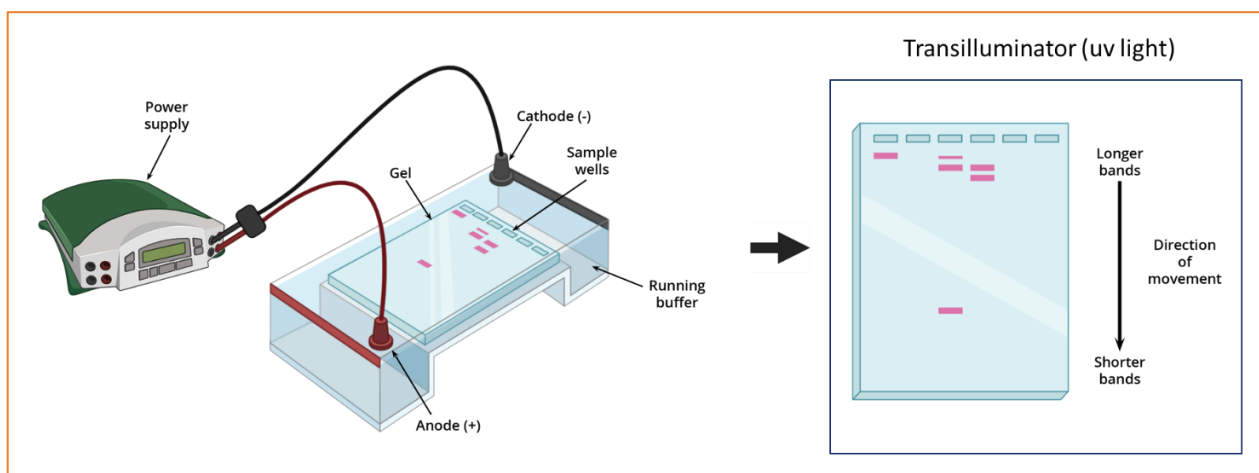


Figure 2-4 Agarose gel electrophoresis workflow.

A gel box container is filled with a 1X Tris-EDTA (TE) buffer (running buffer). This box has a negative terminal (Cathode) fixed in one side, and a positive terminal (Anode) on the opposite side. The gel is made with the running buffer and is allowed to set. It is then introduced to the box with the buffer and using a pipette, the samples are loaded (previously mixed with loading dye) into the sample wells. Afterwards, we attach the cathode and anode to the power supply and apply a voltage across the gel (usually 100 mV) for 20-25 minutes (depending on the gel concentration). This voltage causes charged molecules (DNA fragments) to move through the gel towards the opposite terminal (from cathode to anode). Different molecule sizes travel through the matrix in different rates. Larger fragments travel slower than smaller ones, with the shortest molecules travelling the furthest. To visualize the fragments in the gel, this one is exposed to uv light on a transilluminator (Lee *et al.*, 2012). (Image taken and modified from (Gel Electrophoresis | AAT Bioquest, n.d.)).

In this project, agarose gel electrophoresis is used to run and verify the length of each PCR product, visualized in a E-Box Gel transilluminator (Vilber Lourmat). Usually, an agarose gel between 2 and 3 % with ethidium bromide is used to observe the amplicons and choose the best protocol of PCR for each gene. After the PCR optimisation, each gene amplicon was also sequenced using Sanger sequencing to confirm product specificity and expansion size.

Capillary electrophoresis- Fragment Length Analysis (FLA)

Capillary electrophoresis (CE) is a much more powerful tool for separating DNA fragment than the conventional gel electrophoresis. CE technology has many advantages when separating fragments, such as higher speeds and better resolution, sensitivity and data handling (Aaij and Borst 1972; McCord, Hartzell-Baguley, and King 2008; Righetti and Gelfi 1999). CE separation occurs inside a small diameter (50- to 100- μ m) quartz capillary in the presence of a high voltage (kilovolt-level). While the sample is running through the capillary, the DNA fragments are detected either by UV absorption or by a fluorescent labelling (usually added to the primers) activated by a laser in the detector (Righetti and Gelfi 1999; Smith and Nelson 2003) (Figure

2-5). This eliminates the need of mutagenic substances like Ethidium bromide (Smith and Nelson 2003). For the DNA separation, the amount required of DNA is the nanogram range. The resolution of this technique goes from a single base pair (bp) to several hundred of base pairs (bps) in size (Landers 1996; McCord et al. 2008; Smith and Nelson 2003). There are many different standard sizing markers that are used for CE depending on the different requirements of the sample and electrophoretic mobility. These markers are added to the sample mix before loading it to the capillary (Smith and Nelson 2003). The oligos should have a different dye to label each PCR product so it can be detected, visualized, and separated from the multiple targets. Usually, the number of different colours on a single PCR tube is between 2 to 5 (Smith et al. 1989; Smith and Nelson 2003).

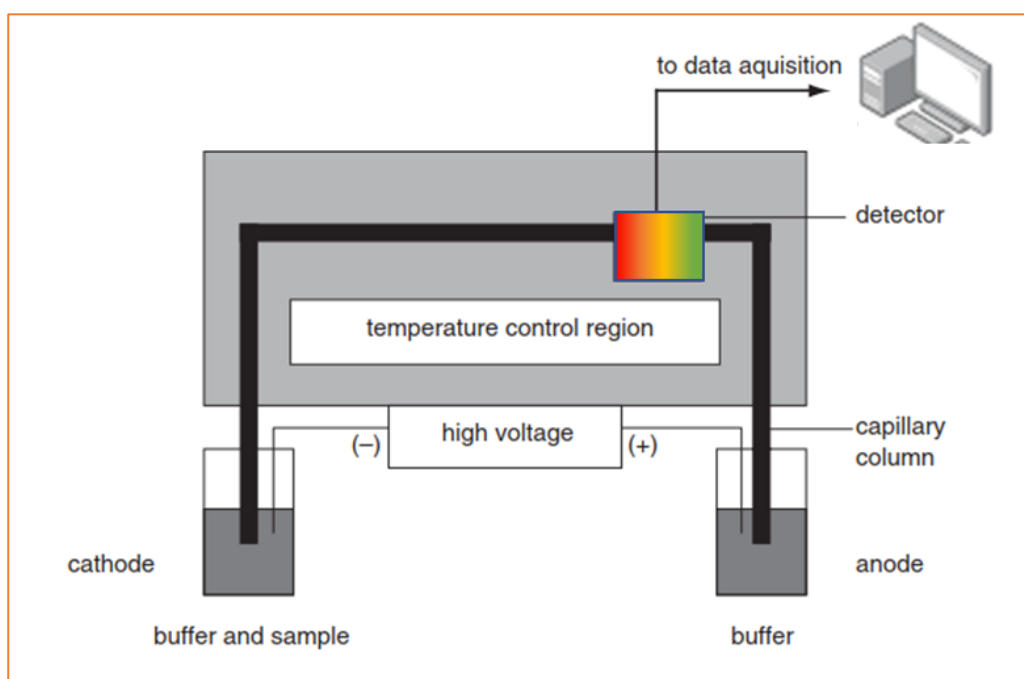


Figure 2-5 Schematic configuration of a CE instrument for DNA separation.

(Taken and modified from Smith & Nelson, 2003).

For each repeat expansion on this project, alleles were determined by capillary electrophoresis, either at Eurofins Medigenomix GmbH (Ebersberg, Germany), or, during protocol development, using an ABI SeqStudio™ Genetic Analyzer System with SmartStart (Applied Biosystems™). Sizing was facilitated by fluorophores on the 5' ends of the forward primers, the choice of which depended on how many genes are in the multiplex PCR. For example, if in

one PCR we have four genes (4 different sets of primers), there will be four different fluorophores of similar spectrum so the laser of the capillary electrophoresis can identify it in one run, plus the size standard (which has its own label) (Figure 2-6).

The resulting electropherogram from the CE run has fragment lengths automatically called in bp by in-house software at Eurofins Medigenomix GmbH Gene Mapper (Figure 2-7) or using GeneMapper v5 with SeqStudio data (Figure 2-8). Sample results are provided as FSA files, which were visualised using Peak Scanner v1.0 software. Using the fragment length analysis (FLA) sizing data, each allele of all NDRE genes was determined using the following general Equation 2-1:

Equation 2-1

$$\# \text{ of repeats} = \text{common repeat count} + \left(\frac{\text{Allele length (bp)} - \text{common repeat length (bp)}}{\text{repeat size (bp)}} \right)$$

Common repeat count: number of repeats present in the reference genome; Common repeat length: size of the in-silico PCR from the reference genome; Allele length: PCR fragment length obtained by CE; Repeat size: repeat motif length.

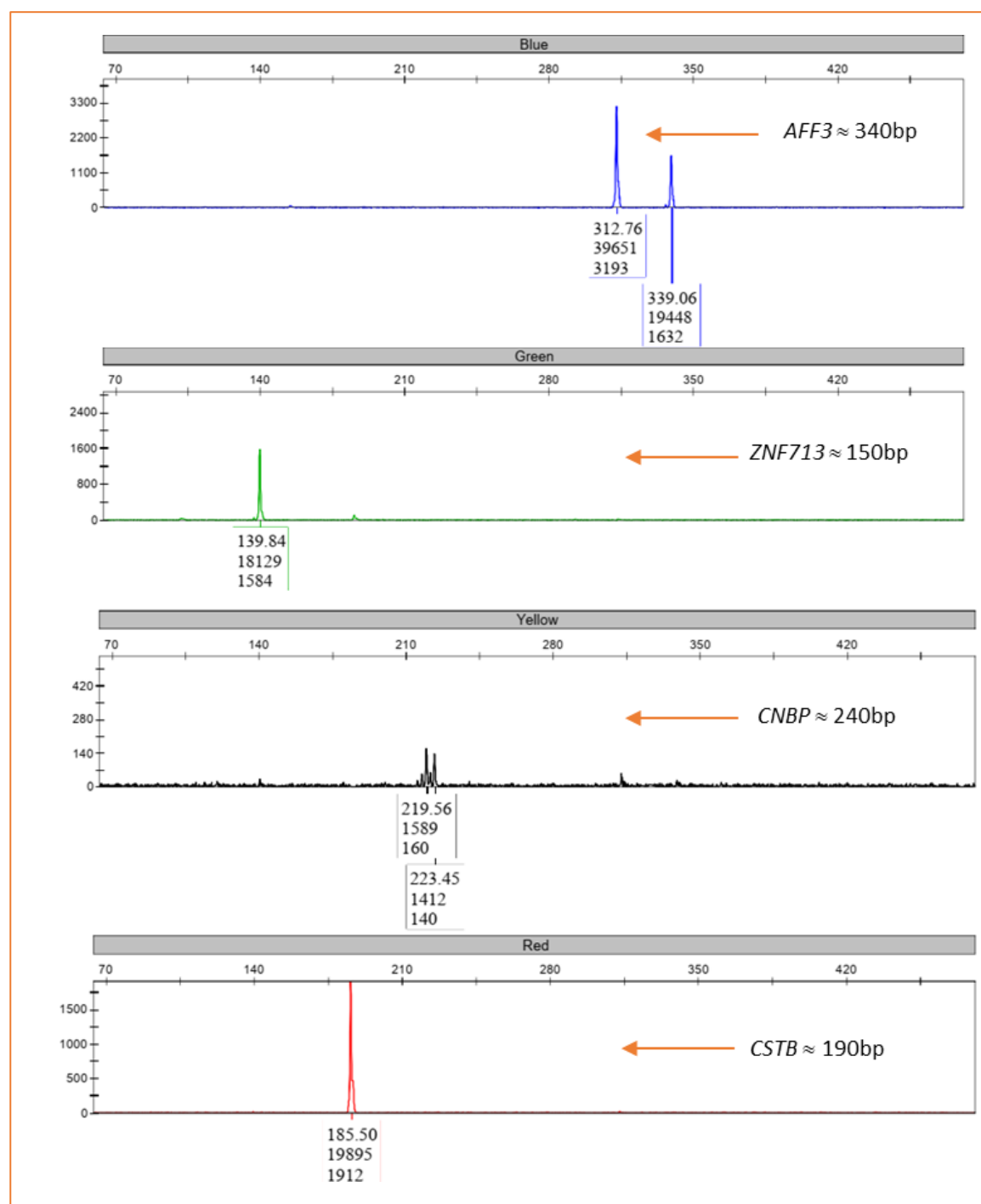


Figure 2-6: Electropherogram of a multiplex PCR.

Example of an electropherogram showing the amplicons of a multiplex PCR resolved by capillary electrophoresis. (Taken and modified from Sobral *et al.*, 2012).

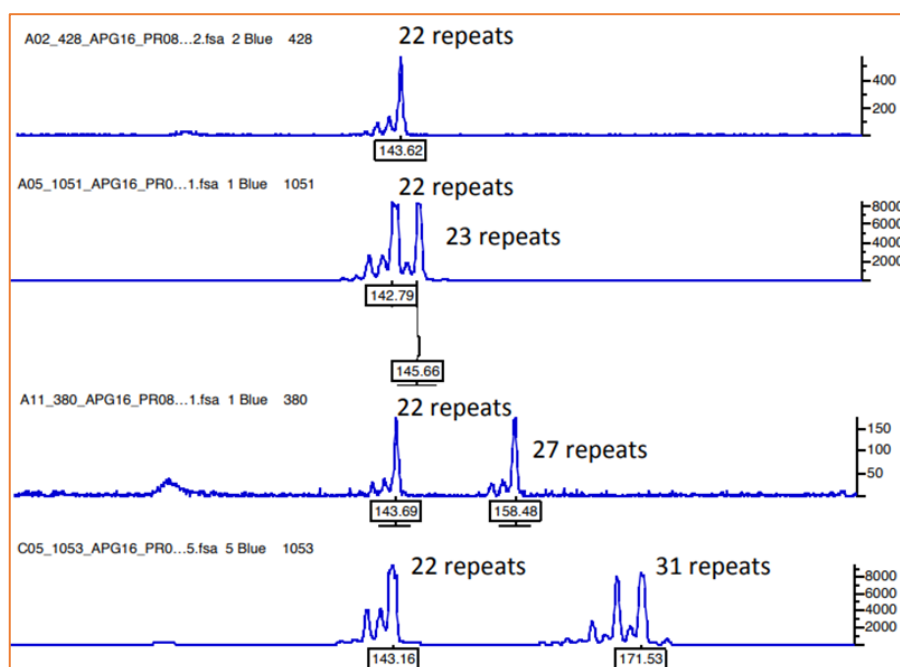


Figure 2-7: *ATXN2* gene electropherogram.

Example of an electropherogram of *ATXN2* alleles for representative Irish ALS patients of each repeat expansion with their respective length (bp).

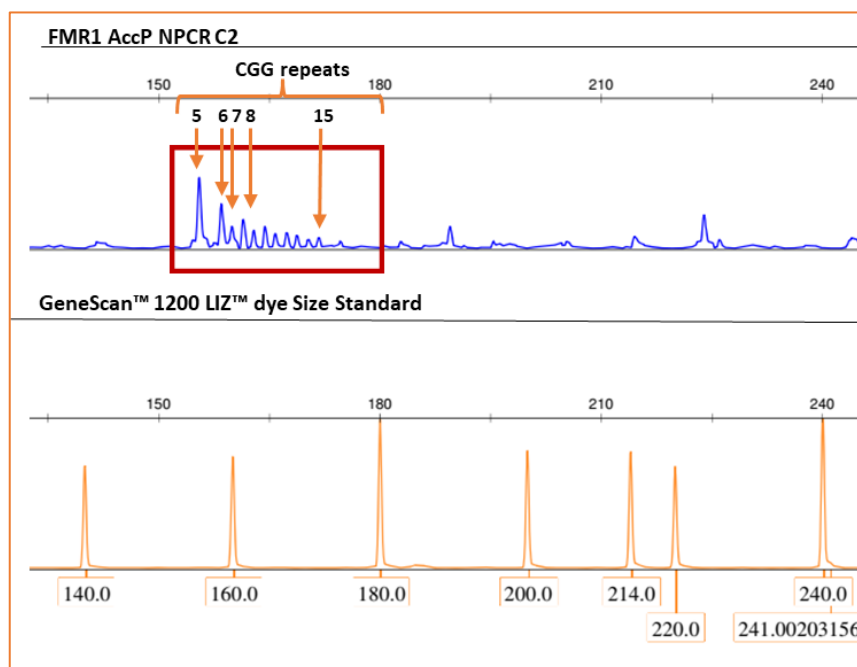


Figure 2-8 *FMR1* rpPCR electropherogram obtained from the SeqStudio™.

Example of an electropherogram of the rpPCR of the *FMR1* CGG repeats of a gDNA control (C2), their respective RE peak counts (upper panel) and the respective size ladder 1200LIZ in bp (lower panel).

Sequencing

After the optimization of each individual set of primers for each gene (Supplementary Table 2-2), samples were sequenced by Eurofins Genomics using their Sanger sequencing service (<https://eurofinsgenomics.eu/en/custom-dna-sequencing/eurofins-services>) to ensure that the amplicons represented target sequence and were sized correctly. The software Snap gene viewer (Version 6.0.5) was used to visualise the sequencing results. Both sequencing and *in-silico* results of individual PCRs were compared and corroborated with each gene amplicon. The same analysis was done after the optimization of the all the multiplex PCRs. During the process of FLA, the different labels have different molecular properties that may change the real length of the amplicon by some bp in a gene-specific manner. Therefore, the formula used to calculate the number of repeats for each gene amplicon was adjusted to reflect deviation of observed sequence-based sizing from our FLA results. To achieve this calibration, we selected two or three homozygotes for the repeat expansion for sequencing. This determined the correct number of repeats for each individual and compared with the PCR results.

Study design

The main steps taken in this study to an optimised multiplex PCR were:

1. Research in the literature all the different RE genes. Extract all the information available for those genes and possible PCR and rpPCR protocols, including the primers. The most important information was the polymerase used, enhancers, primer sequences, oligos concentration and thermal cycles.
2. Design of multiple primers for each gene using primer3 website tool.
3. Optimization individual pairs of primers for each gene. The several primer pairs were key to get different T_m of each gene and different amplicon sizes which would help build a multiplex PCR. For this project, around four to five genes at a time were chosen to be optimized at the same time.
4. Choosing of similar T_m primers with distinctive amplicon sizes up to four different genes and multiplex them on the same PCR. Usually, pairing primers was a good strategy and from there keep adding other optimized genes.

5. After the optimization of the PCR, the amplicons involved were sequenced, both individually and multiplexed. This step confirmed that each amplicon corresponded to the expected in-silico PCR product for each RE gene.

6. After the optimization of each set of multiplex PCRs they were ready for FLA. First, based on the length of the amplicons on the multiplex PCR, one of the four available fluorescent labels was assigned to the 5' end of the forward (F) primer and sent to Metabion for synthesis. At that point, each PCR was executed with the new labelled primers in 4 different controls. The PCR products were sent to Eurofins for FLA. The results of this trial were assessed to ensure that PCR products were showing and they could be differentiated without overlapping or noise.

7. Once all fluorescent labels were correctly assigned and the FLA protocol was selected, a pilot experiment was designed for a 96 well plate for each multiplex PCR. Products were sent to Eurofins and the results were analysed. This confirmed that the experiment and conditions were correct, and could be scaled up to a 96 well plate, which was the format subsequently used for genotyping 1500 ALS patients and 1600 controls.

8. After receiving and analysing results from each gene from different patients and controls, bp sizing was converted to repeat count using Equation 2-1. At least three different samples that appeared to be homozygotes for their specific repeat was chosen to be sequenced; resulting sequence-level data were used to confirm correct sizing and to calibrate FLA results, if necessary.

Overall, these steps summarize how the optimization was carried out. One of the most important parts of the PCR optimization are primers and DNA polymerase. To find a polymerase that can properly amplify REs is challenging. During this project, at least 5 different enzymes were used for different genes and primers to get a good quality PCR product without unspecific amplicons.

Following PCR optimisation, calibration, quality control and final multiplexed PCR panel design, more than 1,500 ALS patients and over 1,700 age- and sex-matched healthy control subjects were genotyped in the following chapter to assess the pleiotropic role of expanded alleles in ALS and FTD risk.

2.3 Results

The final panel consisted of 35 genes in 16 multiplexed PCRs. In this section, two representative examples are provided for the process of optimising individual and multiplex PCRs of 8 genes, along with their sequencing and FLA results. It should be noted that these two examples have been chosen for brevity in this thesis chapter, and the workflows and challenges described are representative of all 16 multiplex PCR setups ultimately optimised to genotype all 35 genes. All final optimised PCR conditions and primers used for each gene can be found in Supplementary Table S 2-1, Supplementary Table S 2-2 and Supplementary Table S 2-3. Supplementary Figures Chapter 2 contains the results of the optimization of the rest of the 15 groups of genes. Among these results are the agarose gels of the optimized PCRs, sequencing, and FLA results of each gene (individual and multiplexed).

PCR optimization

Multiplex 1: HTT, JPH3, ATXN7 and ATXN10 genes

Optimization of HTT and JPH3 genes

The detection of triplet repeats expansion associated with *HD* and *HDL2* diseases was performed by following the protocol from Krause *et al.* (2015). *HD* has a CAG expansion and *JPH3* is a CTG expansion. The primers used were **HU4 F:** 5'-ATGGCGACCCTGGAAAAGCTGATGAA-3' and **HDC2 R:** 5'-CGGCGGTGGCGGCTGTTG-3' for *HTT*; and for *JPH3* **L237-1F:** 5'-GGTTCCTGCACAGAAACCATC-3' and **L237-2R:** 5'-AGATGCCACCGCATTCGG-3' (Krause et al. 2015).

The first PCR was to amplify individually each gene to find the optimum T_m , concentrations of reagents and to verify the length of the amplicons. A gradient of temperatures was performed for each pair of primers and the temperatures ranged from 60-71.2°C and the polymerase used as Q5™ Hot Start High Fidelity DNA Polymerase (New England BioLabs) with/without 1X GC enhancer (provided with the enzyme) and two different gDNA control samples. After running the products in 2.5% agarose gel, the primers amplified unspecific regions of the genomic DNA,

the expected amplicons for each gene were ≈ 126 bp for *HTT* and ≈ 117 bp for *JPH3* (Figure 2-9-A).

Different gradients of $MgCl_2$, DMSO and temperatures were attempted in different PCRs but the primer still amplified unspecific regions, or there were no amplicons (Figure 2-9-B). Different polymerases were used following their specific protocol: KAPA HiFi hot start DNA polymerase, Taq DNA Polymerase (Thermo Scientific), Platinum® Taq DNA polymerase High Fidelity (Life Technologies™) and AccuPrime™ GC-Rich DNA Polymerase (Invitrogen™). For all of them, gradient of temperatures and DMSO were applied but still not a clean PCR with the desire amplicons.

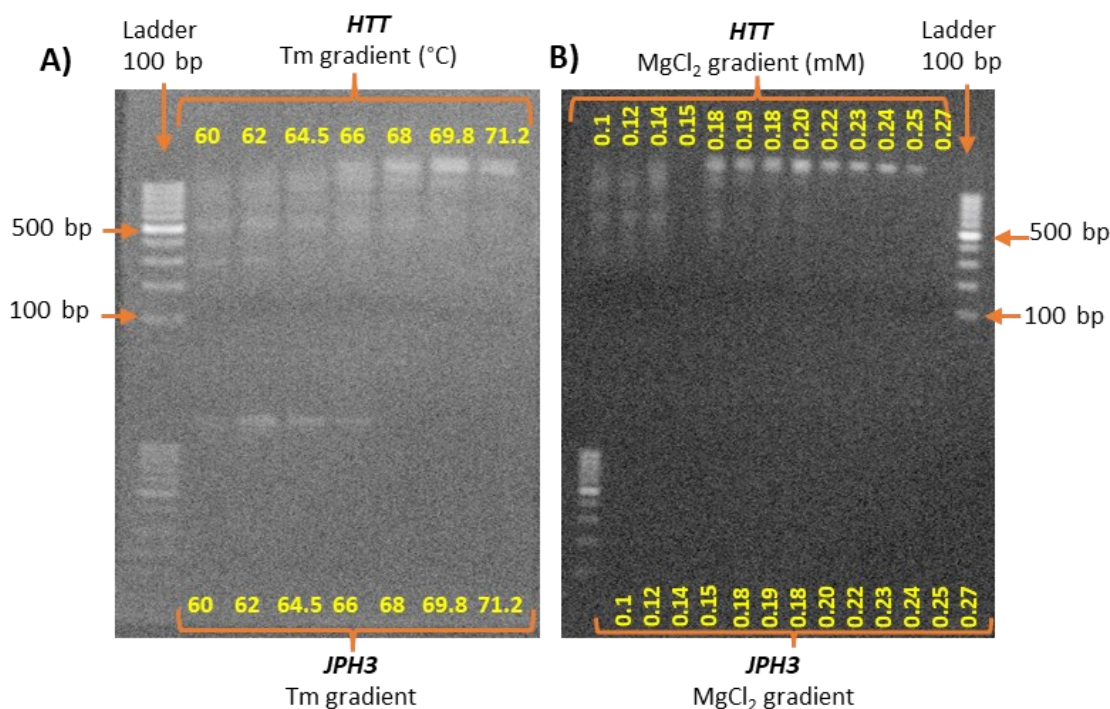


Figure 2-9 Photographic recording of the electrophoretic run on 2.5% agarose gel of the PCR products from *HTT* and *JPH3* primers using Taq DNA polymerase High Fidelity (Life Technologies™).

A) For this experiment, a gradient of temperature between 60 to 71.2°C for both genes. **B)** Gradient of magnesium concentration (mM) for both genes.

A new enzyme, Hot FirePol® DNA Polymerase (Solis Biodyne), was used using a gradient of temperatures and with and without a GC enhancer called solution B at 3%. This enzyme is a

chemically modified DNA polymerase which is activated after 15 minutes incubation at 95°C; preventing the extension of unspecifically annealed primers and the formation of primer dimers. The enzyme has a 5' → 3' polymerization-dependent exonuclease replacement activity and it lacks 3' → 5' exonuclease activity. For each gene, the desired amplicon was amplified and without any unspecific bands (see Figure 2-10). The next step was to multiplex both genes as Krause *et al.* did in their protocol, but it was not possible to differentiate the two amplicons in the three gDNA controls used by agarose electrophoresis.

Therefore, new primers were designed using primer3 online tool (<http://primer3.ut.ee/>) for both genes to get different lengths with similar T_m so they could be multiplexed. Three sets of primers for *HTT* and two sets for *JPH3* were ordered and different PCRs were designed to optimize it using Hot FirePol® DNA Polymerase. For each set of primers 2 different control DNAs were used, 2 different temperatures and with solution S (GC enhancer) of 0X and 3X for each reaction. After running all the product in a 2.5% agarose gel, all the primers amplified the expected amplicons (see Figure 2-11). Primer **HTT-2** (HTT 2F: 5'-CATGGCGACCCTGGAAAAG-3', HTT 2R: 5'-GGGCCCAAACCTCACGGTC-3'; ≈288bp) and **JPH3-1** (JPH3 1F: 5'-CTTTCAGTGGCTGCATCGTT-3', JPH3 1R: 5'-ACTGAGGAGTGGATATCGGAG-3'; ≈233bp) were selected for multiplex PCR and the conditions used were: T_m at 62°C, 3X Solution S and final concentration of each primer at 0.25 μM.

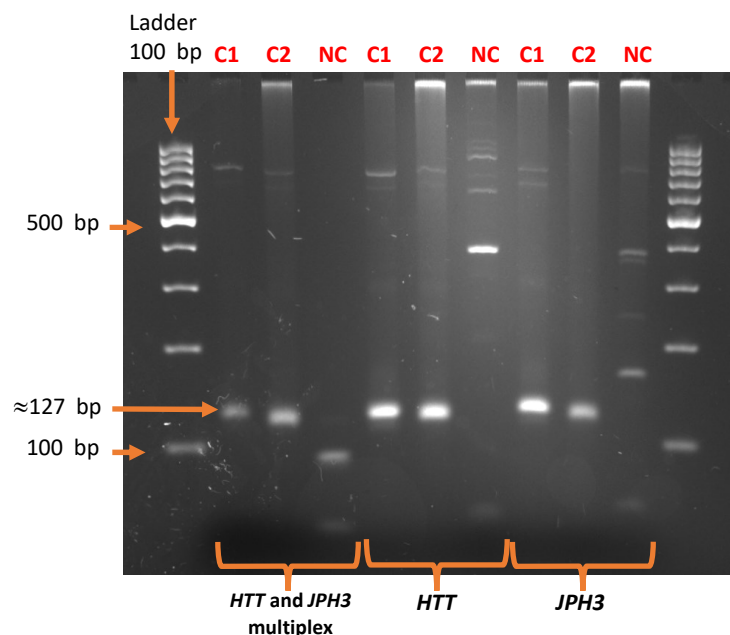


Figure 2-10 Photographic recording of the electrophoretic run on 2.5% agarose gel of the PCR products from *HTT* and *JPH3* primers.

This gel shows two different controls DNA (C1 and C2; NC: Control of PCR without DNA), the amplicons for the multiplex PCR and the individual amplicons for *HTT* and *JPH3* reactions. On this figure the NC has some PCR products. These amplicons are an artefact of primer dimers, in which high concentrations of the primers with high CG can be attached to each other and get amplified during PCR and/or during gel electrophoresis they move more slowly than normal primer dimers.

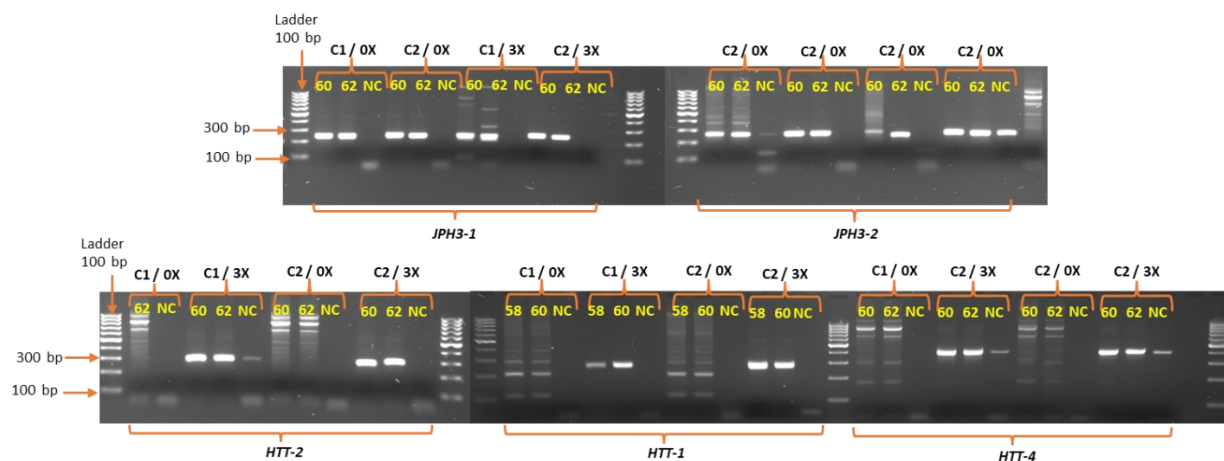


Figure 2-11 Photographic recording of the electrophoretic run on 2.5% agarose gel of the PCR products from *HTT* and *JPH3* new set of primers.

For each set of primers, two different gDNA controls (C1 and C2, NC: Control of PCR without DNA), two temperatures were tested for each set of primers as well as two concentrations of solution S (0 and 3X). On this figure the NC has some PCR products. These amplicons are an artifact of primer dimers, in which high concentrations of the primers with high CG can be attached to each other and get amplified during PCR and/or during gel electrophoresis they move more slowly than normal primers.

Optimization of *ATXN7* and *ATXN10* genes

The detection of fragments containing the REs of *ATXN7*, *ATXN10* and *FXN* genes were amplified using the following set of oligonucleotides:

- *ATXN7* primers (Katagiri et al. 2015), ≈293 bp containing 10 CAG repeats:
ATXN7 F: 5'-TTGTTACATTGTAGGAGCGGAA-3';
ATXN7 2F: 5'-GTGTTACATTGTAGGAGCGGAA-3'
ATXN7 R: 5'-CTTCAGGACTGGGCAGAGG-3'
- *ATXN10* primers (Wang et al. 2008), ≈197 bp containing 14 ATTCT repeats:
ATXN10 F: 5'-AGAAAACAGATGGCAGAATGA-3'
ATXN10 R: 5'-GCCTGGGCAACATAGAGAGA-3'

Initially, *ATXN7* and *ATXN10* primers had similar PCR protocols. It was decided to use a Q5™ Hot Start High Fidelity DNA Polymerase following the recommended PCR cycles, a gradient of Tm between 50-62°C, two different controls gDNA and with/without 1X GC enhancer. For *ATXN7* the set of primers used were ATXN7 F/R. After the electrophoresis in 2.5% agarose gel of the PCR products, for both genes there was unspecific amplifications of the primers for all the different conditions (Figure 2-12). Hence, for both genes different PCRs were carried out using the same enzyme, changing the annealing temperatures and gradients of MgCl₂, as well as two different concentrations of DMSO (2 and 3%). For *ATXN7* two sets of primers were tried (ATXN7 F/R and ATXN7 2F/2R). After running the PCR products, for *ATXN7* the desired amplicon of 293bp was amplified but still with unspecific amplicons, and for *ATXN10* only unspecific bands were obtained. Many other PCRs were designed using Taq DNA Polymerase (Thermo Scientific) and Platinum® Taq DNA Polymerase High Fidelity (Life™ Technologies) and still no clean amplicons were obtained.

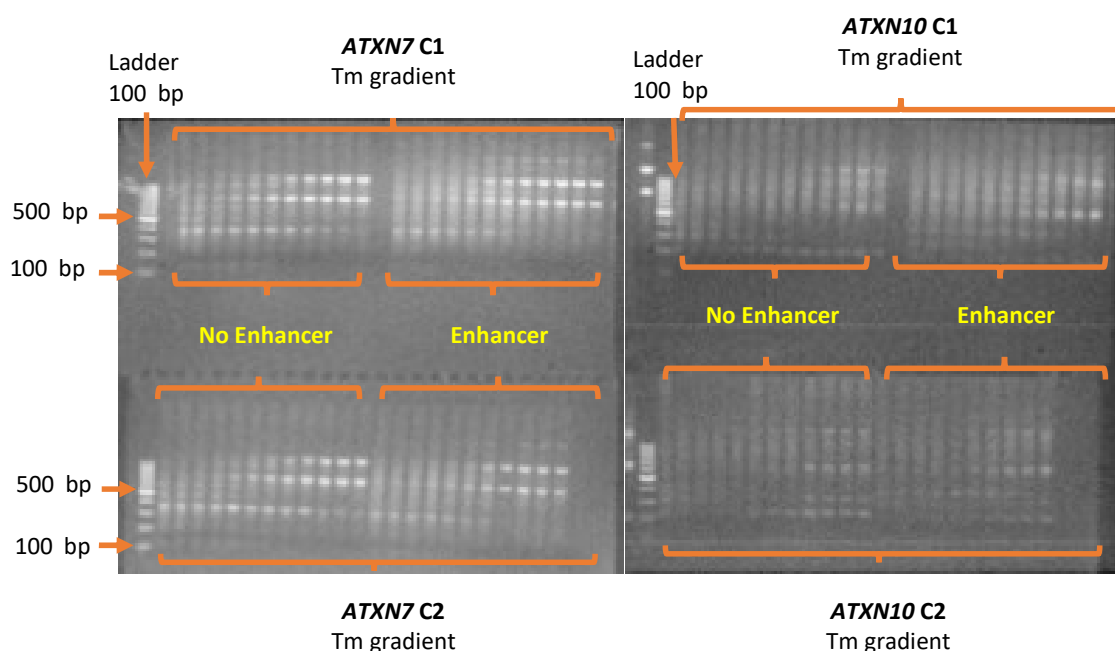


Figure 2-12 Photographic recording of the electrophoretic run on agarose gel 2.5% of the temperature gradient PCR products of *ATXN7* and *ATXN10*.

Temperature gradient PCR for *ATXN7* and *ATXN10* using two different gDNA controls (C1 and C2).

After several attempts, a new PCR was designed for both genes using Hot FirePol® DNA Polymerase, buffer B1 and B2, Tm for *ATXN7* of 63°C (Primer ATXN7 2F/R), and for *ATXN10* 59 and 60°C, a gradient of solution S (0X, 1X, 2X, 3X) and only one control gDNA. For *ATXN7*, it was possible to obtain the desired amplicon of 293 bp using this enzyme with buffer 2, Tm of 63°C and solution S at 3X (Figure 2-13 A). This protocol was done one more time with two different controls gDNA and the PCR was successful (Figure 2-13 B). For *ATXN10* it was not possible to amplify the expected amplicon of 1388 bp without any unspecific products (Figure 2-13 A).

Since different protocols for optimization and many enzymes were used for *ATXN10*, two new sets of primers were designed for the gene using the online tool primer3. Each set of primers were used in PCR using two different Tm temperatures (60 and 62°C), Hot FirePol® DNA Polymerase, buffer 2, solution S at 0X and 3X and two different gDNA controls. After running the PCR products on agarose gels, for *ATXN10* set 2F/2R with the amplicon of 188 bp was amplified successfully with a Tm 62°C, buffer 2, 3X solution S (Figure 2.16).

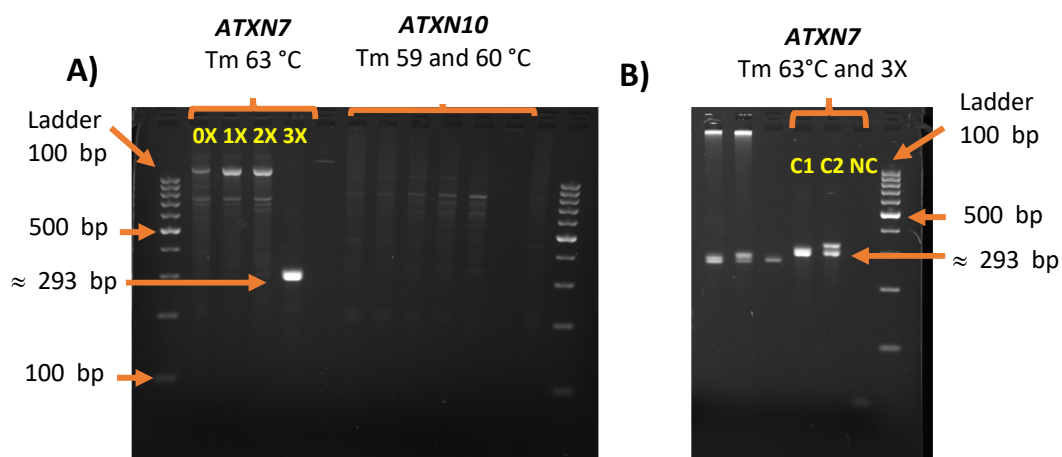


Figure 2-13 Photographic recording of the electrophoretic run on agarose gel 2.5% of the PCR products of *ATXN7* and *ATXN10*.

A) PCR products for *ATXN7* using a T_m of 63°C and 4 different concentrations of solution S (GC enhancer), expected amplicon of ≈293 bp. Temperature gradient PCR of *ATXN10*. **B)** PCR products of *ATXN10* using 2 gDNA controls (C1,C2 and NC: control of the PCR).

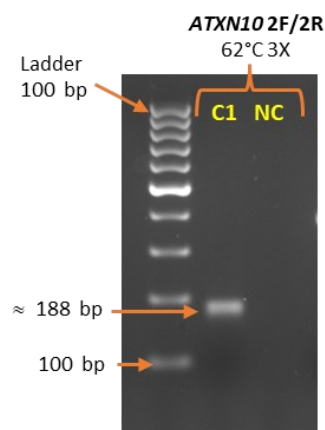


Figure 2-14 Photographic recording of the electrophoretic run on agarose gel 2.5% of the PCR product of *ATXN10* 2F/2R.

Expected PCR product for *ATXN10* of ≈188 bp (C1: gDNA control; NC: Control of PCR without DNA)

Multiplex PCR optimization

After optimising the PCRs for the four genes, all genes had similar PCR conditions when using Hot FirePol® DNA Polymerase, and the same PCR cycles. Hence, it was considered possible to multiplex them in one reaction. A multiplex PCR was designed for these 4 genes using the

same conditions from their individual PCR and a gradient of temperature, between 61 to 64°C. The outcome for this multiplex was very successful, but some of the PCR products had too much or too little product based on the intensity of the amplicons on the agarose gel electrophoresis, and some adjustments on the primer concentrations were made. After several attempts, the multiplex PCR was successful, the 4 amplicons were amplified for each control. The final conditions were: concentration for each primer of 0.18µM for *ATXN7*, 0.31µM *ATXN10*, 0.15µM for *HTT* and 0.28µM for *JPH3*, using 80ng of gDNA for three different controls, a T_m of 63.4°C, solution S at 3X and Buffer 2 (1X) (see Figure 2-15).

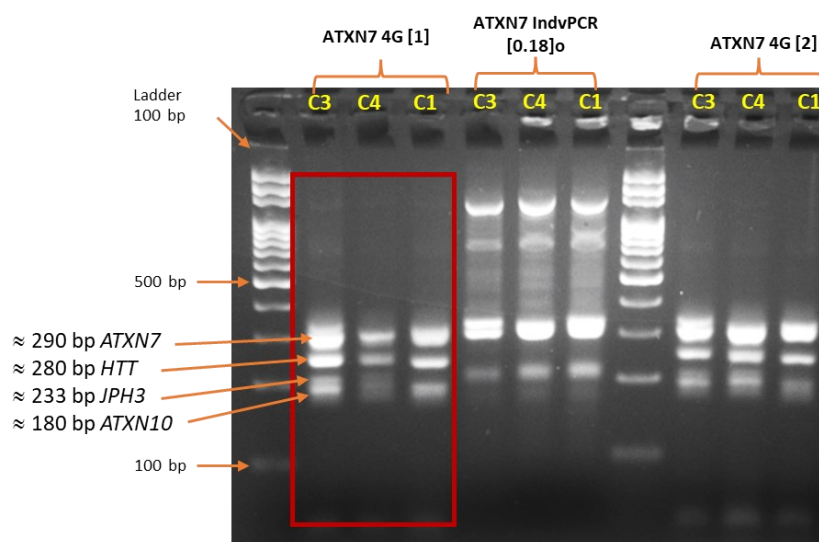


Figure 2-15 Photographic recording of the electrophoretic run on 2.5% agarose gel of the multiplex PCR product of *ATXN7*, *ATXN10*, *HTT* and *JPH3*.

Two different primer concentrations for the multiplex PCR were tested at T_m 63.4°C. The optimum primer conditions of the multiplex were the concentration 1 of primers (*ATXN7* 4G [1], marked by the red square). At the centre of the gel, the PCR products from the individual PCR for *ATXN7* were loaded (initial concentration of primers of 0.18), to check that the primers were working in all the tested PCR's. (gDNA controls: (C3, C4 and C1).

Finally, one of the main objectives of this project was to use the lowest possible amount of gDNA to preserve, as much as possible, the available gDNA on the ALS DNA bank. Hence, a gradient of six different initial concentrations of gDNA (10, 20, 30, 40, 50, 60ng) was set to try on two or three patients from the ALS DNA bank and one random control, using the previous PCR conditions. After viewing the results of the PCR on the agarose gel, the minimum initial quantity of gDNA where all the PCR products are visible is 20ng (final concentration of 1.33

ng/μL) (see Figure 2-16). While this optimization of genes was underway, another gene was added into the optimization panel (*FXN* with ≈378bp). The optimization was initially done using 5 genes but ultimately only four of them could be in the same PCR mix for FLA because the FLA service provider (Eurofins) only allows up to four fluorescent labels plus the chosen size ladder.

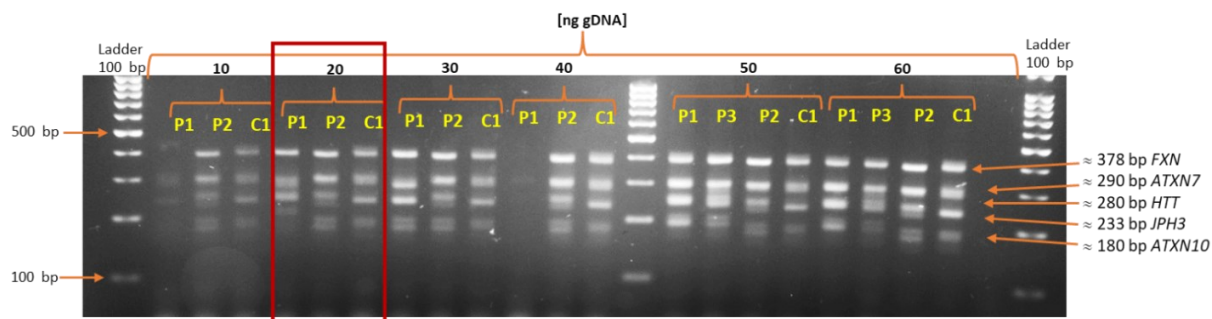


Figure 2-16 Photographic recording of the electrophoretic run on 2.5% agarose gel of the multiplex PCR products from a gDNA gradient of *ATXN7*, *ATXN10*, *HTT* and *JPH3*.

The multiplex PCR was set up having the same conditions but only changing the amount of initial gDNA added to the mix. Two patients and one control (C1) were selected per each concentration of gDNA PCR (Patients: P1, P2, P3). The lowest optimum initial concentration of gDNA for this multiplex is 20ng (marked by the red square).

Multiplex 2: NOP56, PPP2R2B, TBP and TK2-BEANI genes

Optimization of NOP56, PPP2R2B and TBP genes

The detection of fragments containing the REs of *NOP56*, *PPP2R2B* and *TBP* genes were amplified using the following set of oligonucleotides:

- *NOP56* primers (Kobayashi et al. 2011; Lee et al. 2016): **NOP56 fwd F:** 5'-TTTCGGCCTGCGTTCGGG -3'; **NOP56 FRG RP:** 5'-ATCTAGAGCTTTCCAGGCCC-3' (≈ 358 bp containing 7 repeats = (GGCCTG)4(CGCCTG)3).
- *PPP2R2B* primers (Holmes et al. 1999; Lone et al. 2016): **PPP2R2B F:** 5'-CTGCTGGGAAAGAGTCGTG-3' and **PPP2R2B R:** 5'-CGCCAGCGCACTCACCCCTC-3' (≈154 bp containing 10 CAG repeats).

- *TBP* primers (Lone et al. 2015): **TBP F**: 5'-TTTCTCCTTGCTTTCCACAGG-3' and **TBP R**: 5'-ACTCCTAAGTGGGAGGAAACAT-3' ($\approx$ 503 bp containing 38 CAG repeats).

For the three genes, an optimization PCR was designed using Hot FirePol® DNA Polymerase, two different T_m of 60 and 62°C, buffer 2, solution S at 0X, 1.5X and 3X and one gDNA control. After the electrophoresis on agarose gel, for the three genes the desired amplicons were amplified without unspecific bands using a T_m of 62°C, buffer 2 and 3X solution S (Figure 2-17). Since the three genes have the same PCR conditions and cycles, a multiplex PCR was designed for these genes using the same conditions from their individual PCR, a final concentration for each primer of 0.21 μ M for *NOP56* and *PPP2R2B* and 0.28 μ M for *TBP*, and four different controls gDNA at a final concentration of 0.15 ng. After electrophoresis in agarose gel, the multiplex PCR was successful, the three amplicons were amplified for each control (Figure 2-18).

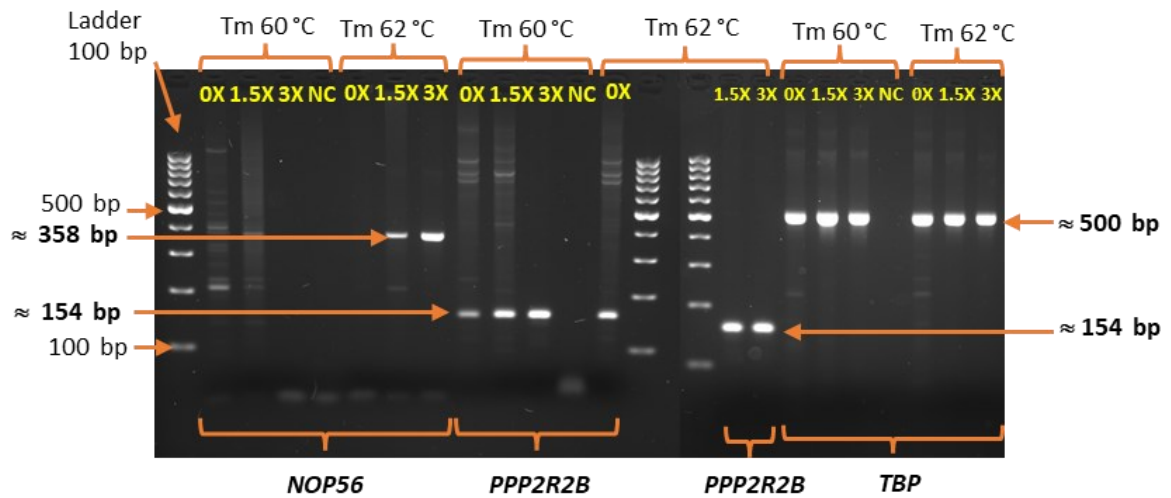


Figure 2-17 Photographic recording of the electrophoretic run on 2.5% agarose gel of the individual PCR products of *NOP56*, *PPP2R2B* and *TBP*.

Gradient of GC enhancer (0X, 1.5X and 3X) and temperature (T_m 60°C and 62°C) for Control 1 (C1). Expected PCR products: *NOP56* of $\approx$ 358 bp, for *PPP2R2B* of $\approx$ 154 bp and for *TBP* of $\approx$ 500 bp (NC: Control of PCR without DNA).

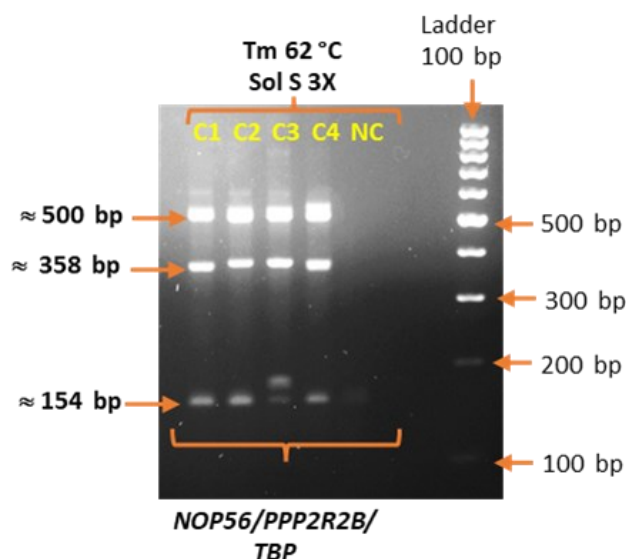


Figure 2-18 Photographic recording of the electrophoretic run on agarose gel 2.5% of the multiplex PCR products of *NOP56*, *PPP2R2B* and *TBP*.

Expected PCR products: *NOP56* of ≈ 358 bp, for *PPP2R2B* of ≈ 154 bp and for *TBP* of ≈ 500 bp. Four gDNA controls (C1-4) and a PCR control (NC: Control of PCR without DNA).

Optimization of *TK2-BEAN1* gene

The detection of fragment containing the RE of *TK2-BEAN1* gene was done by an optimization PCR using Hot FirePol[®] DNA Polymerase, three sets of designed primers, Buffer 2 (1X), two different T_m for each pair of primers and a gradient of solution S at 0X and 3X. After running the PCR products on agarose gels, for *TK2-BEAN1* set F1/R1 with the amplicon of 233bp was amplified successfully with a T_m 60 or 62°C, Buffer 2 (1X) and 3X solution S (see Figure 2-19). The conditions were used in 4 different controls to verify the optimization with a T_m of 62°C (see Figure 2-20).

Designed primers:

- **TK2-BEAN1 F1:** 5'-AAGGTGGAAATTGCAGTGGG-3' and **TK2-BEAN R1:** 5'-CCCTGGGTGGCCAATTCTTA-3'; product size of 223bp; T_m 60°C and 62°C.
- **TK2-BEAN1 F2:** 5'-AGATCACGTCACTGCACTCC-3' and **TK2-BEAN R2:** 5'-CGGGGTGAGAGTTACTGGAT-3'; product size of 239bp; T_m 62°C and 64°C.
- **TK2-BEAN1 F3:** 5'-GGCCAAGATCACGTCACTG-3' and **TK2-BEAN R3:** 5'-CAGCGGGGTGAGAGTTACTG-3'; product size of 247bp; T_m 62°C and 64°C.

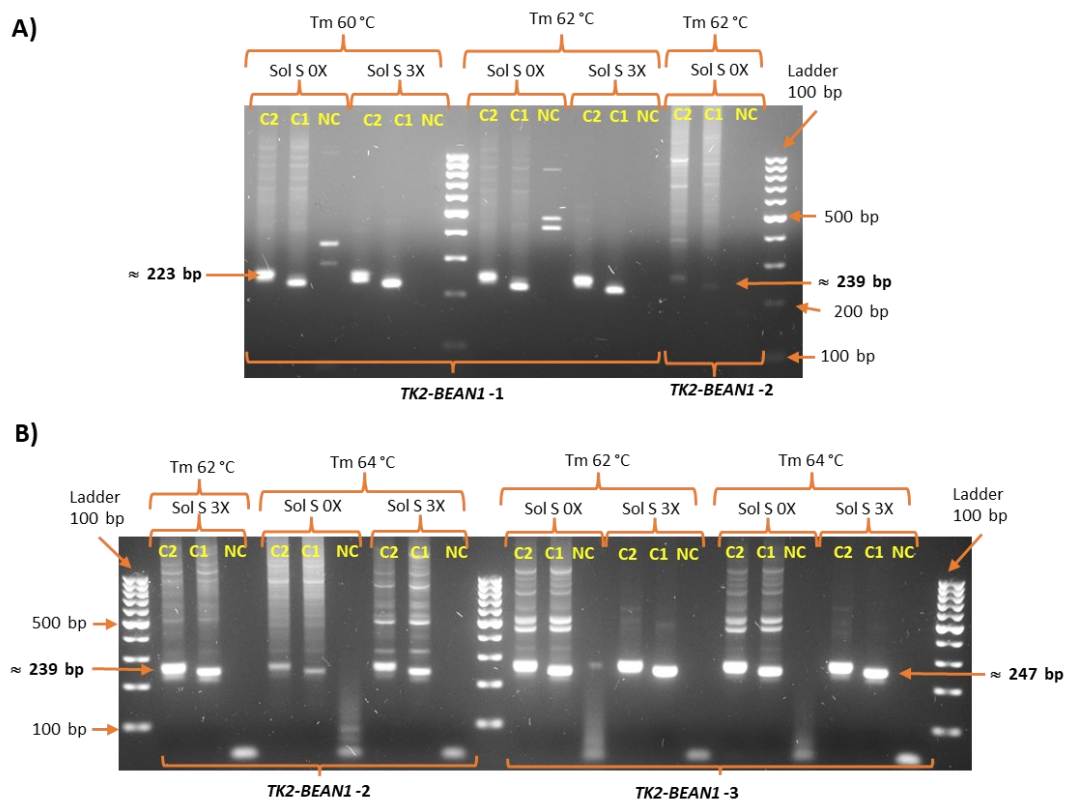


Figure 2-19 Photographic recording of the electrophoretic run on 2.5% agarose gel of the optimisation of three different sets of primers for the RE on *TK2-BEAN1* gene.

A and **B** contains the different tested conditions for *TK2-BEAN1* primers. Each PCR used two gDNA controls (C1 and C2) as well as a control of the PCR without DNA (NC: Negative control). Sol S: Solution S. On this figure the NC has some PCR products. These amplicons are an artefact of primer dimers, in which high concentrations of the primers with high CG can be attached to each other and get amplified during PCR and/or during gel electrophoresis they move more slowly than normal primer dimers.

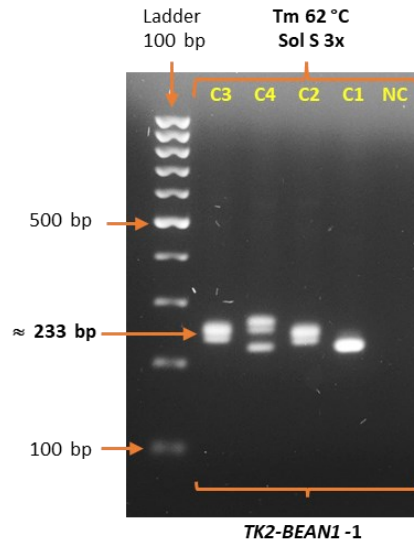


Figure 2-20 2-21 Photographic recording of the electrophoretic run on agarose gel 2.5% of the optimized PCR products of *TK2-BEAN1*.

Four gDNA controls were used to test the PCR conditions (C1, C2, C3 and C4), as well as a control of the PCR without DNA (NC: Negative control).

Multiplex PCR optimization

A multiplex PCR was designed for these 4 genes with a gradient of temperatures of 61, 61, 62, 64°C, solution S 3X and Buffer 2 (1X). The outcome for this multiplex was very successful, the best melting temperature for these genes was 61°C (see Figure 2-22). As before, while this optimization of genes was ongoing, another gene was added into the optimization panel (*CACNA11A* with A PCR product of ≈ 141 bp), but only four could ultimately be included in the FLA run due to a limit on the number of fluorescent labels that could be handled by the FLA provider (Eurofins).

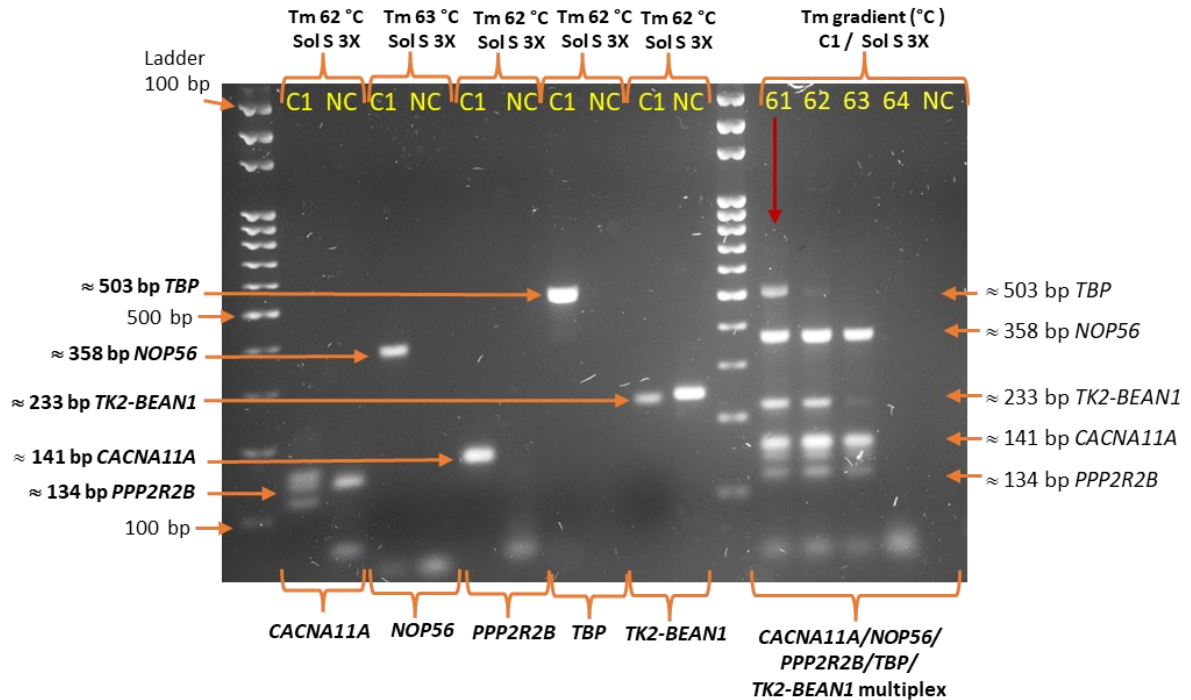


Figure 2-22 Photographic recording of the electrophoretic run on 2.5% agarose gel of the gradient of temperature of the multiplex PCR of *NOP56*, *PPP2R2B*, *TBP*, *TK2-BEAN1* and *CACNA11A* genes and their individual PCRs.

This gel contains the optimised PCRs of each individual gene on the left side of the second ladder. On the right side, the PCR products of the gradient of temperature of the multiplex PCR. Only one control (C1) was used for all the experiments to track that all the PCR products. (NC: negative control, no DNA; Sol S 3X: solution S at 3X). On this figure the NC has some PCR products. These amplicons are an artefact of primer dimers, in which high concentrations of the primers with high CG can be attached to each other and get amplified during PCR and/or during gel electrophoresis they move more slowly than normal primer dimers.

A gradient of six different initial concentrations of gDNA (10, 20, 30, 40, 50, 60ng) was set to try on two or three patients from the ALS DNA bank and one random control, using the previous PCR conditions. After viewing the results of the PCR on the agarose gel, the minimum initial concentration of gDNA where all the PCR products are visible is 20ng (final concentration of 1.33 ng/ μ L) (see Figure 2.25). Since only 4 genes could be on the final multiplex for the FLA run, *NOP56*, *PPP2R2B*, *TBP* and *TK2-BEAN1* were chosen, and various PCR optimizations were designed for different concentration of the primers by only using 20ng of gDNA (final concentration of 1.33 ng/ μ L). The final conditions were: concentration for each primer of 0.18 μ M for *NOP56*, 0.04 μ M *PPP2R2B*, 0.35 μ M for *TBP* and 0.22 μ M for *TK2-BEAN1*, 20ng of gDNA, Tm of 61°C, Buffer 2 (1X) and solution S at 3X (see Figure 2-24).

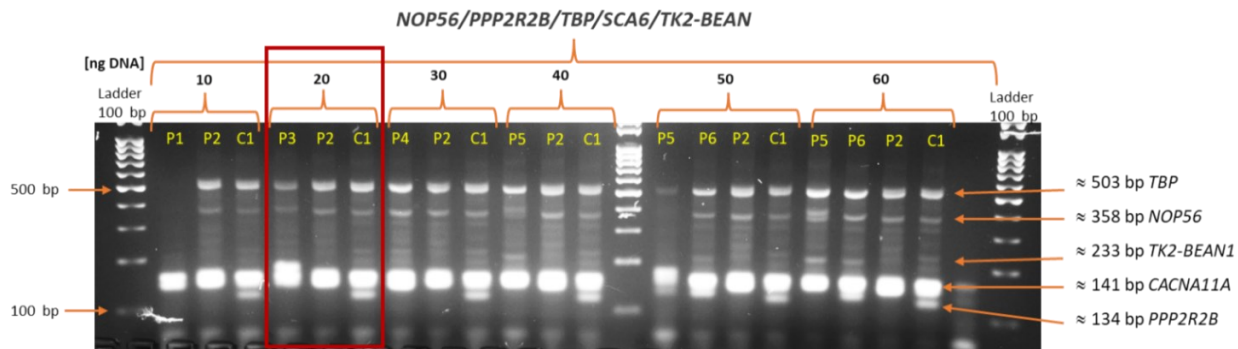


Figure 2-23 Photographic recording of the electrophoretic run on 2.5% agarose gel of the multiplex PCR products from a gDNA gradient of *NOP56*, *PPP2R2B*, *TBP*, *TK2-BEAN1* and *CACNA11A*.

The multiplex PCR was set up having the same conditions but only changing the amount of initial gDNA added to the mix. Two or three patients and one control (C1) were selected per each concentration of gDNA PCR (Patients: P1, P2, P3, P5, P6). The lowest optimum initial concentration of gDNA for this multiplex is 20ng (marked by the red square).

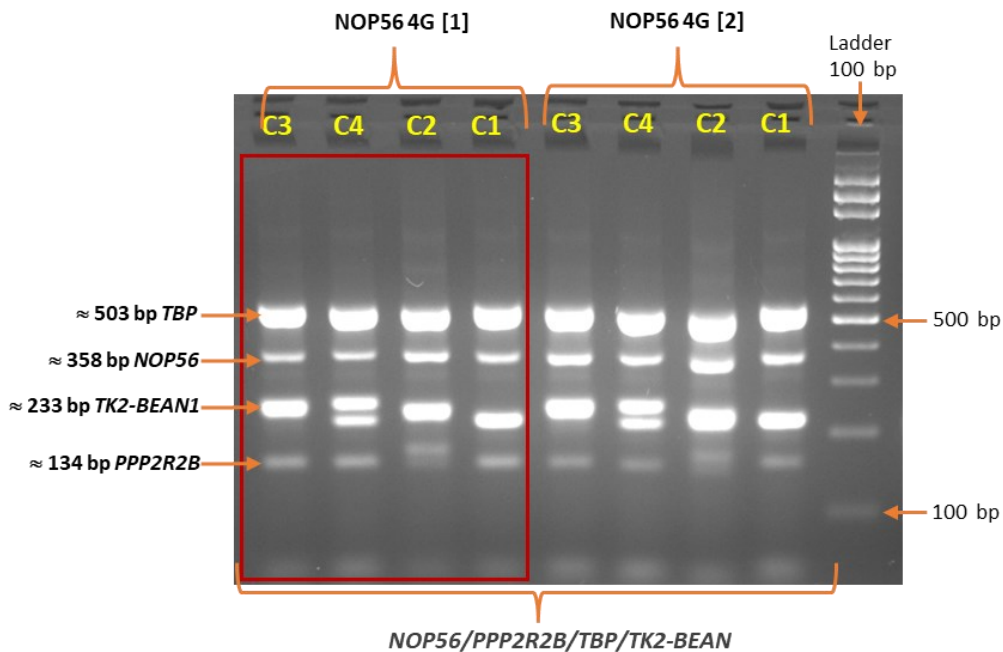


Figure 2-24 Photographic recording of the electrophoretic run on 2.5% agarose gel of the PCR products of the optimized multiplex PCR for of *NOP56*, *PPP2R2B*, *TBP* and *TK2-BEAN1*.

Two different primer concentrations for the multiplex PCR were tested at T_m of 61°C. The optimum conditions of the multiplex were the concentration 1 of primers (*NOP56 4G [1]*, marked by the red square). Four gDNA controls were used to test the two PCR conditions at an initial concentration of 20ng (C1, C2, C3 and C4).

Sequencing results

Each repeat amplicon gene was sequenced using a gDNA control (C1) to corroborate them with their in-silico PCR. These amplicons were obtained using the Snap Gene Viewer software with the GRCh38 Genome reference. All 35 amplicon genes were sequenced in their individual and multiplexed PCRs. All sequences were verified and can be found in Supplementary Figures Chapter 2. Example results are shown in the following subsections for the same eight genes whose optimization is detailed in PCR optimization section. Each individual in-silico PCR sequence (forward and reverse) is shown, followed by the sequencing electropherogram result image. Subsequently, one of either forward or reverse sequencing electropherogram images for that gene is shown for amplicons from their corresponding multiplex PCR.

ATXN7

Primers were taken from Katagiri *et al.*, 2015.

ATXN7 2F: 5'-GTGTTACATTGTAGGAGCGGAA- 3'

ATXN7 R: 5'-CTTCAGGACTGGGCAGAGG- 3'

PCR *in silico*: 290bp → 10 CAG repeats

Sequence forward:

GTGTTACATTGTAGGAGCGGAAAGAATGTCGGAGCGGGCCGCGGATGACGTCAGGGGGGAGCCGC
GCCGCGCGGCGGCGGCGGCGGGCGGAGCAGCGGCCGCGGCCGCCCGGCAGCAGCAGCAGCAGCA
GCAGCAGCAGCAGCCGCCGCTCCGCAGCCCCAGCGGCAGCAGCACCCGCCACCGCCGCCACGGC
GCACACGGCCGGAGGACGGCGGGCCCCGGCGCCGCTCCACCTCGGCCGCCGCAATGGCGACGGTC
GGGGAGCGCAGGCTCTGCCCCAGTCCGAAAG

- Sequencing result of the individual PCR - sequence forward:

ATXN10

Primers were designed using primer3 tool.

ATXN10 2F: 5'- **CAGATGGCAGAATGATAAACTCA** - 3'

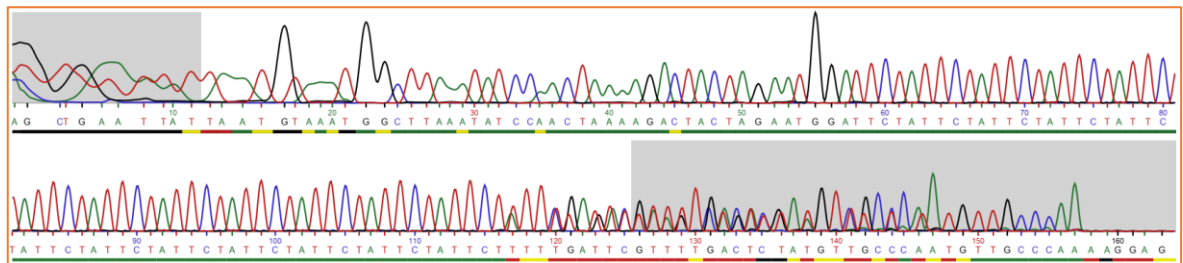
ATXN10 2R: 5'- **TGGGCAACATAGAGAGACTTCA** - 3'

PCR *in silico*: 188bp → 14 ATTCT repeats

Sequence forward:

CAGATGGCAGAATGATAAACTCAATCATGTTGATAAATATATTAAATGTAAATGGCTTAAATATCC
AACTAAAAGACTACTAGAATGGATTCTATTCTATTCTATTCTATTCTATTCTATTCTATTCTATT
CTATTCTATTCTATTCTATTCTATTCTTTTGAGA**TGAAGTCCTCTATGTTGCCCA**

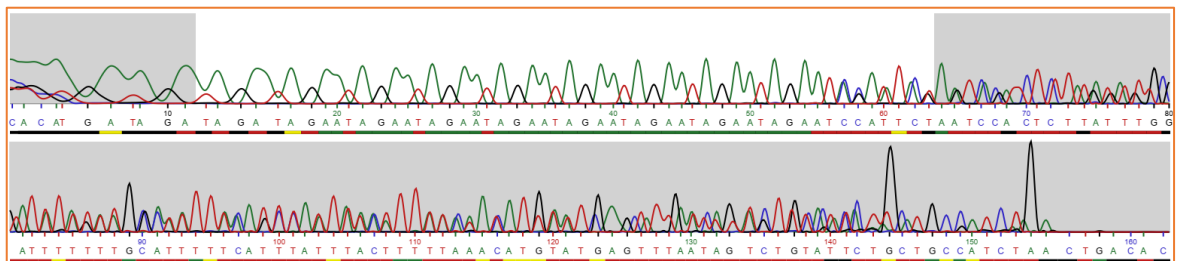
- Sequencing result of the individual PCR - sequence forward:



Reverse complementary sequence:

TGGGCAACATAGAGAGACTTCATCTCAAAAAGAATAGAATAGAATAGAATAGAATAGAATAGAA
TAGAATAGAATAGAATAGAATAGAATAGAATAGAATCCATTCTAGTAGTCTTTTAGTTGGATATT
TAAGCCATTACATTTAATATATTTA**CAACATGATTGAGTTATCATCTGCCATCTG**

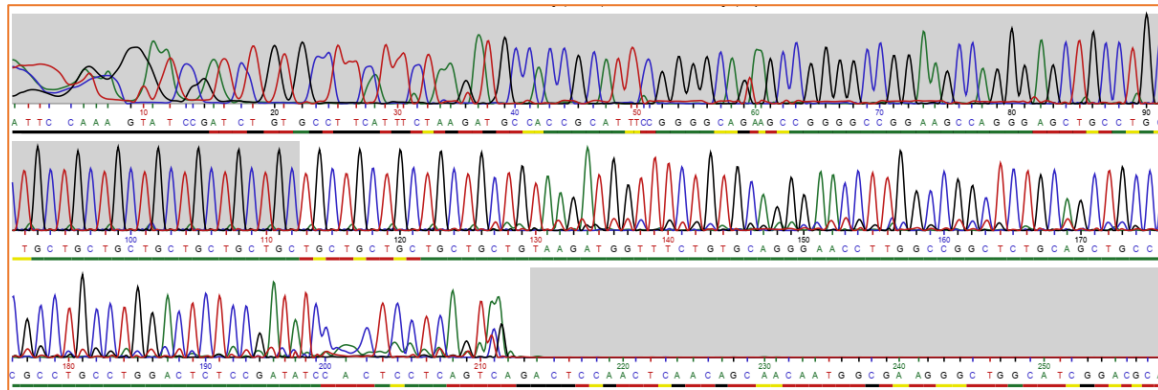
- Sequencing result of the individual PCR - sequence reverse:



- Sequencing result multiplex PCR - sequence reverse:

CTCTCCGATATCCACTCCTCAGT

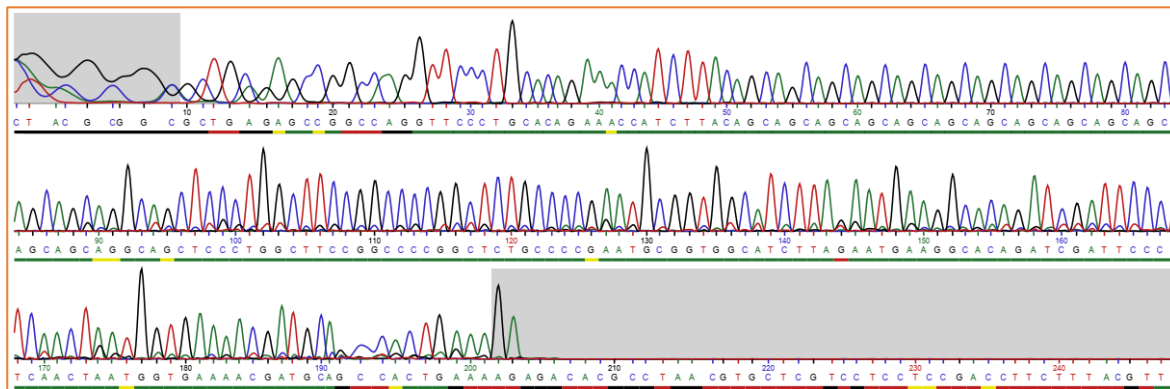
- Sequencing result of the individual PCR - sequence forward:



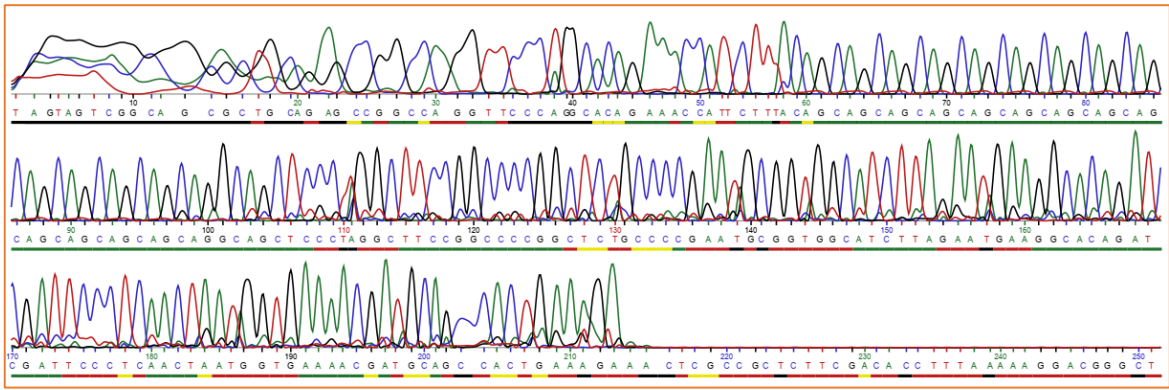
Reverse complementary sequence:

[illegible]

- Sequencing result of the individual PCR - sequence reverse:



- Sequencing result multiplex PCR - sequence reverse:



NOP56

Primers were taken from Kobayashi *et al.*, 2011 and Y.-C. Lee *et al.*, 2016

NOP56 Fwd R : **TTTCGGCCTGCGTTCGGG**

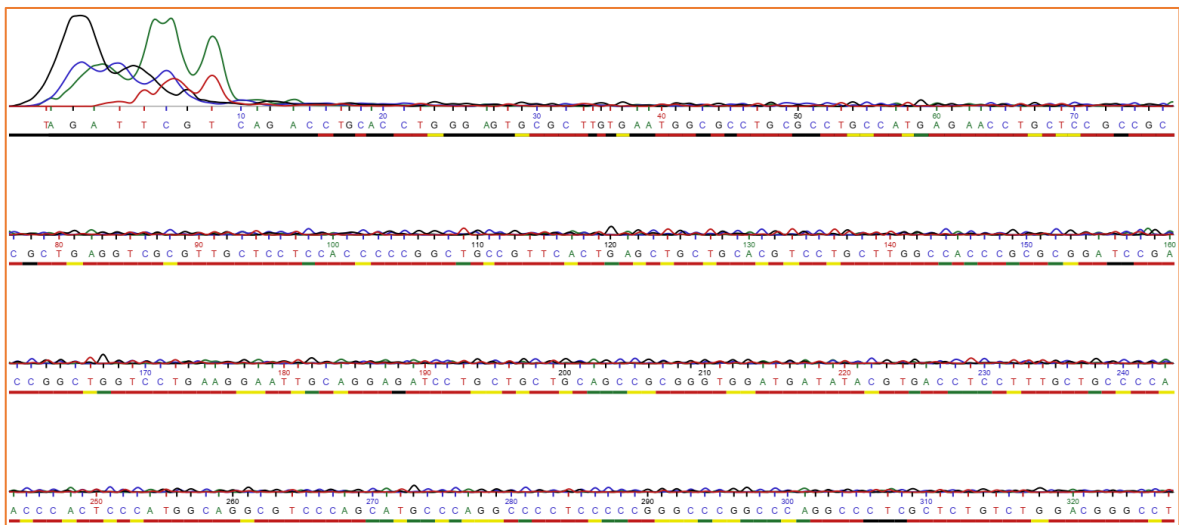
Nop56 Frg RP : **ATCTAGAGCTTTCCAGGCC**

PCR *in silico*: 358bp → (GGCCTG)₄ and (CGCCTG)₃ Repeats = 7 repeats

Sequence forward:

TTTCGGCCTGCGTTCGGG CCGCAGACAG **GGCCTGGGCCTGGGCCTGGGCCTGCGCCTGCGCCT**
CGCCTG CCCTGGGAACGGGTTCGGCAGACGCTGAGGTTGCGTTGACGCTCGCGCCCCGGCTCCC
 GTTCCAGGTGCTGTTGCACGTGCTGTTTAGACACGCGGTCGGCTACGCGCTGCTGGCGCTGAAGGA
 AGTGGAGGAGATCAGTCTGCTGCAGCCGACGGTGGGTGAGATCCGTGGGCTCCTTTGGCGGCCCG
 CAGACCCATCGCGCGGGTCCCAGCATGCACAGCGCGCTCCACGTGGCCGGCCGAGCGGTTCCGG
 GGCGGGGGGAC **GGCCTGGAAAGCTCTAGAT**

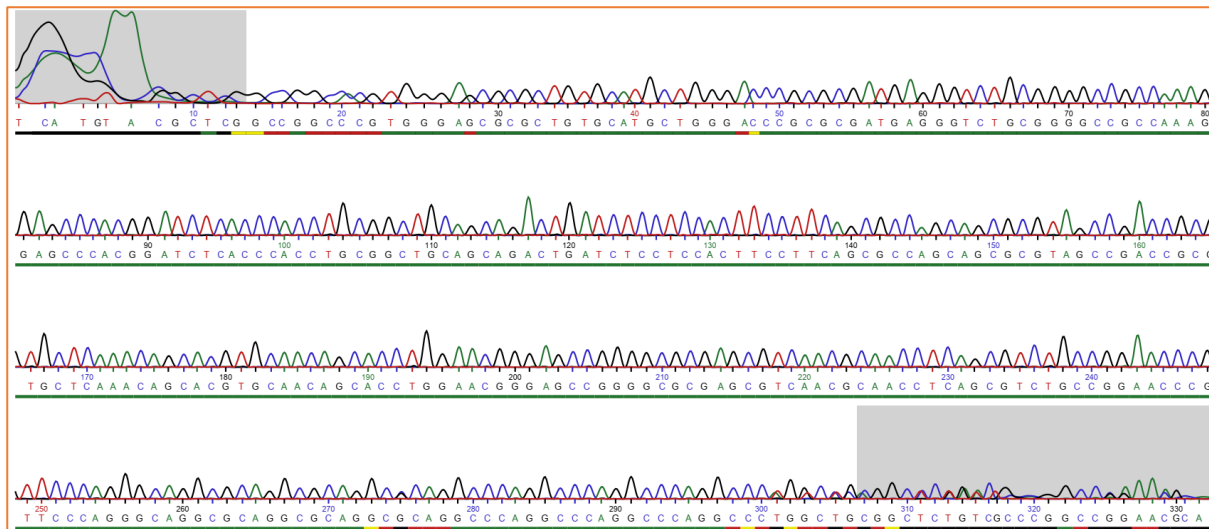
- Sequencing result of the individual PCR - sequence forward:



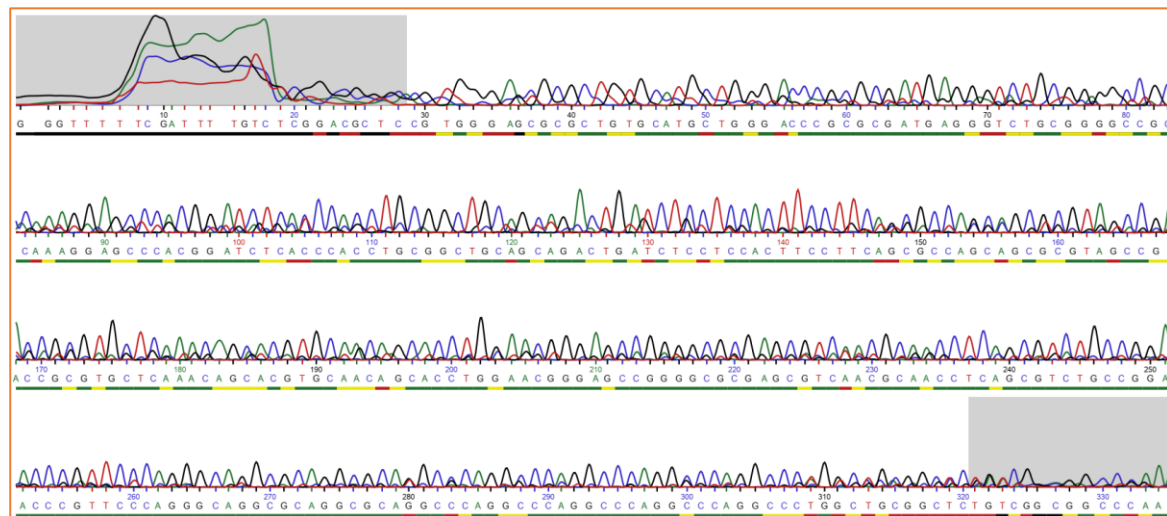
Reverse complementary sequence:

ATCTAGAGCTTTCAGGCCCTGTCCCCCGCCCCGAACCGCTCGGCCGCCACGTGGGAGCGCGCTG
 TGCATGCTGGGACCCGCGCGATGAGGGTCTGCGGGGCCGCCAAAGGAGCCACGGATCTCACCCAC
 CTGCGGCTGCAGCAGACTGATCTCCTCCACTTCTTCAGCGCCAGCAGCGCGTAGCCGACCGCGTGC
 TCAAACAGCACGTGCAACAGCACCTGGAACGGGAGCCGGGGCGCGAGCGTCAACGCAACCTCAGC
 GTCTGCCGGAACCCGTTCCAGGGCAGGCGCAGGCGCAGGCGCAGGCCAGGCCAGGCCAG
 GCCCTGTCTGCGGCCCGAACGCGAGGCCGAAA

- Sequencing result of the individual PCR - sequence reverse:



- Sequencing result multiplex PCR - sequence reverse:



PPP2R2B

Primers were taken from Holmes *et al.*, 1999 and Lone *et al.*, 2016.

PPP2R2B F: 5'-CTGCTGGGAAAGAGTCGTG-3'

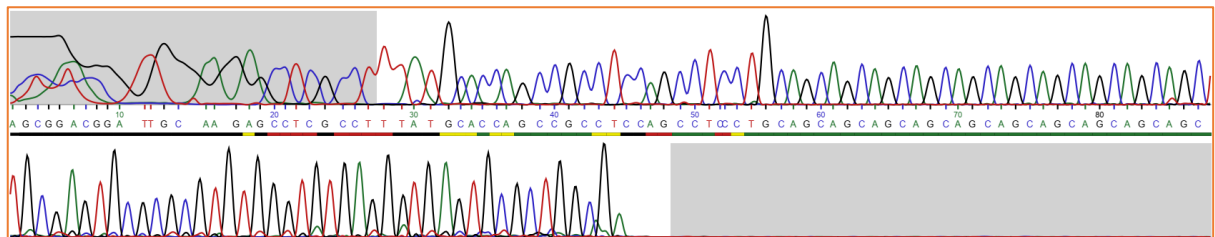
PPP2R2B R: 5'-CGCCAGCGCACTCACCTC 3'

PCR *in silico*: 154bp → 10 **CAG** repeats

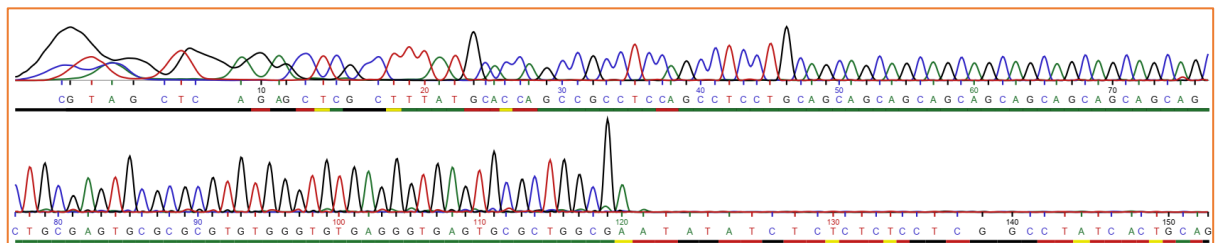
Sequence forward:

[illegible]

- Sequencing result of the individual PCR - sequence forward:



- Sequencing result multiplex PCR - sequence forward:



Reverse complementary sequence:

TCCAAGCGCACTACCC T ACACCCACACGCGCGCACTCGCAG CTGCTGTGCTGTGCTGTGCTGC
TGCTGTGTGCAGGAGGCTGGAGGCGGCTGGTGCATTAAAGGCGCAGGCTCCTCCCAACCGCGTCAGC
AGCCC CACGACTCTTTCCCAGCAG

- Sequencing result of the individual PCR - sequence reverse:

TK2-BEANI

Primers were designed using primer3 tool.

TK2-BEAN1 F1: 5'- AAGGTGGAAATTGCAGTGGG -3'

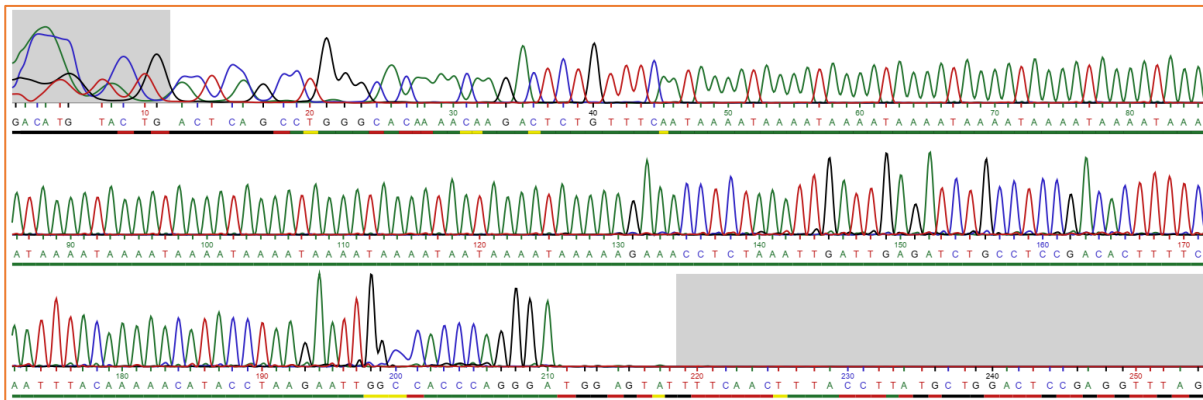
TK2-BEAN1 R1: 5'-CCCTGGGTGGCCAATTCTTA-3'

PCR *in silico*: 223bp → TGGAA (TTCCT repeats (insertion))

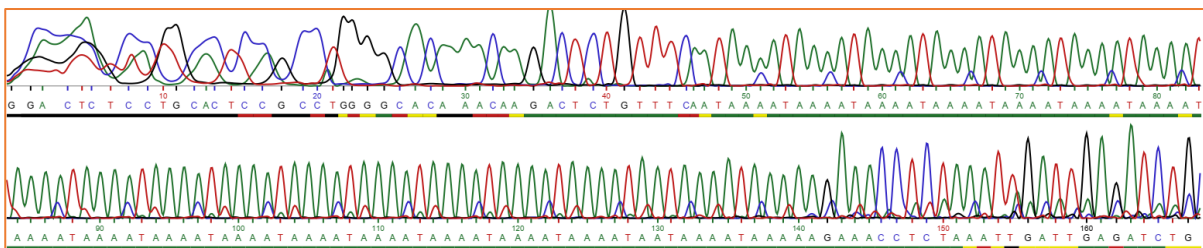
Sequence forward:

AAGGTGGAAATTGCAGTGGGCCAAGATCACGTCACTGCACTCCAGCCTGGGCAACAAAACAAGAC
TCTGTTTCAA[Insertion]TAAATAAAATAAAATAAAATAAAATAAAATAAAATAAAATAAA
ATAAAATAATAAAATAAAAAGAACCTCTAAATTGATTGAGATCTGCCTCCGACACTTTTCAATTTA
CAAAAACATACCTAAGAATTGGCCACCCAGGG

- Sequencing result of the individual PCR - sequence forward:



- Sequencing result multiplex PCR - sequence forward:



Reverse complementary sequence:

TC TGGG TGG CCA ATCT CTA GGTATG TTTTGT AAATTG AAAAGT GTCGGAGGCAGATCTCAATCAA
TTTAGAGGTTTCTTTTATTTATTTATTTATTTATTTATTTATTTATTTATTTATTTATTTA
TTTATTTTA [Insertion] TTGAAACAGAGTCTTGTTTGTGCCCAGGCTGGAGTGCAGTGACGTGATC
TTGG CCCACTGCAATTTCCACCTT

- [illegible]

FLA results

As soon as each individual and multiplex gene amplicon was corroborated, an FLA pilot was designed with the laboratory at Eurofins Medigenomix GmbH, Ebersberg in Germany. There are many different filter set ups that the company offers to properly separate and visualize the fluorescent amplicons on a multiplex PCR. This so-called “filter-set” is how the machine interprets the different light wavelengths of the fluorescent labels on the primers after exposure to a laser. After the combination of fluorescent labels were decided for each multiplex PCR, the laboratory suggested to try two filters sets (DS33 and MWGG5) and several dilutions of the PCR products.

Results from the CE of each of the multiplexes were analysed and one of the filter-sets was selected. For the two previous described multiplexed PCRs, the dilution that worked best for both PCRs was 1:20, and the filter-set that gave a stronger signal for each colour-set was MWGG5.

Figure 2-25 and Figure 2-26 shows the selected filter-set for each of the two multiplexes previously shown in the results section. Not all the genes were able to be optimized by 4-gene multiplexes. Most genes were optimized in pairs and then mixed up with other multiplexes or single PCRs to make combinations to up to 4 different PCR products. PCR amplicons were mixed in different proportions before sending them for CE. This allowed to reduce the number of FLA genes combinations, significantly reducing the costs of the overall analysis. All the conditions for each combination of PCR-FLA of the neurodegenerative repeat expansion genes panel can be found in Supplementary Table S 2-3 and Supplementary Figures Chapter 2.

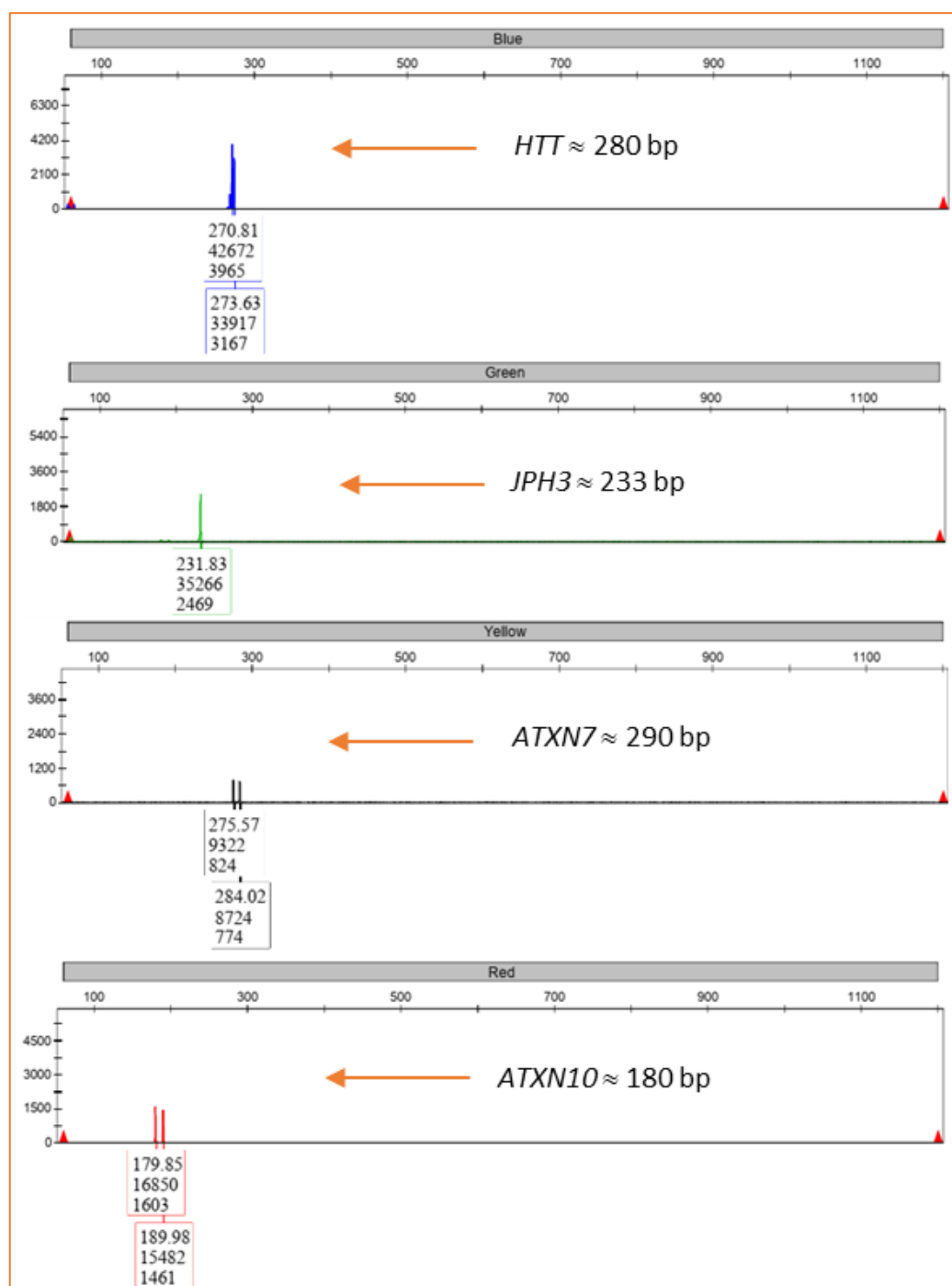


Figure 2-25 Capillary electropherogram of the Multiplex 1: *HTT*, *JPH3*, *ATXN7* and *ATXN10* genes.

Panel 1: Filter-set MWGG5, dilution 1:20, Control 1 gDNA. Blue traces correspond to *HTT* gene with Fluorescein label; C1 is heterozygote with 270.81bp and 273.63bp alleles. Green trace corresponds to *JPH3* gene with Yakima Yellow label; C1 is homozygote with 231.83bp allele. Yellow traces correspond to *ATXN7* gene with Atto550 label; C1 is heterozygote with 275.57bp and 284.02bp alleles. Red traces correspond to *ATXN10* gene with Atto565 label; C1 is heterozygote with 179.85bp and 189.98bp alleles.

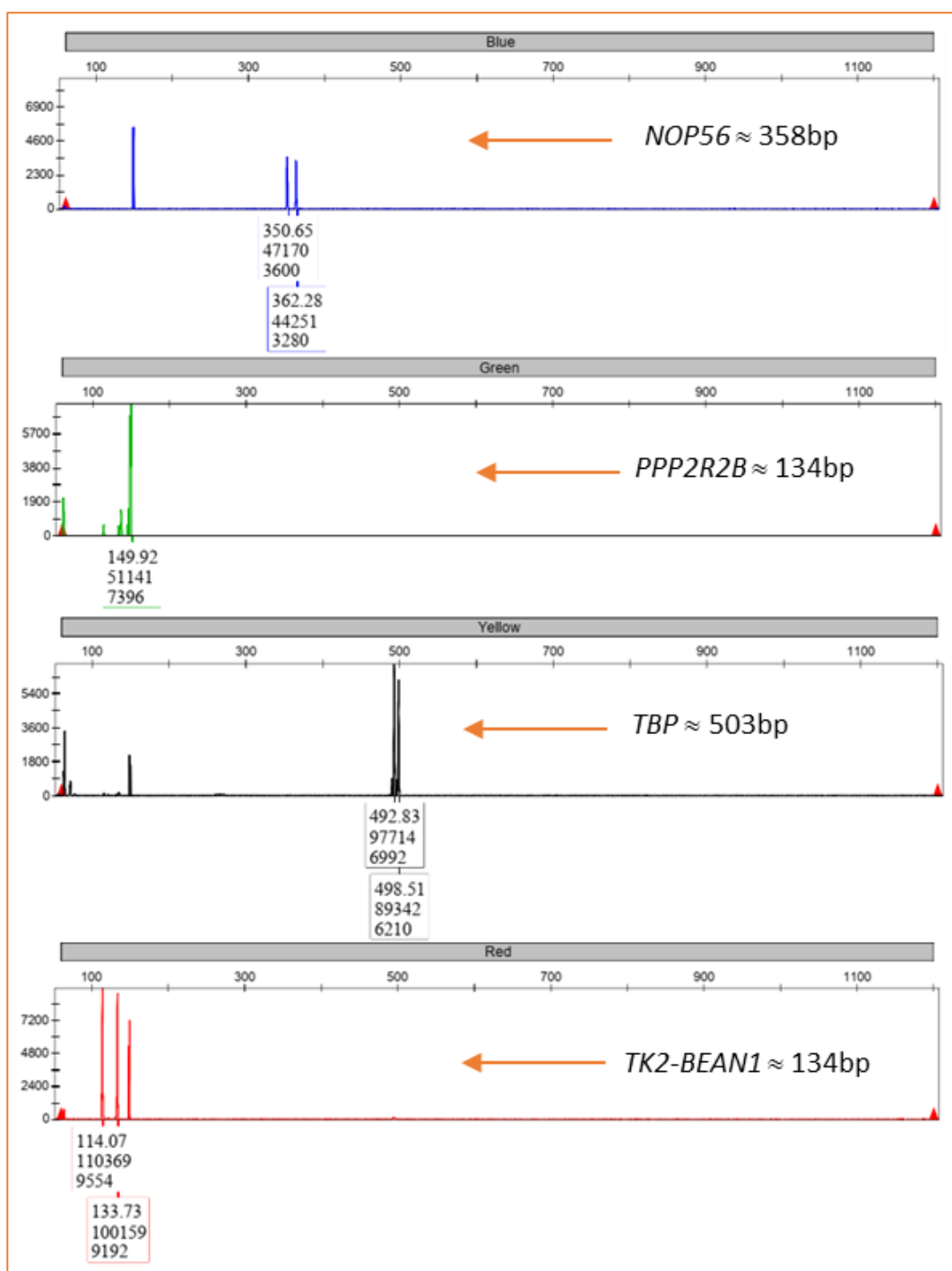


Figure 2-26 Capillary electropherogram of the Multiplex 2: *NOP56*, *PPP2R2B*, *TBP* and *TK2-BEAN1* genes.

Panel 2: Filter-set MWGG5, dilution 1:20, Control 1 gDNA. Blue traces correspond to *NOP56* gene with Fluorescein label; C1 is heterozygote with 350.65bp and 362.28bp alleles. Green trace corresponds to *PPP2R2B* gene with Yakima Yellow label; C1 is homozygote with 149.92bp allele. Yellow traces correspond to *TBP* gene with Atto550 label; C1 is heterozygote with 492.83bp and 498.51bp alleles. Red traces correspond to *TK2-BEAN1* gene with Atto565 label; C1 is heterozygote with 114.07bp and 133.73bp alleles.

Note: During FLA, there is some overlapping between some of the traces that are very similar in the light spectrum. Therefore, while looking into *TK2-BEAN1*, there is an extra peak at 133.7bp that corresponds to the green label

from *PPP2R2B*. Likewise, The same peak is shown in the *NOP56* blue panel. Eurofins has a software that identifies the overlapping and chooses the correct peaks for each gene.

Repeat count calibration

Each allele for every NDRE gene was processed by CE which is the most accurate tool for measuring the bp size of each amplicon. Equation 2-1 (see section Capillary electrophoresis-Fragment Length Analysis (FLA)) is a guideline for converting bp into a repeat count. The bp length of each amplicon in CE is not a rounded number, it has up to two decimal places. Taking this into account, round up and down was applied and this added a potential miscount in the repeat count. In addition to this, each gene has a common repeat interval instead of a common repeat. Initially, the amplicon length and common repeat value used in the formula was the repeat count found at the reference genome (GRCh38) obtained by in-silico PCR. Therefore, corrections in the repeat count formula were made for each gene amplicon.

Sequencing was carried on two or three homozygote patients and/or controls for each repeat gene. The results were compared with the repeat count obtained by the formula. If there was a difference, adjustments were made to the formula by adding/subtracting repeats or changing the common repeat with the known bp value of the amplicon that contains that number of repeats. Table 2-1 contains the all the information of the corrections made for each gene in multiplex 1 and 2. Supplementary Table S 2-4 contains the information of all corrections for each NDREs gene.

For sequencing, patient 1 (P1) and control 5 and 6 (C5, C6) were selected for the *HTT* gene. All three samples were homozygotes for the repeat amplicon. The repeat count obtained with the formula was 13, 18 and 13, respectively, using 21 CAG as the common repeat count and 272bp as the common repeat length. Analysed results of each sequence showed that P1 and C6 had four extra CAGs, and C5 had three more repeats than theoretically (Figure 2-27). Since P1 and C6 had the same allele length and exact sequencing repeat count, the modification applied to the formula for *HTT* was to change the common repeat length (bp) to 261bp (instead of 272bp) and the common repeat to 17 (instead of 21). This modification, when applied to the results, corrected the undercounted repeats.

In the case of *JPH3*, the difference in repeat counts between the selected homozygotes was one repeat count (CAG) (Figure 2-28). The formula correction for the *JPH3* gene was to add an extra repeat to the count obtained by the formulas as shown in red:

Equation 2-2

$$\# \text{ of repeats} = \left[\text{Common repeat count} + \left(\frac{\text{Allele length (bp)} - \text{Common repeat length (bp)}}{\text{Repeat size (bp)}} \right) \right] \\ + (\text{Repeat modification})$$

Common repeat count: number of repeats present in the reference genome; Common repeat length: size of the in-silico PCR from the reference genome; Allele length: PCR fragment length obtained by CE; Repeat size: repeat motif length; Repeat modification: number of extra repeats obtained in the repeat count calibration (consensus modification).

This modification in the formula was used for many of the genes in the NDREs panel. Other genes, such as *ATXN10*, did not require any corrections because the sequencing repeat results matched the repeat count obtained by the formula with the original assumptions (Figure 2-29).

On the other hand, for *TK2-BEANI*, the RE of interest in this analysis is a TGGAA repeat insertion that occurs upstream of another known benign repeat (TAAAA). The results from the homozygotes individuals showed no TGGAA insertion, and instead, the PCR formula measured the TAAAA repeat sequence (Figure 2-30). For this reason, the required modification for the *TK2-BEANI* gene was 242bp as the common repeat length, equivalent to 15 repeats as the “common” repeat count.

Table 2-1 Repeat count calibration results for each gene obtained by sequencing.

<i>Gene</i>	<i>ID</i>	<i>Primer label</i>	<i>Allele length (bp)</i>	<i>Allele repeat count</i>	<i>Sequencing repeat count</i>	<i>Difference repeats count</i>	<i>Consensus modification</i>
<i>HTT</i>	P1	Fluorescin (Blue)	261	13	17	4	Change
	C5		275	18	21	3	261 bp = 17 repeats
	C6		261	13	17	4	
<i>JPH3</i>	P2	Yakimma	230	13	14	1	Add 1 extra repeat to formula
	C7	Yellow	232	13	14	1	
	C8	(Green)	232	13	14	1	
<i>ATXN7</i>	P3	Atto 550 (Yellow)	273	4	10	6	Change
	P4		276	4	10	6	274 bp = 10 repeats
<i>ATXN10</i>	P6	Atto 565	180	12	12	0	No changes
	C9	(Red)	190	14	14	0	
<i>NOP56</i>	P7	Fluorescin (Blue)	349	3	4	1	Add 1 extra repeat to formula
	C10		350	3	4	1	
	C1		363	3 & 5	4 & 6	1	
<i>PPP2R2B</i>	P8	Yakimma	149	8	10	2	Add 2 extra repeat to formula
	C12	Yellow	149	8	10	2	
	C13	(Green)	149	8	10	2	
<i>TBP</i>	P9	Atto 550 (Yellow)	489	33	35	2	Change
	C14		488	33	35	2	488 bp = 35 repeats
<i>TK2-BEANI</i>	P10	Atto 565 (Red)	242	4	15 (TAAAA)	11	Change
	C15		242	4	15 (TAAAA)	11	242 bp = 15 repeats
	C16		246	5	16 (TAAAA)	11	

plates of the same gene. Even though the variance is subtle between plates, the difference originates from differences between the two decimals and how they were rounded.

This issue leads to results with a variance of 1 whole repeat. To address this problem, the replicated results for C1 per gene were filtered by its maximum repeat count. The average allele length for that repeat was then used as the “true value” for the other plate results with different repeat counts. The variance value between their allele length and the “true value” was applied to the formula of the repeat count per individual plate. To verify the plate fix per gene, some of the previous sequenced results for the repeat calibration were used to validate that the change in the formula was correct.

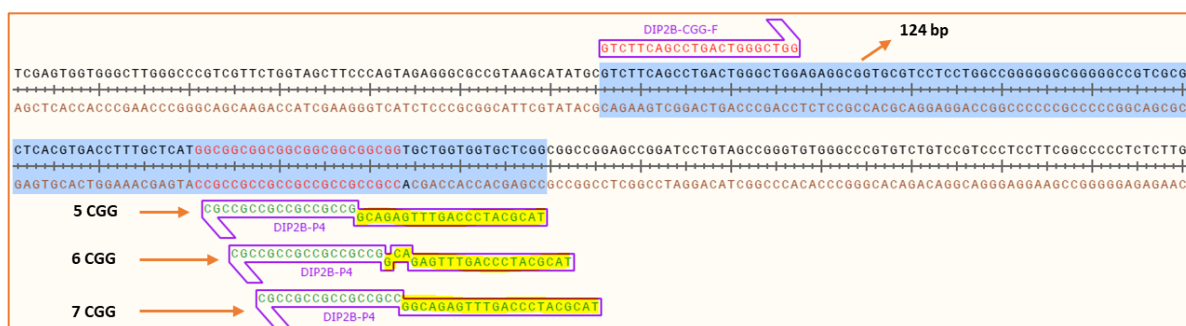
Table 2-2 has an example of three genes with their plate calibration. For the *PPP2R2B* gene, C1 is a homozygote with 10 CAG repeats (Allele_1). The average length from 27 plates with this repeat is 148.48bp. The plate fix was applied to 10 different plates in which the repeat count was over 10 CAG repeats. This solution is also applied to the second allele (Allele_2) because C1 is a homozygote. In the case of heterozygosis for C1, each allele is treated independently, and each fix is applied. In the case of having a plate with two different changes, each fix will take place. Table 2-2 shows the *HTT* gene in which C1 is heterozygote with 20 CAG repeats, and nine plates had an extra repeat for allele 1. While in allele 2, only four plates had one additional repeat. These four plates were paired up, and both individual bp differences were applied to their respective allele. For the other six plates, the fix applied in allele 1 also applied to allele 2. The same rule would apply in the opposite case. The gene *NOTCH2NLC* is an example where only one plate (AR-P01) had one repeat less than the rest plates in allele 2. The difference to calibrate repeat count was 2.83bp for allele 1 and 2. Supplementary Table S 2-5 contains the complete allele-plate count calibration

Table 2-2 Allele-plate repeat count calibration using replicated control 1 results.

<i>Plate number</i>	<i>Gene</i>	<i>Allele 1 repeat</i>	<i>Average allele 1 length (bp)</i>	<i>Allele 1 length (bp)</i>	<i>Allele 1 difference (bp)</i>	<i>Allele 2 repeat</i>	<i>Average allele 2 length (bp)</i>	<i>Allele 2 length (bp)</i>	<i>Allele 2 difference (bp)</i>
<i>NOP56-P12</i>	<i>PPP2R2B</i>	10	148.81	150.07	-1.26				-1.26
<i>NOP56-P15</i>	<i>PPP2R2B</i>	10	148.81	149.93	-1.12				-1.12
<i>NOP56-P17</i>	<i>PPP2R2B</i>	10	148.81	149.84	-1.03				-1.03
<i>NOP56-P19</i>	<i>PPP2R2B</i>	10	148.81	149.57	-0.76				-0.76
<i>NOP56-P25</i>	<i>PPP2R2B</i>	10	148.81	150.03	-1.22				-1.22
<i>NOP56-P26</i>	<i>PPP2R2B</i>	10	148.81	150.08	-1.27				-1.27
<i>NOP56-P27</i>	<i>PPP2R2B</i>	10	148.81	149.59	-0.78				-0.78
<i>NOP56-P29</i>	<i>PPP2R2B</i>	10	148.81	149.92	-1.11				-1.11
<i>NOP56-P32</i>	<i>PPP2R2B</i>	10	148.81	150.21	-1.40				-1.40
<i>NOP56-FTD</i>	<i>PPP2R2B</i>	10	148.81	150.17	-1.36				-1.36
<i>ATXN7-P10</i>	<i>HTT</i>	20	271.03	271.59	-0.56	21	273.91	274.61	-0.70
<i>ATXN7-P18</i>	<i>HTT</i>	20	271.03	271.72	-0.69	21	273.91	274.6	-0.69
<i>ATXN7-P19</i>	<i>HTT</i>	20	271.03	271.54	-0.51				-0.51
<i>ATXN7-P31</i>	<i>HTT</i>	20	271.03	271.66	-0.63				-0.63
<i>ATXN7-P07</i>	<i>HTT</i>	20	271.03	271.66	-0.63	21	273.91	274.62	-0.71
<i>ATXN7-P08</i>	<i>HTT</i>	20	271.03	271.57	-0.54				-0.54
<i>ATXN7-P17</i>	<i>HTT</i>	20	271.03	271.65	-0.62				-0.62
<i>ATXN7-P36</i>	<i>HTT</i>	20	271.03	271.51	-0.48				-0.48
<i>ATXN7-P32</i>	<i>HTT</i>	20	271.03	271.82	-0.79	21	273.91	274.73	-0.82
<i>AR-P01</i>	<i>NOTCH2NLC</i>	16	205.53	202.7	2.83				2.83

Similarly to the previously described PCR methodology, rpPCR protocols for ten genes were optimised using different polymerases and protocols, a gradient of temperatures and concentration of reagents, such as magnesium, DMSO, GC-buffer enhancer, primers, and DNA concentration. Each rpPCR product was viewed by CE using the SeqStudio device. This section shows an optimisation for rpPCR of the *DIP2B* gene as an example. The complete optimised genes conditions for each rpPCR protocol can be found in the Supplementary Table S 2-6.

PCR *in silico*: 124bp (highlighted in blue) → 5 **CGG** repeats



"begins" for this gene on the in-silico PCR. This tallest peak includes 5 CGG repeats, and the next adjacent peak will increase by 1 CGG (Figure 2-32).

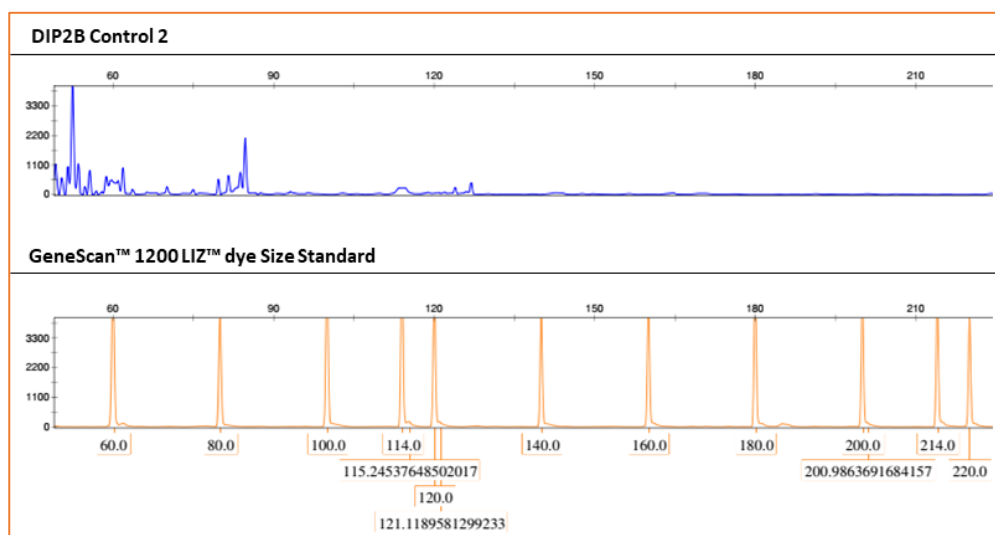


Figure 2-31 Capillary electropherogram results obtained from the SeqStudio™ of *DIP2B* rpPCR optimisation.

Upper panel shows the CE results of the rpPCR products for the *DIP2B* gene. The conditions for this PCR were: Hot FirePol® DNA Polymerase (Solis Biodyne) with a TA of 68°C, gDNA from Control 2 and size ladder GeneScan™ 1200LIZ™ dye size standard in bp (orange peaks, lower panel).

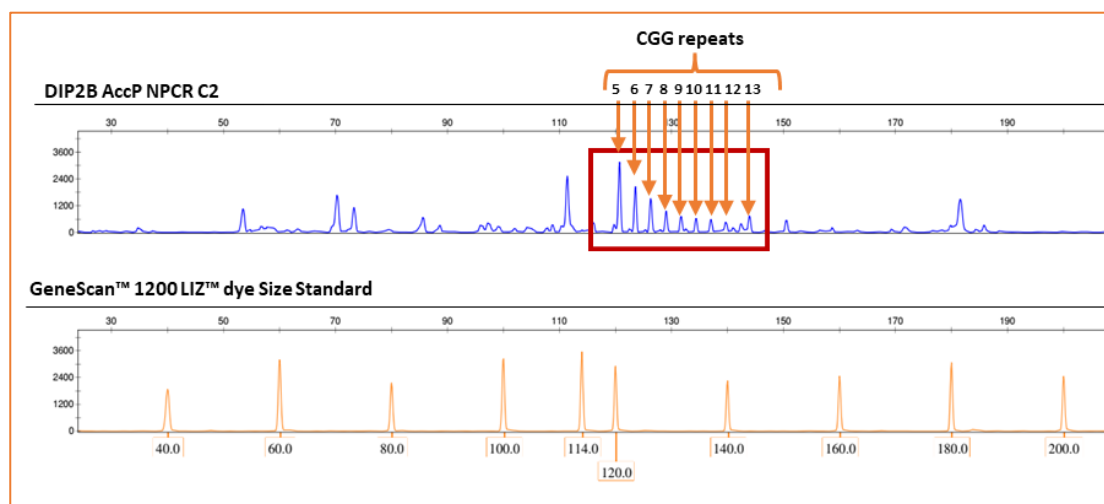


Figure 2-32 Capillary electropherogram from the optimised rpPCR for the *DIP2B* gene obtained from the SeqStudio™.

Upper panel shows the smooth peaks pattern obtained by the rpPCR of the *DIP2B* gene starting around 120 bp. The first peak has 5 CGG repeats, increasing by 1 repeat every following peak. C2 has 13 repeats for *DIP2B*. Lower panel corresponds to the size ladder GeneScan™ 1200LIZ™ dye size standard shown with orange peaks. (gDNA Control 2: C2).

rpPCR is a great tool to count long repeats in a specific gene region. This technique is used for research, and diagnostic laboratories on *C9orf72* GGGGCC repeats, in which counting more than 30 repeats means you have a pathological expansion in this gene, increasing your risk for ALS. This project aims to amplify intermediate REs on 35 NDREs genes adequately. For this reason, rpPCR was used to verify the suspected expanded or long repeats obtained in the PCR panel results. After all the samples were amplified for the *DIP2B* gene, an ALS case and control showed expanded alleles of 24 repeats. Eurofins Germany processed the rpPCR products for both samples to confirm these expansions using CE. Results from the electropherograms confirmed that both individuals had the RE obtained by the multiplex PCR panel protocol (Figure 2-33).

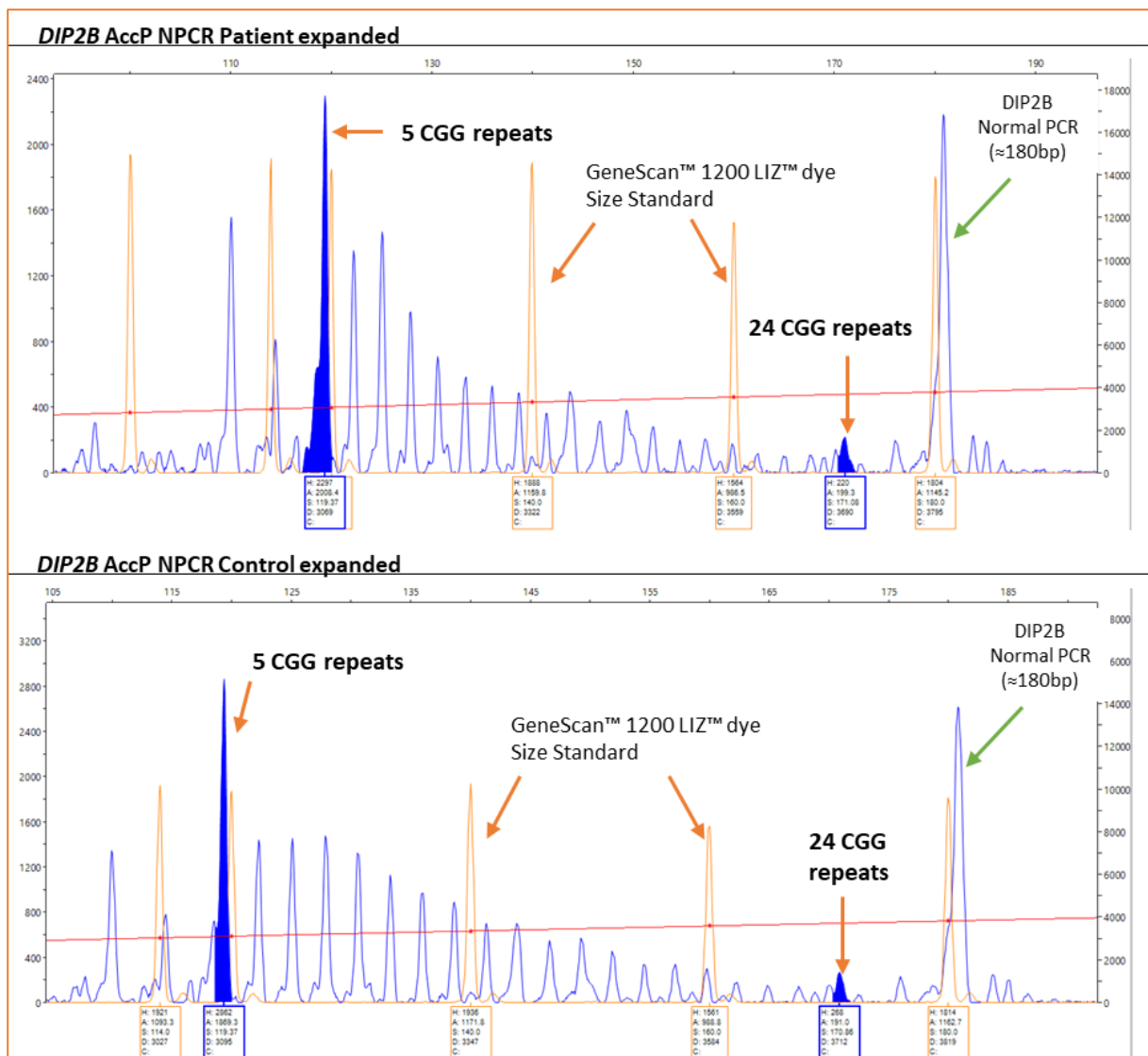


Figure 2-33 Capillary electropherogram of the *DIP2B* rpPCR products of two individuals with suspected CGG expansions.

This figure shows the schematic results of a CE run from an ALS patient and a control (upper and lower panel, respectively) with 24 CGG repeats. Each panel show the pattern of the smooth peaks obtained by the rpPCR of the *DIP2B* gene starting around 120 bp (blue peaks). The first peak has 5 CGG repeats, increasing by one repeat with each subsequent peak. Peaks in orange are from the size ladder GeneScan™ 1200LIZ™ dye size standard.

Note: The shown RP-PCR for *DIP2B* gene includes the normal PCR, in which we are not only getting the FLA for the RP-PCR, but also full length of the complete amplicon that contains the repeat region. The result of this PCR corresponds to the high peak after the 24 CGG (~180bp, green arrow)

2.4 Discussion

Many neurodegenerative diseases are caused by an expanded RE motif, whose length varies between genes and between people. Similarly, RE expansions significantly influence human phenotypic variation and disease severity. For *ATXN2*, only the intermediate length of the RE (between 24 to 34 repeats) has been associated with increased ALS risk, while harbouring 34-49 has been associated with parkinsonism and more than 34 with SCA2. This RE demonstrates the different phenotypes that a gene can confer with different lengths of RE that may increase the risk of various neurodegenerative diseases. Hence, the identification and measurement of REs can improve our understanding of neurodegenerative diseases, significantly speed up patient diagnosis, and enhance patient care.

The primary aim of this project is based on the premise that different lengths of the repeat motif in a neurodegenerative gene may confer different phenotypes and contribute to the risk of developing other diseases, including ALS. Therefore, developing a low cost, reliable and reproducible methodology to genotype each RE from the NDRE genes in an ALS cohort will help identify the possible intermediate REs associated with increased risk.

Since the invention of PCR in the 1980s, this technique has been a cornerstone of molecular genetics, from the molecular cell biology to Sanger sequencing. PCR enables focus on individual loci instead of the whole genome and is an efficient and cheap technique for rapidly detecting mutations in the lab. On the other hand, next-generation sequencing has been improving so much that it has been considered the new gold standard for molecular analysis. However, this is not necessarily the case for long REs, as aligning sequence reads that derive from expanded alleles cannot be done accurately and unambiguously, thus impairing the ability to count the RE motif correctly. Furthermore, although whole exome sequencing and whole genome sequencing are procedurally straightforward and could in principle read small REs residing within sequence reads, sequencing more than 3,000 samples (the cohort size analysed in the next chapter) would be prohibitively expensive and likely still of inadequate accuracy.

For example, previous work carried out in the Complex Trait Genomics Laboratory, TCD (where the work detailed in this thesis was also carried out), using whole-genome sequences along with data from this thesis, showed that several 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, often could not identify large REs and frequently had a high number of false positives (Mark Doherty, PhD thesis, Trinity College Dublin, 2022). Even though these tools are promising, it was observed that the accuracy is gene-dependent, as seen in a different study by Rajan-Babu *et al.* (2021) with the *TBP* gene, indicating that these findings are not unique to the dataset (Rajan-Babu *et al.* 2021). These findings highlight the potential of our NDRE PCR panel to accurately genotype and classify REs in the general population and the importance of validating results to evade misdiagnosis.

For this project, preserving the DNA ALS bank was a priority. This DNA bank is unique in Ireland, is meant to be used for research purposes and is very limited since we cannot resample DNA from patients who have died. The second objective of this project was to find a protocol that can genotype many genes by using the lowest amount of DNA to preserve it for future work. During the development of the PCR panel, a high-performance polymerase was essential for correctly amplifying the high GC region using the lowest DNA concentration per sample. In the literature, numerous PCR protocols for genotyping these RE genes use at least 100-200ng per individual PCRs, around 5µg per patient. The multiplex PCR panel for 35 neurodegenerative genes uses a total of **412ng** of DNA per individual, around 92% less DNA usage than previously described protocols.

Based upon each NDRE gene and its genotype, defining each RE was based upon their discovery paper or, on some occasions, its most recent research. The idea of defining what the repeat motif is and its location can be challenging; there could be "interruptions" in the sequence or just a change of a base that biologically does not change the amino acid. The human genome has at least 3% repeat sequence, of which repeats often with a high GC content are found in genes related to neurodegeneration. The pathogenicity of these motifs comes from being expanded within a flanking area in the gene, not from a specific starting point. Therefore, the definition of each RE gene in this project could differ from other researchers.

Defining the number of repeats within normal, intermediate, and expanded alleles is challenging, and it changes within population-based studies throughout multiple available papers. For most of the genes, there is no consensus on what determines a repeat length as intermediate, or if such a threshold exists (Table 1-3). Also, across the literature, the cut-offs for normal and expanded alleles vary considerably. For ALS, only a couple of the NDRE genes in this study had some defining RE length associated with ALS. Hence, the thresholds of each gene repeat motif were built using the most recent research or discovery papers associated with their associated diseases (Table 1-3).

During the development of the NDRE PCR panel and genotyping of the *TK2-BEANI* gene, it was noticed that the TGGAA repeat insertion, previously associated with SCA37 (Sato et al. 2009), was not seen on any of the controls and patients that were sequenced. It is impossible to know by only PCR that the repeat count is from the insertion of the TAAAA repeat. Therefore, further analysis of these genotyping results in this thesis does not distinguish the TAAAA repeat from the TGGAA repeat insertion. Further work to establish the composition of any potentially pathogenic allele will require sequencing data.

The developed NDRE PCR panel can be used in the future by others and can also be updated with even more neurodegenerative RE. To introduce more genes into the panel is by matching their PCR conditions and to use this optimization chapter as a guidance to keep up to date new genes that harbour RE.

Early diagnosis of neurodegenerative diseases is vital for the diagnosis and new genetic-specific treatments for various conditions. The genotyping of RE is getting more popular, and several private or research laboratories offer a panel of several NDREs genes. Table 2-3 shows four different labs which offer this service, the panel cost, the number of genes included and the charge per gene. The most expensive panel is from The University of Chicago, which offers a panel of 13 NDRE genes for €1252, equating to €96.30 per individual gene. The cheapest panel in the market belongs to NHS Wales, which has a panel of 6 genes for €186 (€31 per gene). While comparing all these panels with our multiplex PCR genotyping protocol, the price for the 35 genes is **€50.05** (€1.43 per gene) when using a 96 well plate; if using a single tube (one

individual at the time), the full cost is **€226.45 (€6.47** per gene). The budget for the 3,000 samples and 35 genes between the developed multiplex panel and no panel was 82% less expensive (multiplex panel: €152,490 vs no panel with individual FLA: €846,495). When considering WGS for all the included individuals, the multiplex PCR panel is 95% more affordable than sequencing the whole genome (see Table 2-4 for more information). Taking all of this together, developing a multiplex PCR panel was the most valuable and affordable method to accurately measure intermediate RE from 35 neurodegenerative genes.

Table 2-3 Gene/cost comparison of multiple NDREs panels.

<i>Laboratory / company name</i>	<i>Cost per panel (€)</i>	<i>Genes included in panel</i>	<i>Cost per gene (€)</i>	<i>Reference</i>
DNA LABS INDIA	255	7	36.43	(Anon n.d.-f)
Victorian Clinical Genetics Services (VCGS)	500	6	83.33	(Anon n.d.-e)
NHS WALES	186	6	31	(Anon 2018)
The University of Chicago	1,252	13	96.30	(Anon n.d.-a)
Trinity College Dublin	226.45 (single tube) 50.05 (96 well)	35 35	6.47 1.43	Current developed panel

Table 2-4 Budget comparison of several protocols for the NDREs project.

<i>Genotyping method</i>	<i>Project costs (€)</i>
Whole Genome Sequencing (WGS)	≈ 1000 per individual/ <u>$\approx 3,000,000$ project</u>
Individual PCR panel, one gene per FLA run	846,495
Individual PCR panel, 4 genes per FLA run	427,350
Multiplex PCR panel, 4 genes per FLA run	<u>152,790</u>

2.5 Summary

The developed PCR panel is a practical and valuable methodology to accurately detect intermediate repeat motifs in 35 NDREs genes. The panel was created with the primary objective of preserving gDNA and the ability to detect multiple intermediate repeat expansions. Using PCR as the gold standard, the data obtained is precise and will be the base for various studies, such as bioinformatics, and will create a biological impact in the research of neurodegenerative diseases.

Chapter 3

Population genetic analysis of the genetic pleiotropy of neurodegenerative repeat expansions in ALS and FTD

3.1 Introduction

Recent studies have shown that different complex traits are genetically linked, including neurodegenerative diseases (Anttila et al. 2018). This correlation may partly be explained by pleiotropy (Sivakumaran et al. 2011), when variation at a given locus affects more than one trait, and it is one possible underpinning cause for cross-phenotype association (Solovieff et al. 2013). Since neurodegenerative conditions are often comorbid, with specific pairs of neurodegenerative conditions being genetically more similar than others, the analysis of phenotypically overlapping phenotypes has the potential to uncover new associations. Identification of pleiotropic loci may also point to shared biological mechanisms. Importantly, pleiotropic single nucleotide polymorphisms (SNPs) identified in genome-wide association studies are more likely to be genic (exonic or intronic) and less likely to be tissue-specific (Watanabe et al. 2019). Also, pleiotropic loci have the potential to serve as targets for interventions that contemporaneously help or treat multiple conditions. Various methods have been developed that exploit pleiotropic effects, and several analyses have been conducted to identify new inheritable associations in neurological diseases.

Heterogeneity of ALS sub-phenotypes

The site of onset of ALS refers to the location in the body where the first symptoms of ALS appear. ALS can begin in different parts of the body, and the site of onset can vary from person to person. The most common site of onset is in the limbs, typically in the hands or feet. This is known as limb-onset ALS or spinal onset ALS. About 70% of people with ALS experience limb-onset symptoms. Another type of onset is bulbar onset ALS, which affects the muscles used for speech, swallowing, and breathing. It is less common, accounting for about 30% of cases. Bulbar onset ALS can cause symptoms such as slurred speech, difficulty swallowing, and shortness of breath. Less commonly, ALS can begin in the respiratory muscles, or the muscles used for eye movement. These are known as respiratory onset and ocular onset ALS, respectively. Cognitive and behavioural changes can occur in some people with ALS, but cognitive onset is less commonly acknowledged, despite the relatively high prevalence of comorbid cognitive involvement in the disease. In particular, ALS-FTD accounts for about 15% of all cases of ALS, but the percentage can be higher in certain populations, such as people with a family history of the disease. In this form, the initial symptoms are related to changes in behaviour, personality, and/or thinking skills, rather than weakness or problems with movement (Testa and Jankovic 2019).

The phenomenon of ALS spreading from the initial site of onset to other parts of the body over time is a critical aspect to note in understanding the progressive nature of the disease. This is known as disease progression, and it can result in widespread muscle weakness and disability. Overall, the site of onset for ALS can vary from person to person, but the most common site is in the limbs, followed by the bulbar region (Testa and Jankovic 2019).

The age of onset is a crucial factor that influences the clinical phenotype and disease progression in ALS patients. The incidence of ALS is low among individuals below 40 years of age and rises significantly with increasing age. Sporadic ALS typically has a mean age of onset of 58-63 years, while familial ALS has a mean onset age ranging from 40-60 years (Ingre et al. 2015; Testa et al. 2004; Uenal et al. 2014). Despite significant advancements in understanding the genetics and underlying pathophysiology of ALS, the factors that contribute to the wide variability in age of onset remain unclear. A better understanding of the factors that influence

age of onset in ALS can provide important insights into the disease mechanisms, help predict disease progression, and identify therapeutic targets (Uenal et al. 2014).

The course of ALS can vary from person to person, but it is generally a progressive disease that worsens over time. The average survival time after onset of symptoms is approximately 2-5 years, although some people may live longer. A small percentage of people with ALS may survive for 10 years or more after onset of symptoms. The duration of the disease depends on several factors, including the age at which symptoms first appear, the site of onset, and the rate of disease progression. In general, people who develop symptoms at a younger age tend to have a longer disease duration than those who develop symptoms at an older age. Additionally, people with limb-onset ALS tend to have a longer disease duration than those with bulbar-onset ALS, which is associated with more rapid disease progression (Testa and Jankovic 2019; Uenal et al. 2014).

There is currently no cure for ALS, and treatments are primarily focused on managing symptoms and improving quality of life. Some treatments may slow the progression of the disease or provide temporary relief of symptoms, but they do not stop or reverse the underlying neurodegenerative process. In some cases, people with ALS may be eligible to participate in clinical trials of new treatments or therapies (Testa and Jankovic 2019; Uenal et al. 2014).

Common features between RE-associated diseases

Despite their diversity, RE diseases share many genetic features. For example, all REs arise from a healthy population of polymorphic repeats, and the degree of the polymorphism around the normal range varies between the diseases. The mutable normal repeats or “normal repeat range” have a higher tendency to expand during transmission, moving into the pathogenic range. This means that unaffected families can become affected if a de novo mutation occurs (Paulson 2018). Other general features that are common in diseases associated with REs include: 1) REs are unstable (dynamic), they often change size when transmitted to the following generation. 2) Longer repeats often cause the disease to be more severe and with earlier age of onset. 3) Clinical

anticipation is very common in successive generations, where the disease becomes more severe.

4). The overall phenotypes are variable, which can be associated with variation in repeat size.

Diseases caused by repeat expansions tend to cluster around certain phenotypes. For example, most RE disorders exhibit one or more of the following phenotypes: tremor, cerebellar ataxias, cognitive impairment, myoclonic seizures, muscular dystrophies, peripheral neuropathies, and chorea. Table 3-1 contains descriptive information on the most relevant NDREs genes in the present study, including disease pathway-related symptoms and association with ALS.

Research aims

The aim of this project is to explore the hypothesis that many rare REs causes a significant proportion of ALS cases. If many rare REs causes ALS, this would be a convergent genetic mechanism for the disease, since the disease would not be caused by only one or many genes, but instead by a particular mutation type, appearing on any transcript expressed in the central nervous system. These transcripts could be coding or noncoding RNA, which have been associated with ALS and can create a large amount of genetic variation in the genome for ALS. Furthermore, if the contribution of REs is substantial to ALS risk, the pathogenic mechanisms of the RE will be a convergent target for pharmaceutical intervention and treatments for patients.

Since hexanucleotide repeats in *C9orf72* causes ALS-FTD, and REs in *ATXN2* cause SCA2 and are risk factors for ALS, possibly other NDREs may exhibit similar pleiotropy, so the proposed study will investigate all 35 NDRE genes listed in Table 1-3 in a cohort of more than 1,500 ALS patients and over 1,700 controls from the Irish ALS DNA bank, to improve earlier underpowered studies that have examined some of the NDREs. The measure of all the 35 NDREs genes in Table 1-1 will be by implementing the previously described multiplex PCR genotyping panel from Chapter 2.

Table 3-1 Summary of the most relevant NDREs genes disorders for this project.

GENE	REPEAT MOTIF	DISEASE	RELATED DISEASE PATHWAYS	RELATED SYMPTOMS	PREVIOUSLY ASSOCIATED WITH ALS	REFERENCES
<i>AFF2</i>	CCG	FRAXE	May act as a transcription factor regulating other genes (unknown). Protein may control alternative splicing.	Learning disabilities, delayed speech, hyperactivity, short attention span, autistic behaviors.	Yes, expression of expanded AFF2/FMR2 regulates transcription and toxicity of <i>C9orf72</i> G4C2 repeats in C9ORF72-ALS/FTD neurons.	(Yuva-Aydemir et al. 2019)
<i>AR</i>	CAG	SBMA	Androgen receptor (AR) binds to DNA and regulates androgen-responsive genes. Expanded CAG region changes structure of AR, protein accumulates in cells and may interfere with normal cell functions.	Spinal and bulbar muscular atrophy. Atrophy in arms and legs resulting in cramping, leg muscle weakness.	No, but SBMA is often a differential diagnosis for ALS /Yes, AR antagonism accelerates disease onset in SOD1 mouse model of ALS	(McLeod et al. 2019; Monks et al. 2008; La Spada et al. 1991)
<i>ARX</i>	GCC	XLID	Transcription factor. ARX protein involved in migration and communication of neurons.	Intellectual disabilities, muscle stiffness, weak muscle tone, feeding difficulties, epilepsy.	No	(Kitamura et al. 2002)
<i>ATXN1</i>	CAG	SCA1	Ataxin-1 may be involved in the first stage of Transcription and processing of RNA. Expanded CAG in Ataxin-1 protein forms aggregates in the nucleus, damaging cells, specifically in Purkinje cells.	Numbness, pain or tingling in legs and arms, dystonia, atrophy, and muscle twitches. Rarely rigidity, involuntary movements, tremors.	Yes	(Lattante et al. 2018; Orr et al. 1993; Tazelaar et al. 2020)

<i>ATXN2</i>	CAG	SCA2	Ataxin-2 protein may be involved in processing RNA and play a role in productions of proteins from RNA. Abnormal protein with CAG expansions may lead to cell death.	Loss of limbs sensation and weakness, atrophy, dystonia, and involuntary jerking movements.	Yes	(Van Damme et al. 2011; Daoud et al. 2011; Groen et al. 2012; Ross et al. 2011)
<i>ATXN3</i>	CAG	SCA3	Expanded PolyQ Ataxin-3 aggregates reducing cellular Ataxin 3 content and impairing its functions in the cell.	Clumsiness in legs and arms, spasticity in lower limbs, difficulty swallowing and speech, lurching gait, and impaired eye movements.	GWAS results in 2021: Potential involvement in ALS. Ataxin 3 also interacts with valosin-containing protein (VCP/p97) which is a rare ALS gene.	(Doss-Pepe et al. 2003; Yuva-Aydemir et al. 2019)
<i>ATXN8OS/ATXN8</i>	CTA/CAG	SCA8	Expanded polyglutamine protein that forms a 1C2-positive inclusions in Purkinje cells and other neurons.	Progressive incoordination of hands, speech, and eye movements.	Yes	(Hirano et al. 2018)
<i>C9ORF72</i>	GGGGCC	FTD / ALS	Protein likely to be involved in processing RNA. Protein may influence the production, translation, and transport of RNA. Aberrant GGGGCC expanded protein disrupts RNA and protein production in the cell causing protein aggregates, leading to motor neuron cell death.	ALS: Progressive muscle weakness, inability to control movement and loss of muscle mass. FTD: progressive disorder affecting language, behaviour and personality.	Yes	(Chan 2014; Woollacott and Mead 2014; Xu et al. 2013)
<i>CNBP</i>	CCTG	DM2	Pathogenic expanded CUG sequesters the muscleblind-like protein (MBNL), causing alterations in the metabolism of multiple RNAs. Alterations on the MBNL splicing disrupts muscle development and function.	Muscle weakness and wasting. Primarily to proximal muscles, such as neck, elbows, shoulders, and hips. Prolonged muscle contractions (myotonia).	No	(Christina L. Liquori et al. 2001; Zu et al. 2017)
<i>FMR1</i>	CGG	FXS	FMR protein (FMRP) is present in several tissues (brain, ovaries, testes). In the brain,	Severe learning problems intellectual disability.	No	(Groen et al. 2012; Grosso et al. 2021;

	CGG	FXPOI	development of synapses between nerve cells. May help to regulate synaptic plasticity. Expanded CGG causes silencing of the gene.	Irregular menstrual cycles, early menopause, infertility, high FSH hormone.		Rajan-Babu et al. 2021)
	CGG	FXTAS		Progressive problems with movement (ataxia), tremors, peripheral neuropathy, memory loss, and mental-behavioral changes.		
<i>FRA10AC1</i>	CGG	Mental Retardation, FRA10AC1 type	Fragile site associated with cerebrospinal fluid $A\beta_{1-42}$ level.	Autism. Delayed development of speech and motor skills. ADHD.	No	(von Elsner et al. 2022; Li et al. 2015; Sarafidou et al. 2004)
<i>FRA11A</i>	CGG	Mental retardation, FRA11A	Expanded CGG may cause DNA hypermethylation in the C9orf80 gene and may be associated to reduce its expression.	Mentally retarded patients.	No	(Debacker et al. 2007)
<i>HTT</i>	CAG	HD	Nucleus-cytoplasm toxicity.	Trouble speaking, walking, and swallowing. Changes in personality, irritability, decline in thinking and reasoning abilities.	Yes	(Lieberman, Shakkottai, and Albin 2018; Manini et al. 2022)
<i>JPH3</i>	CTG	HDL2	Protein is a component of junctional complexes and mediates crosstalk between surface and intracellular iron channels of all excitable cell types.	Abnormalities of voluntary movements, involuntary movements, weight loss, psychiatric symptoms and dementia.	No	(Figley, Thomas, and Gitler 2014; Krause et al. 2015)
<i>NIPA1</i>	CGC/GCG	SPG6	Transport of inorganic cations/anions and amino acids/oligopeptides.	Difficulty with balance, weakness and stiffness in the legs, muscle spasms, and dragging the toes when walking.	Yes	(Blauw et al. 2012; Svenstrup et al. 2011; Tazelaar et al. 2019)
<i>PABPN1</i>	CGC	OPMD	Protein involved in polyadenylation, transport mRNA out of nucleus, regulating mRNA	Dysphasia, tongue atrophy, pneumonia.	No	(Blumen et al. 2009; Richard et al. 2015, 2017)

			production and breakdown of poor-quality mRNA. Extra alanines cause to form intranuclear inclusions in muscles cells, they accumulate and impair normal functions of muscle cells, eventually triggering cell death.			
<i>TK2- BEAN1</i>	TGGAA/T TCCA	SCA31	BEAN direction transcribed formed RNA foci in Purkinje cell nuclei in SCA31 patients. Degenerating Purkinje cells, abnormal dendrites, and somatic Purkinje cells.	Ataxia cerebellar dysarthria, limb ataxia, decreases muscle tone, mild to moderate bilateral sensorineural hearing loss.	No	(Ishiguro, Nagai, and Ishikawa 2021; Sakai et al. 2010; Sato et al. 2009)

3.2 Methods

Sampling

For this project, more than 1,500 ALS patients, 44 FTD patients and over 1,700 unrelated control subjects with no family history of ALS participated in this study. All samples were Irish and were drawn from the Irish ALS DNA bank. 11.9% of cases had familial ALS, defined as patients with at least one first or second degree relative also affected. This number increased to 16% when considering patients with any relative also affected. The sex and the average and standard deviation of the age for controls (when the samples were taken) and the age of onset for patients are shown in Table 3-2. For ALS patients, data for site and age of onset along with disease duration were obtained from the Irish ALS register.

Table 3-2 General demographics

<i>Sample</i>	<i>Sex</i>	<i>Age</i> [†] (years old ± Std)		
		Female (%)	Male (%)	
<i>ALS</i> (<i>n</i> =1550)		40.07	59.93	61.72 ± 11.55
<i>FTD</i> (<i>n</i> =44)		54.55	45.45	64.23 ± 8.32
<i>Control</i> (<i>n</i> =1719)		48.42	51.58	63.14 ± 10.3

Abbreviations: n: is the number of individuals; Std: is the standard deviation.

[†] Age at onset for ALS and FTD patients; age at sampling for controls.

PCR Genotyping

The protocol for genotyping all 35 NDREs genes is described in Chapter 2. Briefly, the panel consists of 16 different multiplex PCRs, which includes all 35 genes. After the PCR is completed,

the product of the PCRs were organized in 10 different panels of maximum 4 genes per set and sent for FLA at Eurofins Medigenomix GmbH, Ebersberg in Germany. All the conditions for each combination of PCR-FLA of the NDRE panel can be found in Supplementary Table S 2-3 and Supplementary Figures Chapter 2.

The resulting electropherogram had fragment lengths automatically called in bp by in-house software at Eurofins Medigenomix GmbH using Gene Mapper software. Each individual electropherogram was revised and bp data was collected in an Excel sheet. The allele length of each repeat expansion peak is converted into number of repeats using the Equation 2-2:

$$\# \text{ of repeats} = \left[\text{common repeat count} \times \left(\frac{\text{Allele length (bp)} - \text{common repeat length (bp)}}{\text{repeat size (bp)}} \right) \right] + (\text{Repeat modification})$$

Common repeat count: number of repeats present in the reference genome; Common repeat length: size of the in-silico PCR from the reference genome; Allele length: PCR fragment length obtained by CE; Repeat size: repeat motif length; Repeat modification: number of extra repeats obtained in the repeat count calibration (consensus modification).

Ultimately, the obtained number of repeats obtained by the formula went through a repeat count calibration (explained in detail in Chapter 2, Section 0 Repeat count calibration) to ensure the most accurate number of allele repeats.

Statistical analysis and plotting

Each gene was analysed using a thresholded approach, following previous works (Elden et al. 2010; Tazelaar et al. 2020) . Considering the longer allele per patient, each STR length was taken as a threshold and the population of case-control data was analysed by comparing individuals below this threshold vs above or at the threshold. Odds ratios (OR) for ALS risk were calculated as follows:

		Case/control status	
		Case	Control
Allele vs threshold	$\geq$	a	b
	$<$	c	d

$$OR = \frac{(n \text{ cases} \geq \text{threshold}) \times (n \text{ controls} < \text{threshold})}{(n \text{ cases} < \text{threshold}) \times (n \text{ controls} \geq \text{threshold})}$$

$$OR = \frac{ad}{cb}$$

Confidence intervals (CI):

$$\text{Upper 95\% CI} = e^{\ln(OR) + 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

$$\text{Lower 95\% CI} = e^{\ln(OR) - 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

As well as calculating odds ratios, case/control counts and site of onset were analysed using Fisher's exact test. Age at onset and disease duration were analysed using the Mann-Whitney-Wilcoxon test. All statistical tests compared counts or distributions for data below each threshold length vs above or at the threshold.

Data were simultaneously analysed and plotted using an in-house script (Appendix B) in RStudio version 2022.12.0+353. This permitted a standardised visualisation of the data, with the plots containing the following elements:

1. Carrier frequency plot: schematic representation of the carrier frequency (long or short allele) of each obtained repeat count for the studied populations (ALS/FTD vs. controls).
2. OR plot: A representative $\log_{10}(OR) \pm 95\% \text{ CI}$ analysis for each repeat count and above (count $\geq$), thresholded by each observed allele as a lower limit. Fisher exact test p -value labels are included to estimate the association between the repeat expansion length count and above (repeat count $\geq$) and patient/control status.

3. Site of onset stacked bar plot: this figure shows the distribution of the site of onset categories (%) per each repeat count and above (count $\geq$). Fisher exact test p -value labels are included to estimate the association between the repeat expansion length count and above (repeat count $\geq$) and site of onset.

4. Disease duration box-plot and 5) Age of onset box-plot: these show distributions of disease duration (in months) and age of onset (in years) for ALS patients per each repeat count and above (count $\geq$), following standard conventions: boxes show the range of the lower and upper quartile; whiskers show the extent of the most extreme values $< 1.5 \times$ the interquartile range from the boxes; outlier points were not plotted. Mann-Whitney-Wilcoxon p -value labels were included to test whether the observed distributions for data $\geq$ each threshold were significantly different to (unplotted) data $<$ the threshold (as opposed to chance variation within the same distributions).

Reported p -values consider a critical α of 0.05 as statistically significant in the narrative accompanying the results; however, it should be noted that a considerable multiple testing burden exists for the data produced by these experiments. The decision to maintain an α value of 0.05 reflects the substantial non-independence of tests when using thresholded statistics as described above, as well as the likely partial correlation between overlapping clinical features such as age at onset and disease duration. Where such descriptions have been used, the underlying p -value upon which these descriptions are based are also reported for completeness.

3.3 Results

Genotyping results (ALS)

Within the scope of our investigation encompassing 35 NDRE genes, a cohort comprising 1550 ALS cases and 1719 control subjects underwent PCR amplification. It is imperative to note that, during the subsequent genotyping analysis, a subset of PCRs experienced technical challenges, resulting in varying genotyped population sizes for each gene within our comprehensive panel.

Short Allele

Although this study focuses on alleles with intermediate length and more extended, short alleles were also observed and analysed. These alleles allow us to observe the baseline distribution of each gene in the entire population and to rule out some errors in obtaining the results. For each gene, a representative graph of the allele frequencies for ALS cases and controls for the shorter allele observed per individual can be seen in Supplementary Figure S 3-1. The baseline distribution could be observed for all genes, and the frequencies of the most common alleles for both populations correspond to the literature.

Longer repeat allele vs. ALS risk

This section presents the results of the longer allele for 35 NDREs genes using the multiplex PCR-FLA protocol. The obtained results were classified into five categories by their gene results with ALS risk.: previously reported ALS risk genes, novel ALS risk genes, probable ALS risk genes, possible ALS risk genes and genes not associated with ALS risk genes

Previously reported ALS risk genes

Once all the data for each of the genes were analysed, it was observed that in the Irish population, similar results were observed for five genes that have previously been associated with increased risk of ALS in other populations. The results between Figure 3-1 and Figure 3-5 show the results of *ATXN1*, *ATXN2*, *C9orf72*, *HTT* and *NIPA1*, including an explanatory paragraph of each result.

All statistical results are in Supplementary Table S 3-2 (ORs and proportions), Supplementary Table S 3-3 (site of onset statistical results) and Supplementary Table S 3-4 (statistical results for age of onset and disease duration).

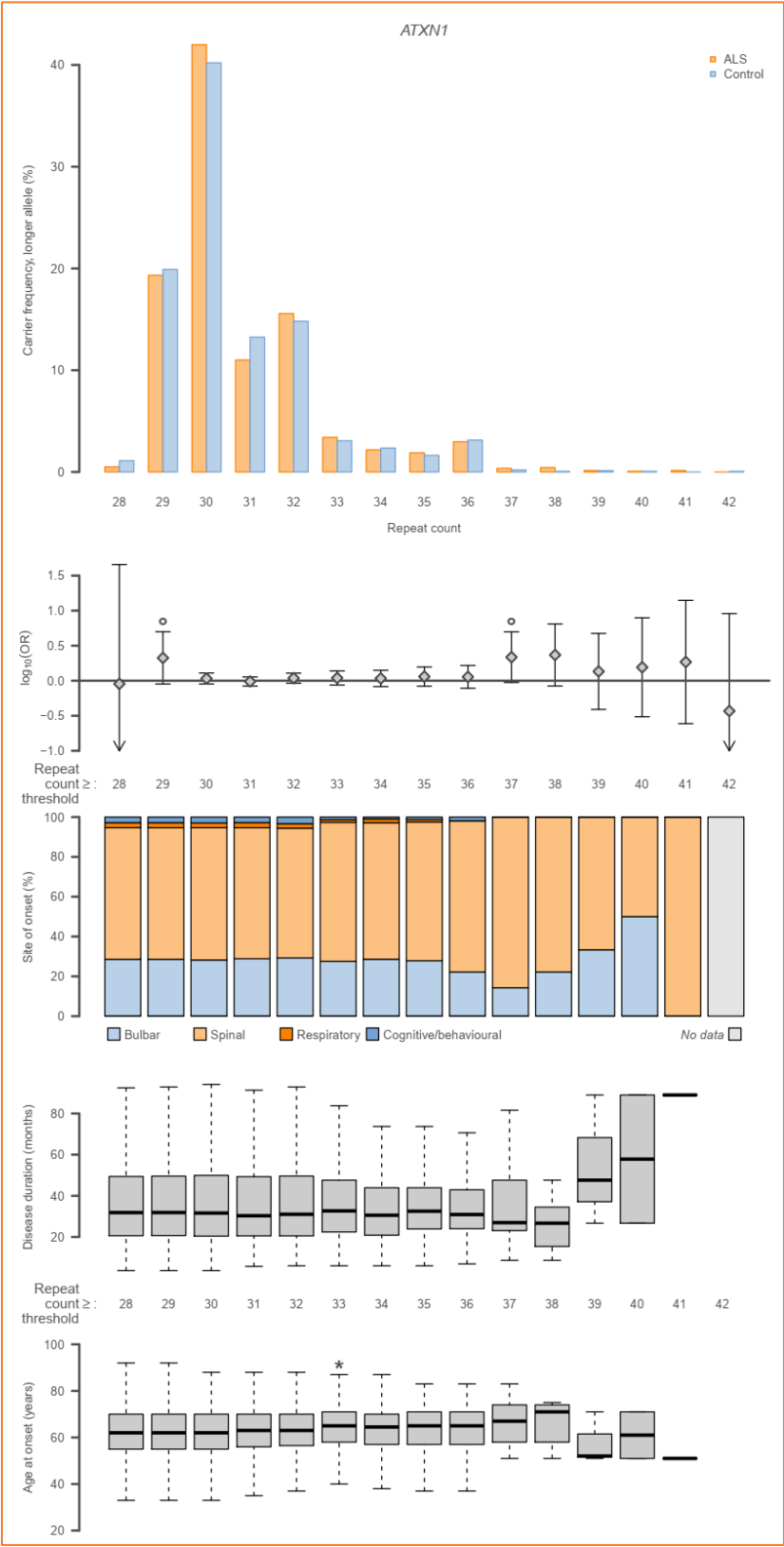


Figure 3-1 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *ATXN1* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-1 shows genotyping results for the *ATXN1* CAG repeat expansion in a large sample of ALS and control individuals. The most frequent allele was 30 repeats, followed by 29 and 32 repeats, with a higher percentage of CAG repeats at 30 and 32 observed in the ALS cases. Our analysis using Fisher exact test to estimate the association between ALS case/control status revealed that the probability that ALS and controls were sampled from the same population did not statistically signify any difference between cases and controls with the current sample size (e.g., p -values for ≥ 30 , ≥ 32 , and ≥ 37 CAG repeats: 0.435, 0.332 and 0.066, respectively). Furthermore, we investigated the association between CAG repeat length and ALS risk by calculating odds ratios (ORs) for various repeat counts. Our results indicate that there is no increase in ALS risk when harbouring a normal CAG repeat length between ≥ 28 and ≥ 35 . However, an increased ALS risk of 2 times was observed for individuals with CAG repeats of ≥ 29 and ≥ 37 (OR ≥ 29 = 2.12, 95% CI 0.90 – 4.99; OR ≥ 37 = 2.16, 95% CI 0.944 - 4.97). Although these results are not statistically significant, they provide important insights into the potential role of CAG repeat expansion in the pathogenesis of ALS. In addition, our analysis of the site of onset bar plot suggests that ALS cases with CAG repeats have a higher percentage of spinal onset across all repeats, with a potential increase in bulbar onset observed for individuals with 38 or more repeats, although this finding was not statistically significant. Furthermore, we found no significant changes in disease duration or age of onset when harbouring intermediate CAG repeats in the *ATXN1* gene (31.87 months). However, individuals with more than 37 repeats exhibited a decrease in disease duration by 1.9 months, although this was not statistically significant. A significant ($p < 0.014$) increase in age at onset by 5 years when harbouring more than 33 repeats was observed. Overall, these findings contribute to our understanding of the potential genetic basis of ALS and provide important insights into the potential role of CAG repeat expansion in this disease. Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

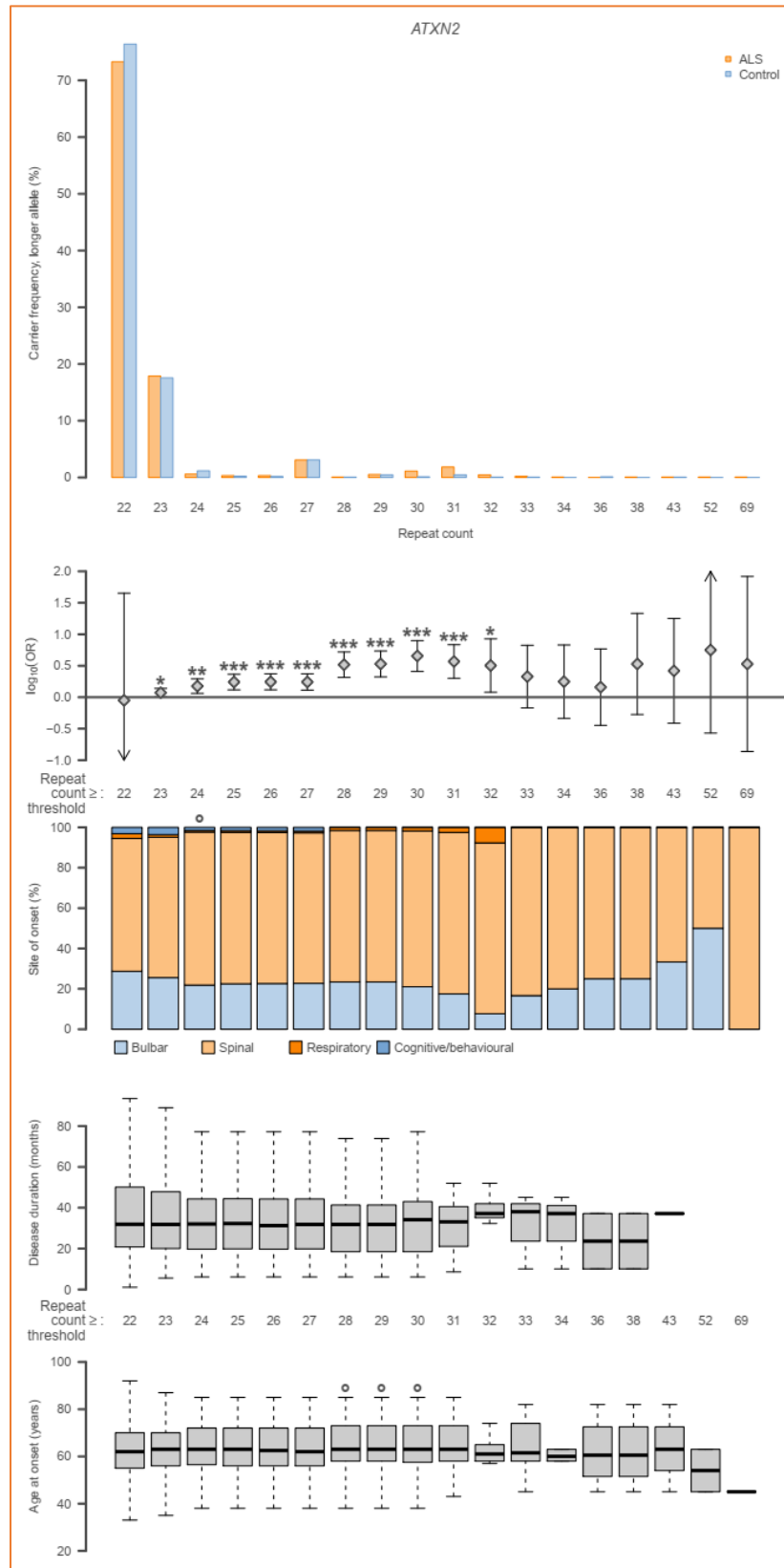


Figure 3-2 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *ATXN2* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-2 presents the results of genotyping 1527 ALS and 1715 control individuals for the *ATXN2* CAG repeat expansion. The analysis revealed that the highest frequency of the PolyQ repeat in both populations was for 22 repeats (73.28% ALS and 76.39% Controls) and 23 repeats (17.88% ALS and 17.55% Controls). Statistical analysis using Fisher's exact test indicated that ALS and controls are significantly different in terms of their repeat count for all categories from ≥ 26 to ≥ 32 (p -values: 1.39×10^{-4} , 1.95×10^{-4} , 1.22×10^{-7} , 8.74×10^{-8} , 5.49×10^{-9} , 8.73×10^{-6} , respectively). The second panel shows ORs calculated for the range of more or equal to a specific CAG repeat count. The analysis shows a significant increase in ALS risk when harbouring intermediate repeat lengths of the PolyQ expansion in the *ATXN2* gene (between 24 to 34 repeats). The highest obtained OR was when harbouring more than 30 CAG, which increased ALS risk by 4.5 times (OR=4.52, 95% CI 2.58-7.94). The site of onset bar plot shows that ALS cases with CAG repeats have a higher percentage of spinal onset across all repeats followed by bulbar onset. There was a non-significant increase in respiratory ALS with 32 or more repeats. Disease duration did not show significant variation across all repeats (approximately 31.9 months), while interestingly, the mean of survival increased slightly by 5.34 months when harbouring 32 repeats or more, although this was not statistically significant. Age at onset showed a decrease by 1 year when harbouring more than 28, 29 and 30 CAG repeats, which was almost statistically significant with the current sample size (p -value: 0.051, 0.051 and 0.073, respectively). Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

C9orf72

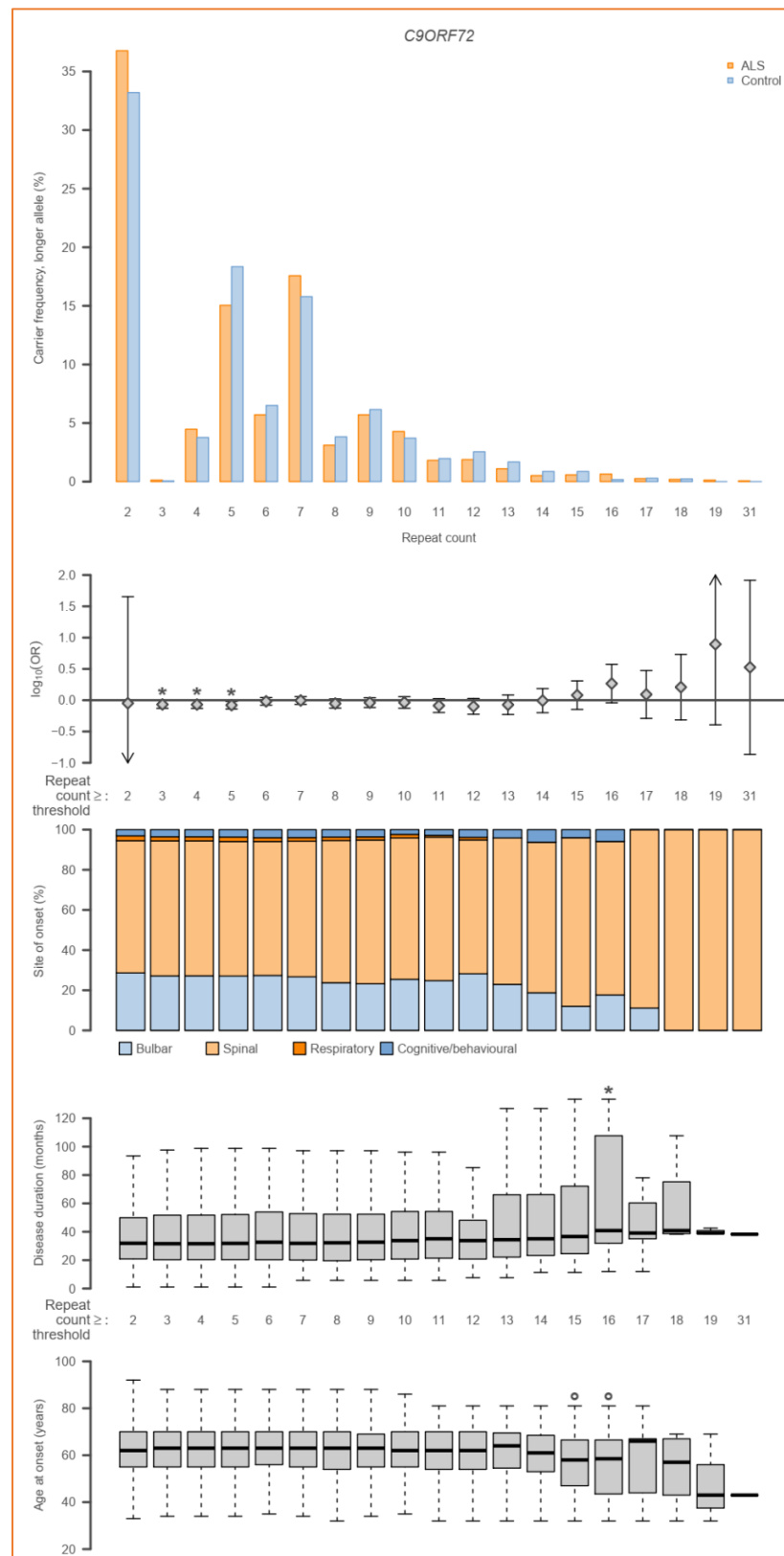


Figure 3-3 The composition of the population of ALS and Control samples for the GGGGCC repeat expansion in the *C9orf72* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-3 provides a detailed overview of the 1542 ALS and 1723 control individuals genotyped using conventional PCR for the *C9orf72* GGGGCC repeat expansion. The results show that the highest longer allele for both populations were 2 repeats (36.77% ALS and 33.20% Controls) followed by 7 repeats (17.58% ALS and 15.79% Controls). Fisher's exact test was used to estimate the association between ALS case/control status. For individuals with more than 3, 4, and 5 repeats, the probability that ALS and controls are sampled from the same population is 0.033, 0.030, and 0.012, respectively. This indicates that both populations are significantly different in terms of their repeat count (p -values: 0.033, 0.030, and 0.012, respectively). On the other hand, for individuals harbouring 6 to 31 repeats, both populations are not significantly different (p -values > 0.05). The ORs calculated across the repeat counts found in our study for the *C9orf72* gene show an increase in ALS risk for intermediate repeats between ≥ 15 to ≥ 20 (Supplementary Table S 3-2). Specifically, harbouring ≥ 19 GGGGCC repeats in the *C9orf72* gene increases ALS risk by 7.8 times (OR=7.84, 95% CI 0.41-151.84), although these results have wide confidence intervals and are thus not statistically significantly different from an odds ratio of 1. It is also notable that this analysis excludes repeat-primed PCR results of *C9orf72*, meaning that some substantially expanded cases are not included in these statistics. The site of onset bar plot indicates that ALS cases with GGGGCC repeats have a higher percentage of spinal onset across all repeats, followed by bulbar onset. There is an increase in disease duration of 8.87 months when harbouring 16 or more repeats, which is statistically significant (p -value: 0.026). This suggests that the increase in disease duration is due to an effect of harbouring 16 or more intermediate GGGGCC repeats in the *C9orf72* gene. Interestingly, the age at onset decreases by around 5 years when harbouring 15, and 16 repeats from 63 to 58 years, although these results are not statistically significant (p -value: 0.069 and 0.099, respectively). Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

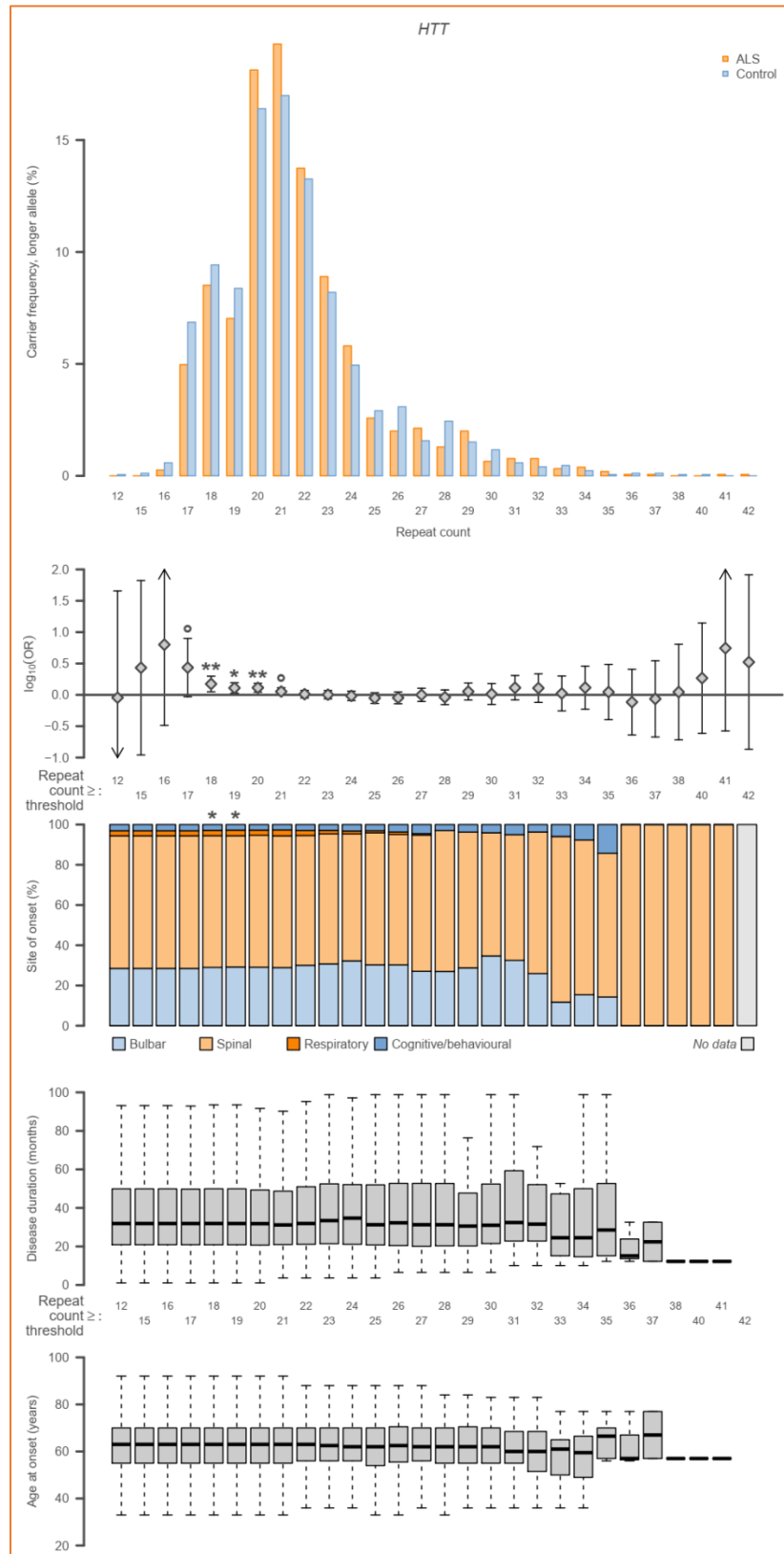


Figure 3-4 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *HTT* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-4 provides a detailed representation of the *HTT* CAG repeat expansion in 1550 ALS and 1719 control individuals. Our findings indicate that the most common longer allele for both populations were 20 and 21 repeats. Carriers between 20 and 24 repeats had a higher allele frequency in ALS cases than controls. Fisher's exact test revealed a significant association between ALS case/control status and the PolyQ carriers with 18, 19, and 20 repeats or more (p -values: 0.006, 0.01, and 0.002, respectively). Conversely, for repeats with ≥ 17 and between ≥ 22 - ≥ 42 , both populations were not significantly different. The second panel displays the odds ratios (ORs) across the repeat counts found in our study for the *HTT* gene. We observed a modestly increased risk of ALS when harbouring intermediate-length repeats between ≥ 29 - ≥ 35 , and ≥ 38 CAG repeats in the *HTT* gene. The highest increase of ALS risk (30%) was found at ≥ 31 CAG repeats (OR=1.30, 95% CI 0.083-2.03, not statistically significant). For pathogenic-length repeats for HD (≥ 39 repeats), our results suggest an increase in ALS risk of up to 5.5 times when harbouring ≥ 41 CAG repeats in the *HTT* gene, although this was not statistically significant (OR=5.55, 95% CI 0.266-115.74). Our site of onset bar plot revealed that ALS cases with CAG repeats have a higher percentage of spinal onset across all repeats followed by bulbar onset. Although, when considering more than 33, 34 and 35 CAG repeats, cognitive/behavioural onset increases but results are not statistically significant (p -value: 0.3). We also found that disease duration decreases by 7.4 months when harbouring 33 or more repeats, not statistically significant. Lastly, age at onset appeared to increase by 4 years when harbouring more than 35 repeats in the *HTT* gene, which was also not statistically significant (median age of onset for <35 : 62.5 years, ≥ 35 : 66.5 years; p -value 0.45). Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

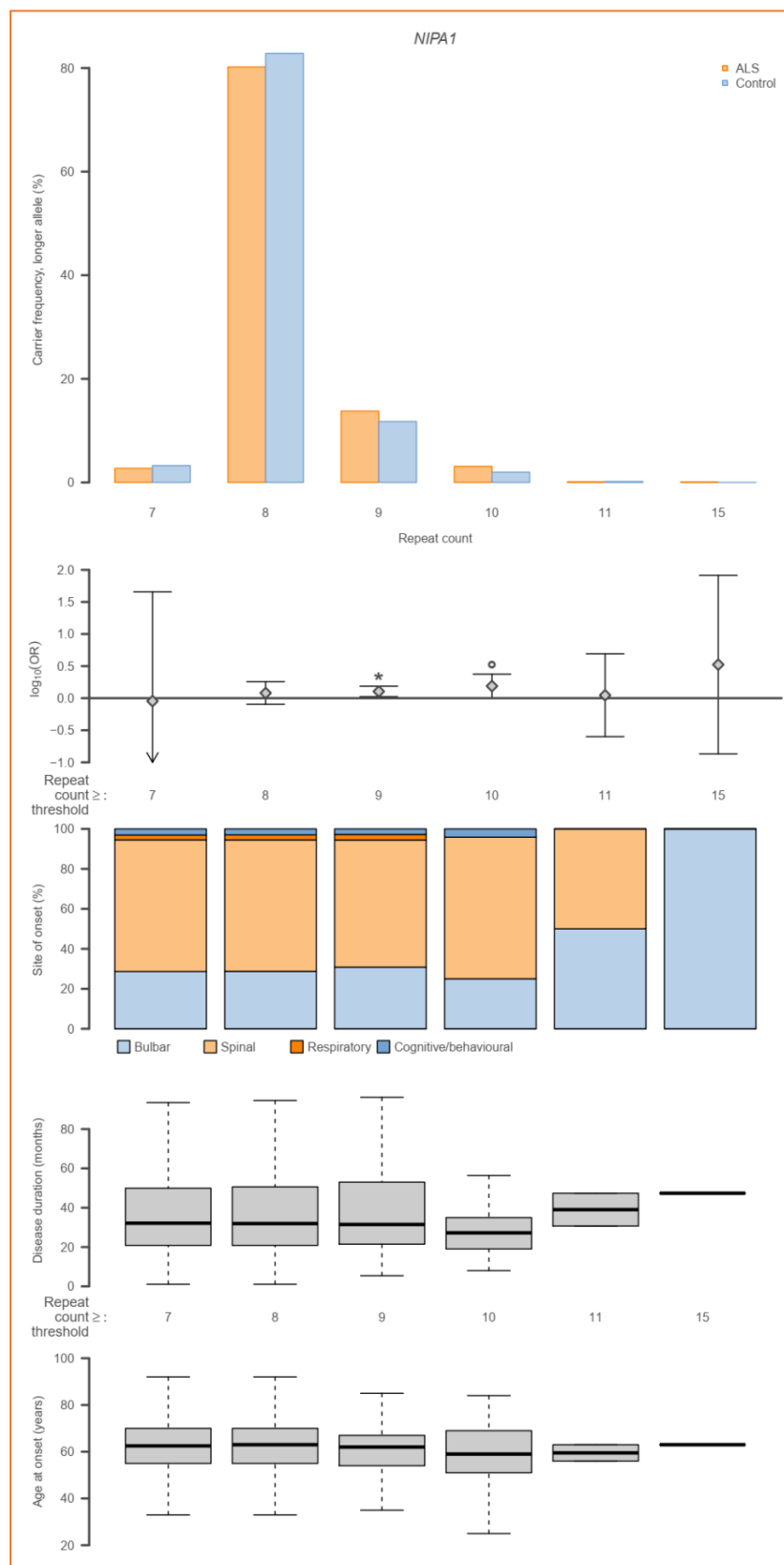


Figure 3-5 The composition of the population of ALS and Control samples for the CGC repeat expansion in the *NIPA1* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-5 shows genotyping results for 1546 ALS cases and 1718 controls with respect to the *NIPA1* CGC repeat expansion. Our analysis revealed that the most frequent longer allele for both populations was 8 repeats, with no significant difference between ALS cases and controls. However, carriers with 9 and 10 repeats had a higher allele frequency in ALS cases than in controls, and Fisher exact test indicated that the two populations are significantly different with respect to these repeats. In particular, for carriers with 9 repeats or greater, the *p*-value was 0.013. The odds ratio (OR) analysis demonstrated that carrying more than 9 CGC repeats in the *NIPA1* gene is associated with a 27% increase in ALS risk, and carrying more than 10 repeats leads to a 54% increase, which is nearly statistically significant. The bar plot for site of onset indicated that spinal onset is more frequent than bulbar onset for all repeats, with no respiratory onset observed in cases with more than 10 repeats. Disease duration did not vary significantly across all repeats, with a mean of around 32 months. However, for carriers with more than 10 repeats, disease duration decreased by 6.89 months, although the *p*-value was not statistically significant (0.627). Age at onset decreases by 3 years when considering more than 11 repeats in *NIPA1* gene, but this was not statistically significant (median age of onset for <11: 62.5 years, ≥11: 59.5 years; *p*-value 0.618). Overall, these findings suggest that *NIPA1* CGC repeat expansion is associated with ALS risk and site of onset, but not disease duration or age of onset. Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

Novel ALS risk genes

During the analysis of the longer allele of each NDRE gene, it was observed that 3 of them provided sufficient evidence of increased ALS risk in the Irish population to be considered potentially novel ALS risk genes. For these genes (*ARX*, *PABPN1* and *FRA10AC1*) there are no previous studies that have shown an association between their repeat expansion and ALS risk. Figures 3.7 to 3.9 shows the results of the longer repeat allele of these novel genes.

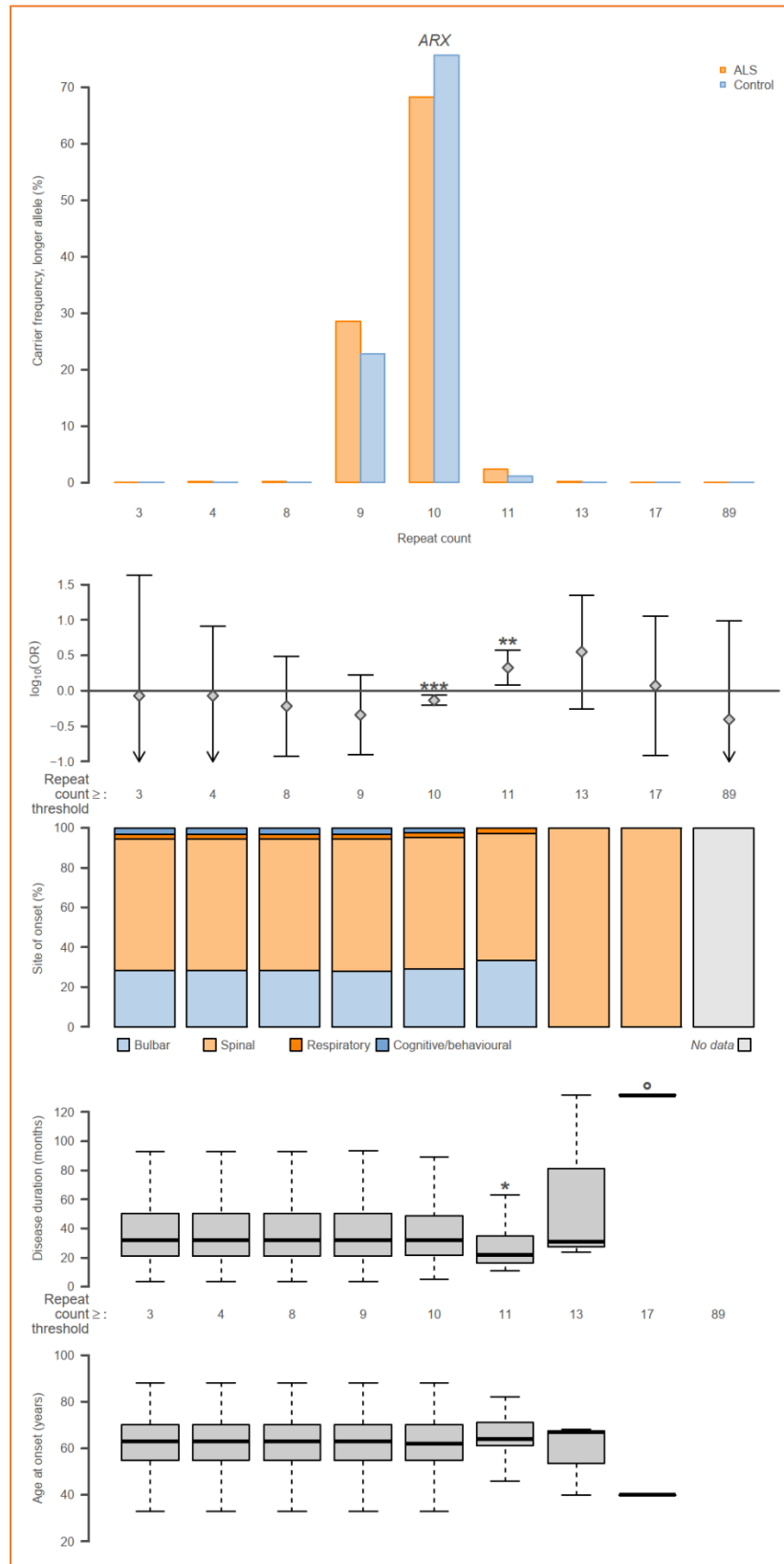


Figure 3-6 The composition of the population of ALS and Control samples for the CGC repeat expansion in the *ARX* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-6 provides genotyping results of 1288 ALS and 1518 control individuals for the *ARX* CGC repeat expansion. The analysis revealed that the highest longer allele for both populations were 10 repeats (68.22% ALS, and 75.69% controls), followed by 9 repeats (28.57% ALS, and 22.86% controls). Our protocol successfully amplified pure expanded polyalanine repeats (>23 repeats) and used Fisher exact test to estimate the association between ALS case/control status. For most repeat categories, ALS and controls were not significantly different. However, when considering 10 and 11 CGC repeats, there were significant differences between ALS and controls (p -values: 3.26×10^{-4} , and 7.97×10^{-3} , respectively). Notably, when considering ≥ 11 repeats, the OR suggests a significant increase in ALS risk of 2.12 times (OR= 2.12, 95% CI 1.21-3.70). Surprisingly, our data revealed a possible protective effect against ALS of 73% when harbouring ≥ 10 CGC repeats, which was statistically significant (OR=0.73, 95% CI 0.618-0.87). Multiple scientists have investigated the alanine tract in *ARX* over the years, and some published the pure CGC RE and others the CGN tracts. Either way, all are measuring the alanine tracts in the second exon. The conversion from pure CGC to CGN is six extra CGN to add to the pure tract. Our results can also be interpreted as harbouring more than 17 GCN repeats increase the risk of ALS by 2.12 times, while 16 GCN suggests a protective effect against ALS of 73%

While site of onset showed no significant variation, our analysis revealed a significant decrease in disease duration when harbouring more than 11 repeats, reducing the life of ALS patients by 10.32 months (p -value: 0.011). Moreover, when harbouring more than 11 repeats, age at onset increased by 2 years, although this was not statistically significant (p -value: 0.124). In summary, our findings suggest that intermediate CGC repeats increase the risk of ALS, while 10 CGC repeats may be protective. However, a significant increase in ALS risk was observed for individuals with ≥ 11 CGC repeats, leading to reduced life expectancy in ALS patients. Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4

FRA10AC1

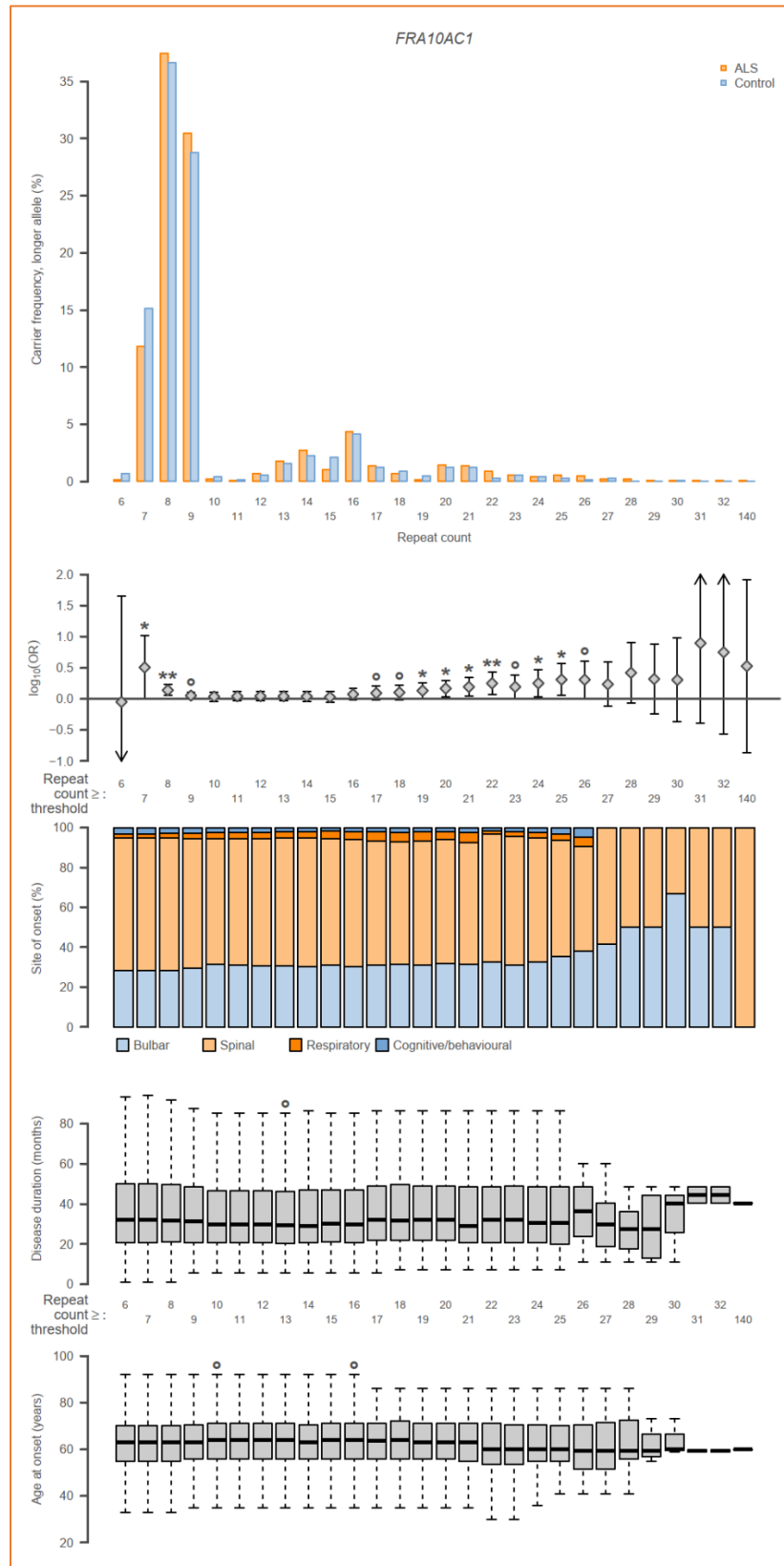


Figure 3-7 The composition of the population of ALS and Control samples for the CGG repeat expansion in the *FRA10AC1* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-7 provides genotyping results for the *FRA10AC1* CGG repeat expansion in 1530 ALS cases and 1714 controls. We found that the most common allele size for both populations was 8 repeats (37.45% ALS, and 36.64% controls), followed by 9 repeats (30.46% ALS, and 28.76% controls), with no expanded (>200) repeats detected. An intermediate repeat range between ≥ 17 and ≥ 24 showed a statistically significant increase in ALS risk, with odds ratios ranging from 1.23 (95% CI 0.96-1.59; $p = 0.096$) for ≥ 17 repeats to 1.78 (95% CI 1.07-2.97; $p = 0.029$) for ≥ 24 repeats, with a lowest p -value of 0.006 for ≥ 22 repeats. Site of onset and disease duration were consistent across repeat thresholds. However, we observed a non-statistically significant decrease in disease duration by 2.97 months and an increase in age at onset by 1 year for individuals with more than 21 CGG repeats. Our findings are summarized in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4. Overall, our study provides new insights into the association between *FRA10AC1* CGG repeat expansion and ALS risk, highlighting the importance of intermediate repeat ranges in ALS pathogenesis.

PABPN1

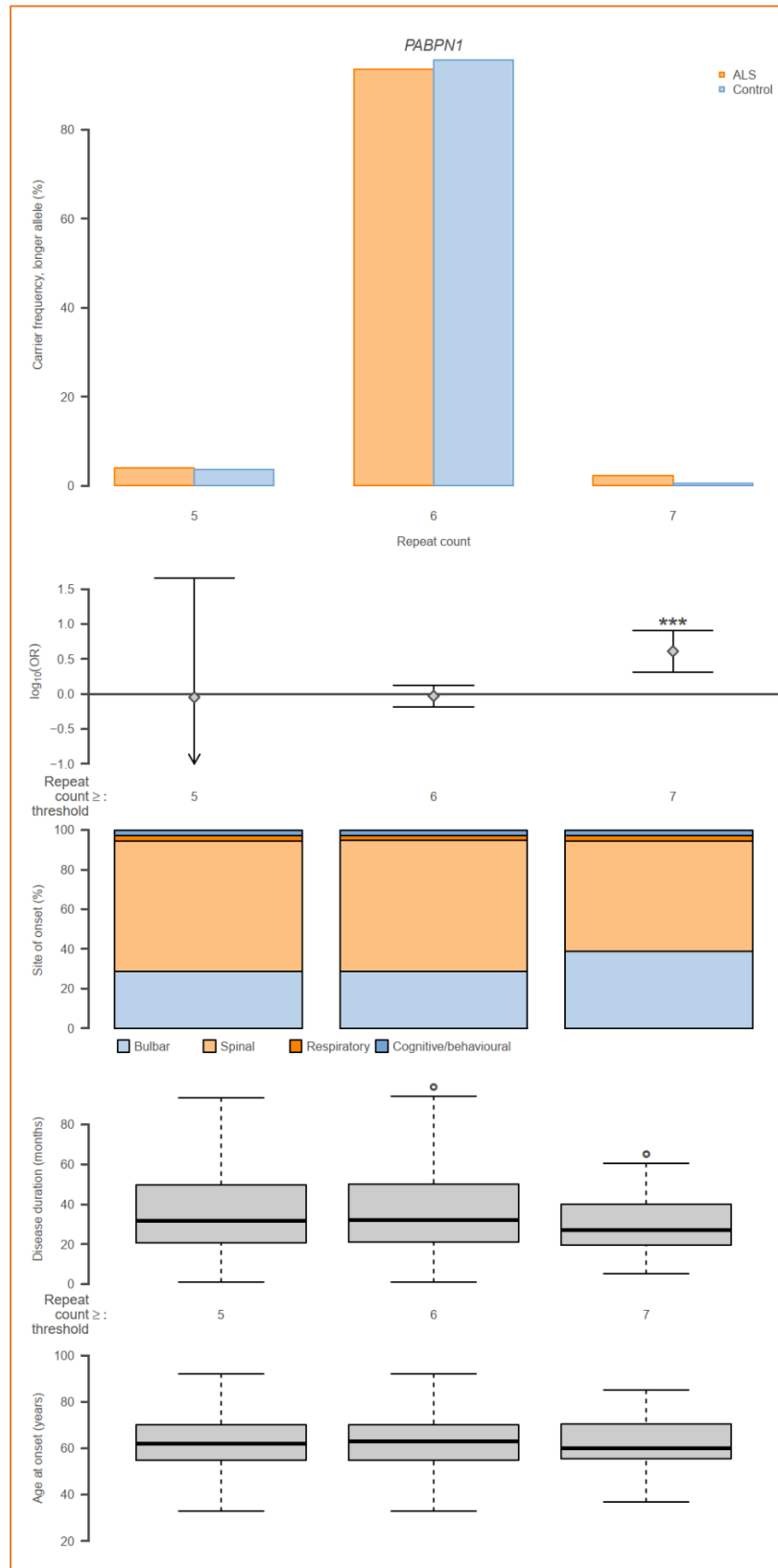


Figure 3-8 The composition of the population of ALS and Control samples for the GCG repeat expansion in the *PABPN1* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-8 presents genotyping results from 1543 ALS and 1726 control individuals for the *PABPN1* GCG repeat expansion. The analysis revealed that the highest longer allele for both populations was 6 repeats (93.58% ALS and 95.66% controls). Using Fisher's exact test, we found a statistically significant difference between ALS and controls for harbouring 7 GCG repeats, with a probability of 1.13×10^{-5} . This means that the *PABPN1* GCG repeat count is a significant contributor to ALS risk, with a 4-fold increase in ALS risk for ≥ 7 repeats compared to those with < 7 repeats (OR= 4.07, 95% CI 2.05-8.10). For patients with more than 7 GCG repeats, there was an increase in bulbar onset, although this was not statistically significant. The mean disease duration decreased by 4.96 months when patients had more than 7 repeats, which was almost statistically significant (p -value: 0.071). There was also a decrease of 3 years in age at onset when patients had more than 7 repeats, although, again, this was not statistically significant. Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4. Overall, these findings suggest that the *PABPN1* GCG repeat count is a crucial factor in predicting the risk of developing ALS.

Probable ALS risk genes

We designated nine genes as probable ALS risk genes based on either marginal or borderline statistical support for increased risk of ALS, or evidence for a contribution of expanded alleles to specific ALS sub-phenotypes. Figures 3.10 to 3.18 shows the results of the longer repeat allele of the following 9 genes: *AFF2*, *AR*, *ATXN3*, *ATXN8*, *CNBP*, *FMR1*, *FRA11A*, *JPH3*, and *TK2-BEANI*.

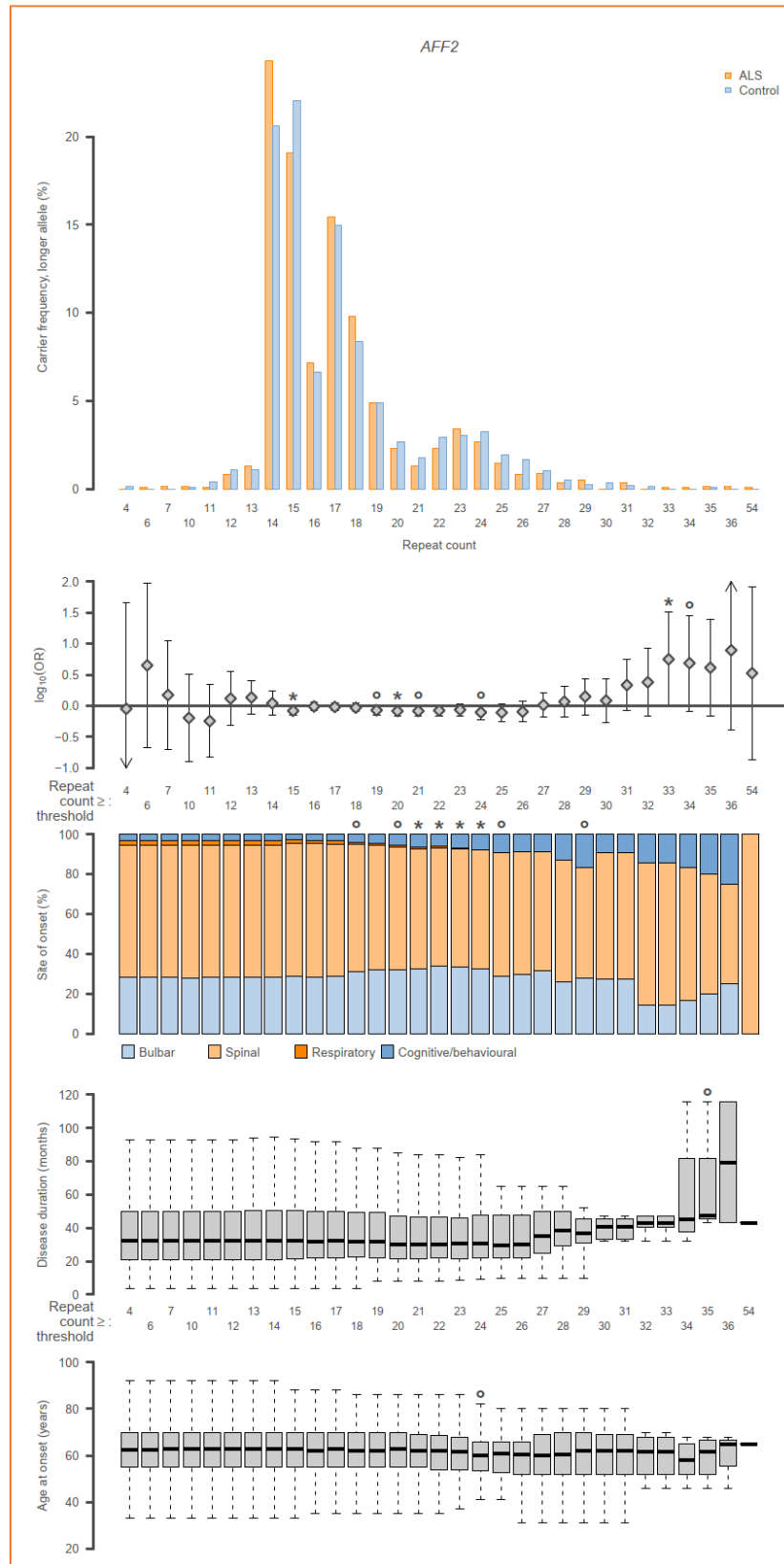


Figure 3-9 The composition of the population of ALS and Control samples for the CCG repeat expansion in the *AFF2* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-9 shows genotyping results for the *AFF2* CCG repeat expansion in 1387 ALS patients and 1549 control individuals. The highest longer allele frequency for both populations was 14 repeats (24.30% ALS, and 20.59% Controls), followed by 15 repeats (19.11% ALS, and 22.01% Controls). This protocol amplified longer repeats than the normal range (54). Fisher's exact test provided significant support for an ALS risk allele range ≥ 33 repeats ($p = 0.031$), with an associated odds ratio 5.61 (95% CI 0.97-32.42). The proportion of cognitive-onset ALS increased significantly from 21 repeats upwards. Disease duration did not vary significantly across all repeats, although the distribution qualitatively shifts upwards for higher repeat counts, with marginal support ($p = 0.086$) at ≥ 35 repeats. Age at onset decreased by 3 years when harbouring more than 24 repeats, although this finding was not statistically significant. Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

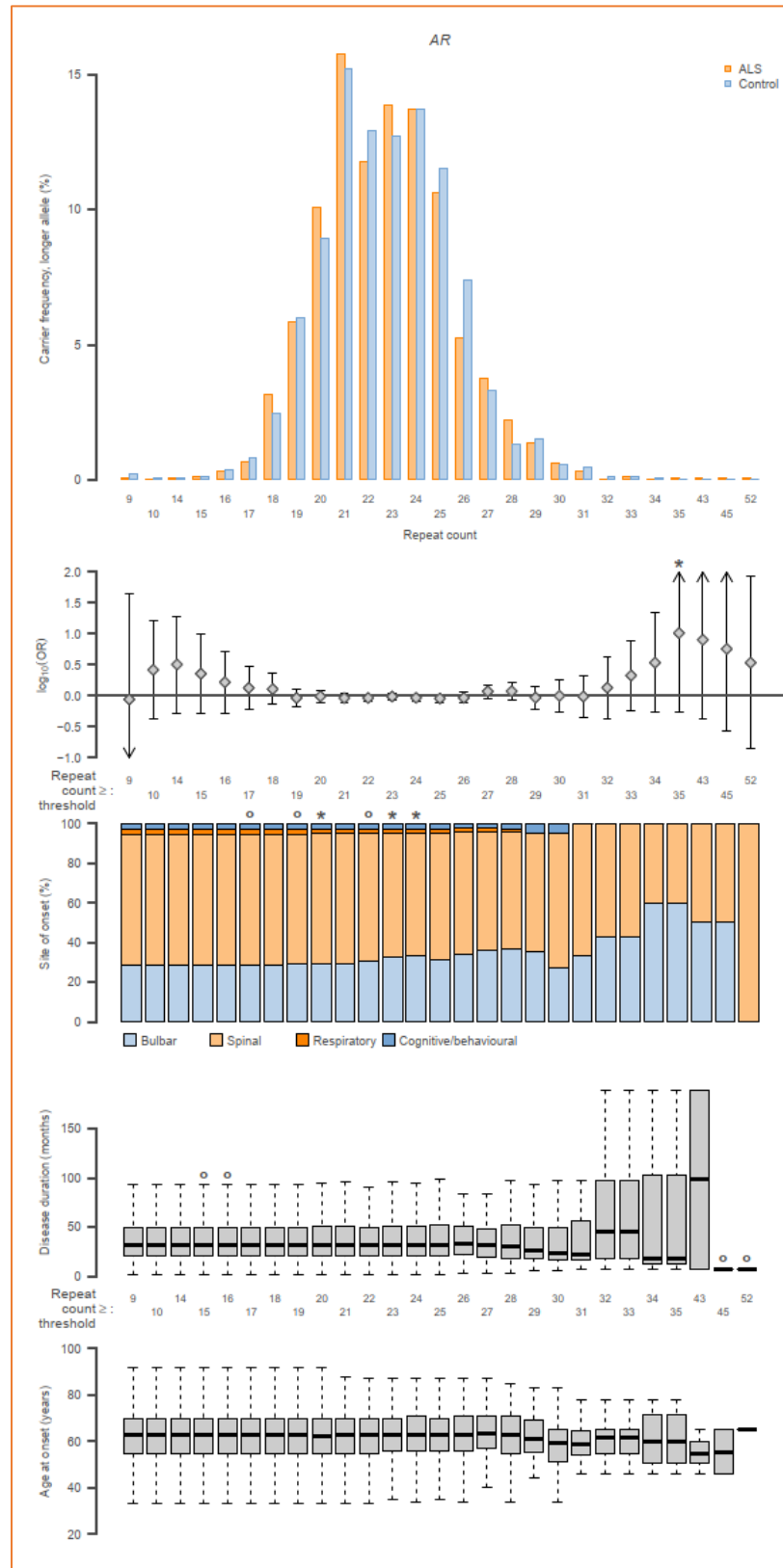


Figure 3-10 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *AR* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-10 presents the results of genotyping 1487 ALS and 1701 control individuals for the *AR* CAG repeat expansion. The highest longer allele frequency for both populations was 21 repeats (15.74% ALS, and 15.23% Controls), and carriers with 23 and 24 repeats had a higher allele frequency in ALS cases than controls (13.85% vs. 12.70%, 13.72% vs. 13.70%, respectively). This study amplified much longer repeats than the normal range (≥ 38 repeats). Fisher exact test revealed that for the PolyQ carriers with ≥ 35 repeats, both populations are significantly different (p -value: 0.047; OR=10.32, 95% CI 0.56-191.90). Interestingly, there is an increase of bulbar ALS with more than 22 repeats, which becomes statistically significant when considering more than 23 and 24 CAG repeats. Furthermore, disease duration decreases by 5.9 months, and age at onset decreases by 2 years when harbouring more than 29 repeats, but these results were not statistically significant. Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

ATXN3

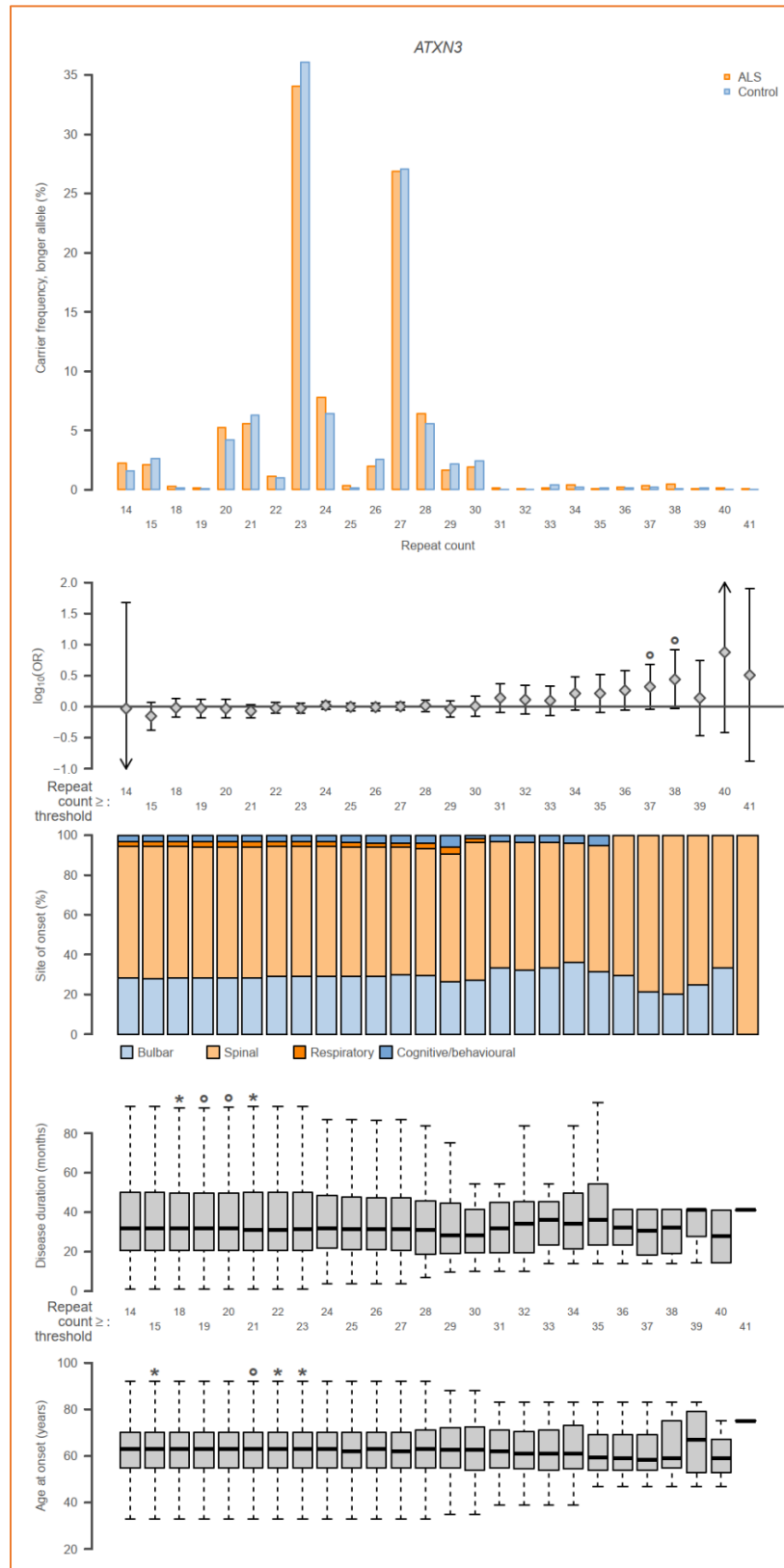


Figure 3-11 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *ATXN3* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-11 presents the findings from genotyping 1459 ALS and 1563 control individuals for the *ATXN3* CAG repeat expansion. The study revealed that the highest longer allele frequency for both populations were 23 repeats (34.06% ALS, and 36.09% Controls), followed by 27 repeats (26.87% ALS, and 27.03% controls). Notably, the protocol did not amplify expanded repeats between 61-84 repeats. Fisher's exact test and odds ratios did not indicate any significant association between ALS risk and expanded *ATXN3* repeat counts, but a trend with borderline statistical support is evident in the range of ≥ 34 repeats. An increase in cognitive/behavioural and respiratory ALS was observed with more than 29 repeats, although not statistically significant. Disease duration remained constant across all repeats without significant variation (approximately 31.9 months). Notably, when harbouring more than 29 repeats, disease duration decreased by 3.65 months, but the result was not statistically significant. Finally, the age at onset analysis revealed no significant differences across all genotyped repeats, with the age of onset medians approximately 63 years. Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4

ATXN8

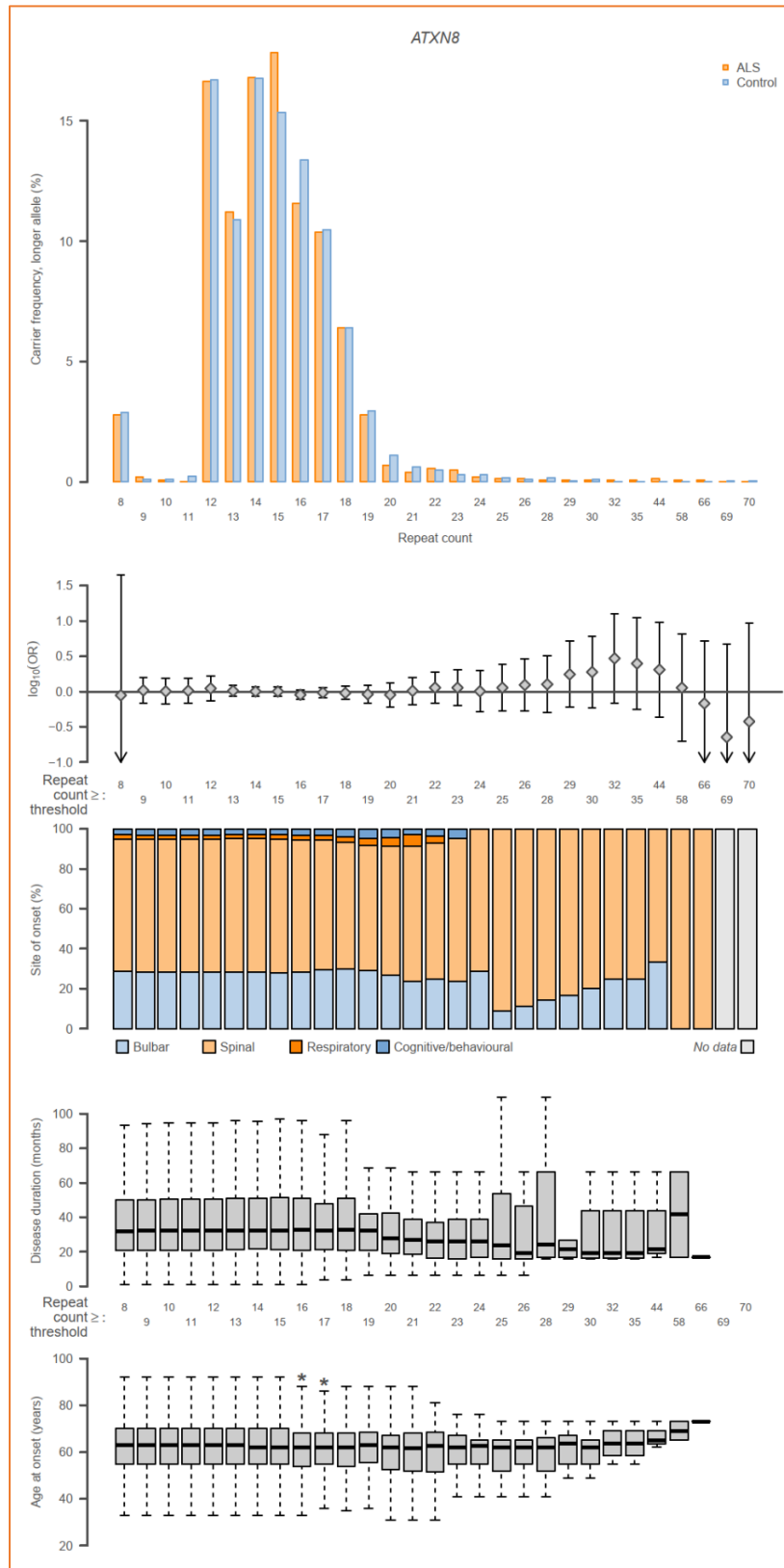


Figure 3-12 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *ATXN8* gene.

See section 3.2 for a full description of the information contained within the plot.

Genotyping of 1435 ALS and 1622 control individuals for the *ATXN8* CAG repeat expansion are presented in Figure 3-12. Our results showed that the highest longer allele frequency for both populations were 15 (17.84% ALS vs. 15.35% controls) and 14 repeats (16.79% ALS vs. 16.77% Controls), with similar frequencies in ALS and control groups. We were able to detect longer repeats beyond the normal range (>50). ORs reveal an increase in ALS risk when harbouring intermediate CAG repeats between ≥ 29 to ≥ 44 . Notably, the OR for those with more than 32 repeats was almost 3 times higher than those with lower repeat counts (OR=2.95, 95% CI 0.68-12.70), although the results were not statistically significant. The site of onset analysis suggested a non-significant increase of spinal ALS when harbouring more than 25 CAG repeats. Interestingly, disease duration decreased when harbouring between ≥ 20 and ≥ 44 repeats, but again, the results were not statistically significant. Finally, age at onset analysis revealed no significant differences across all genotyped repeats, with an average age of onset of approximately 63 years old. Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4.

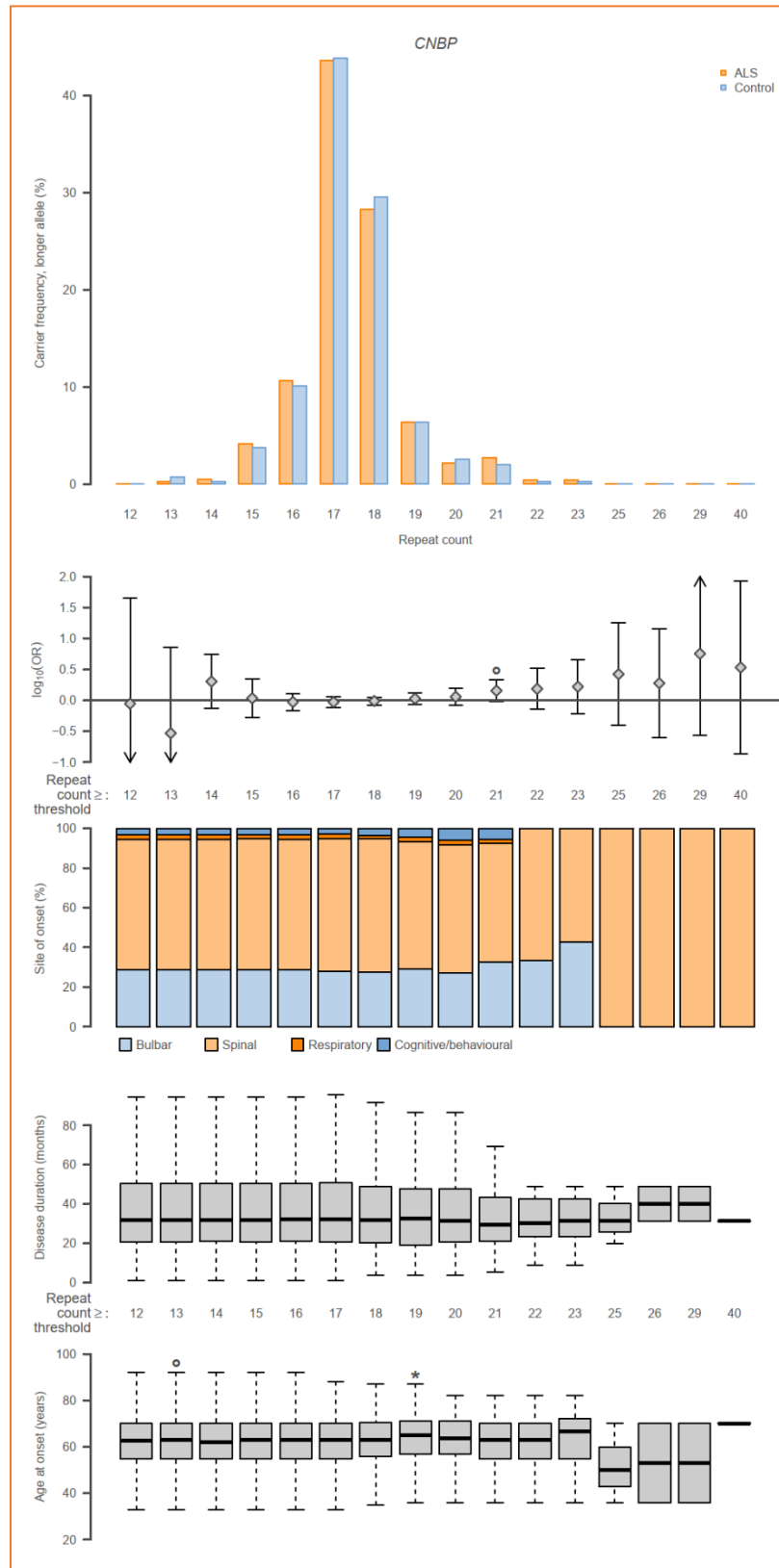


Figure 3-13 The composition of the population of ALS and Control samples for the CCTG repeat expansion in the *CNBP* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-13 presents genotyping results for the *CNBP* tetranucleotide repeat expansion in 1461 ALS cases and 1655 control individuals. The analysis revealed that the highest longer allele frequency for both populations was 17 repeats (43.67% ALS and 43.87% controls), followed by 18 repeats (28.27% ALS and 29.55% controls). Interestingly, ALS patients had a higher frequency of carriers than controls at 16 (10.68% vs. 10.15%) and 21 (2.74% vs. 1.99%) repeats. Notably, the protocol did not amplify expanded pathogenic repeats. The results showed a marginal increase in ALS risk starting at 21 repeats ($p = 0.083$; OR=1.43. 95% CI 0.96-2.13). There was no significant variation in site of onset or disease duration across all repeats (approximately 31.9 months). However, the age at onset increased by 0.6 years when harbouring more than 19 repeats, which was statistically significant (p -value 0.031). Overall, the results suggest a possible increase in ALS risk when harbouring intermediate CCTG repeats in the *CNBP* gene, but the sample size was too low to determine statistically significant association. Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4

FMRI

After revising all the *FMRI* FLA electropherograms, it was noticed that at least around 50% of the *FMRI* gene traces had several peaks. This could be background noise or a mosaicism of multiple numbers of the repeat for this gene, which could happen with gDNA obtained from peripheral blood. Whatever the reason, results were not fully reliable and should be verified by rpPCR. For completeness, the obtained results are presented in Figure 3-14 and discussed below; however, these should not be considered as fully reliable.

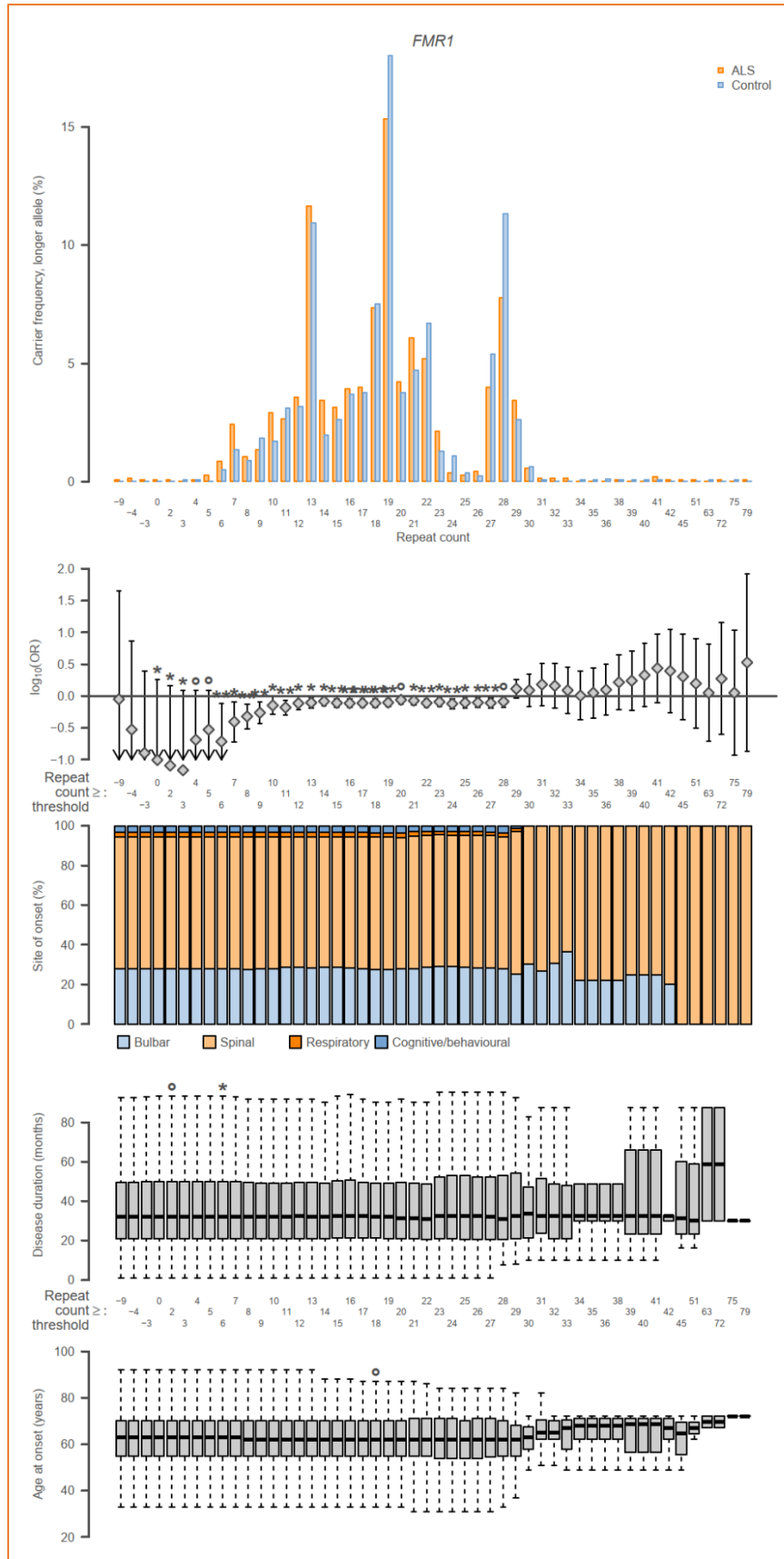


Figure 3-14 The composition of the population of ALS and Control samples for the CGG repeat expansion in the *FMR1* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-14 presents the findings from genotyping 1401 ALS and 1572 control individuals for the *FMR1* CGG repeat expansion. The analysis revealed that the highest longer allele frequency for both populations was 19 repeats (15.35% ALS, and 18.00% Controls), followed by 13 repeats (11.65% ALS, and 10.94% Controls). Notably, the protocol used in this study amplified much longer repeats lengths than the normal range (> 53). Fisher's exact test suggested an apparent protective effect for lower-length alleles, but the reliability of the data must be considered when interpreting these results. When considering repeats between ≥ 29 to ≥ 79 , the overall ORs suggest an increase in ALS risk, with the highest risk of 3 times with 79 or more repeats, although the results are not statistically significant (OR=3.37, 95% CI 0.14-82.76). Site of onset and disease duration stay the same across all repeats without significant variation, but duration increases by 1.5 months when harbouring more than 30 repeats, albeit not statistically significant. Age at onset increases by 3 years when harbouring more than 31 repeats, but this was not statistically significant. Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4

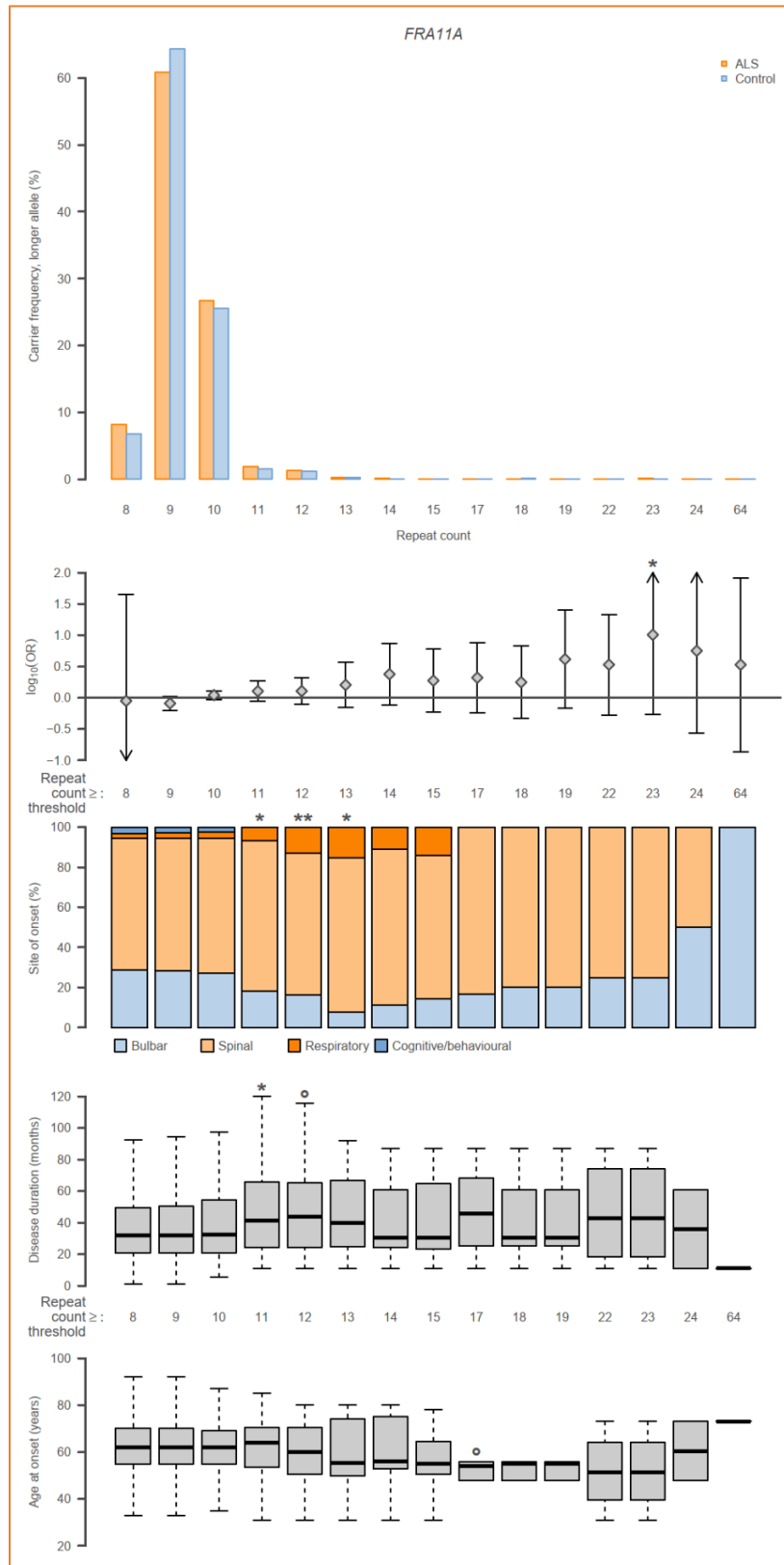


Figure 3-15 The composition of the population of ALS and Control samples for the CGG repeat expansion in the *FRA11A* (*C11orf80*) gene.

See section 3.2 for a full description of the information contained within the plot.

Our analysis of the *FRA11A* CGG repeat expansion included 1526 ALS cases and 1716 control individuals. As shown in Figure 3-15, the highest longer allele frequency for both populations were 9 repeats (60.88% ALS and 64.34% controls), followed by 10 repeats (26.74% ALS, and 25.58% controls). Our protocol did not amplify expanded repeats beyond 500 repeats. Although we found that ALS and controls were not significantly different for most *FRA11A* CGG repeat counts (p -value >0.05), we observed a trend towards increased risk conferred by expanded alleles, with ≥ 23 repeats significantly associated with ALS ($p = 0.0489$; OR=10.15, 95% CI 0.55-188.63). We found that respiratory onset ALS significantly increased when considering more than 11, 12 and 13 CGG repeats (p -values: 0.022, 0.009, and 0.036, respectively). We also observed an increase in disease duration when harbouring ≥ 11 and ≥ 12 repeats (p -values: 0.045 and 0.069, respectively). Interestingly, the median age of onset decreased by 2 years when harbouring more than 12 repeats, but the results were not statistically significant. Results of all the statistical analysis are in Supplementary Table S 3-3 and Supplementary Table S 3-4.

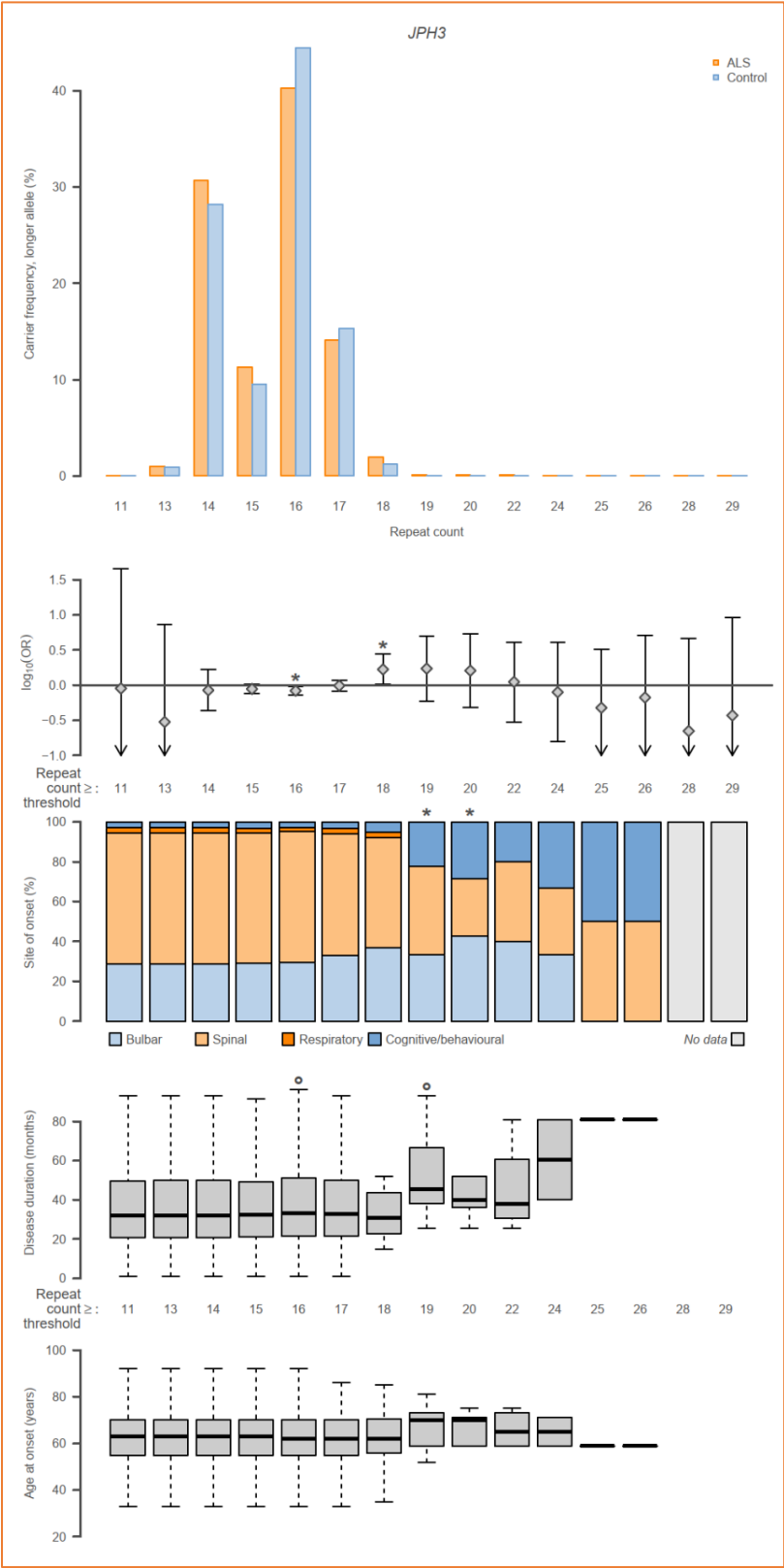


Figure 3-16 The composition of the population of ALS and Control samples for the CTG repeat expansion in the *JPH3* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-16 shows genotyping results for the *JPH3* CTG repeat expansion in a population of 1545 ALS cases and 1719 controls. The highest frequency of number of longer alleles for both populations was found to be 16 repeats (40.26% ALS and 44.44% controls), followed by 14 repeats (30.68% ALS and 28.21% controls). Amplification of greatly expanded repeats was not possible using this protocol, which excluded the amplification of repeats between 66-78. Greater than 18 CTG repeats was significantly associated with ALS risk ($p = 0.044$; OR = 1.68; 95% CI 1.02-2.75). Interestingly, the data suggested a protective effect in ALS risk of 83% when harbouring ≥ 16 CTG repeats in the *JPH3* gene, and this result was statistically significant (OR=0.83, 95% CI 0.72-0.96). When considering more than 19 and 20 CTG repeats, the proportion of cognitive/behavioural ALS increases significantly (p -values: 0.049 and 0.015, respectively), and respiratory ALS disappears. Disease duration increases by 13.76 months on average when harbouring more than 19 repeats, although this result is not statistically significant ($p = 0.067$). Similarly, age at onset increases by 7.5 years when harbouring more than 19 repeats, but this result is not statistically significant. Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4.

TK2-BEANI

During the sequencing analysis for the *TK2-BEANI* gene, the repeat motif of interest TGGAA (previously associated with SCA31) was not found in any of the sequenced individuals (Chapter 2, 0 *TK2-BEANI*). The TGGAA repeat is a motif that is located in the start of another known benign repeat expansion, TAAAA, which is the one amplified using the NDREs PCR-FLA protocol. Therefore, rpPCR and sequencing is the best methodology to amplify this new repeat expansion securely. The obtained results from our study of the TAAAA repeat expansion are represented in the Figure 3-17.

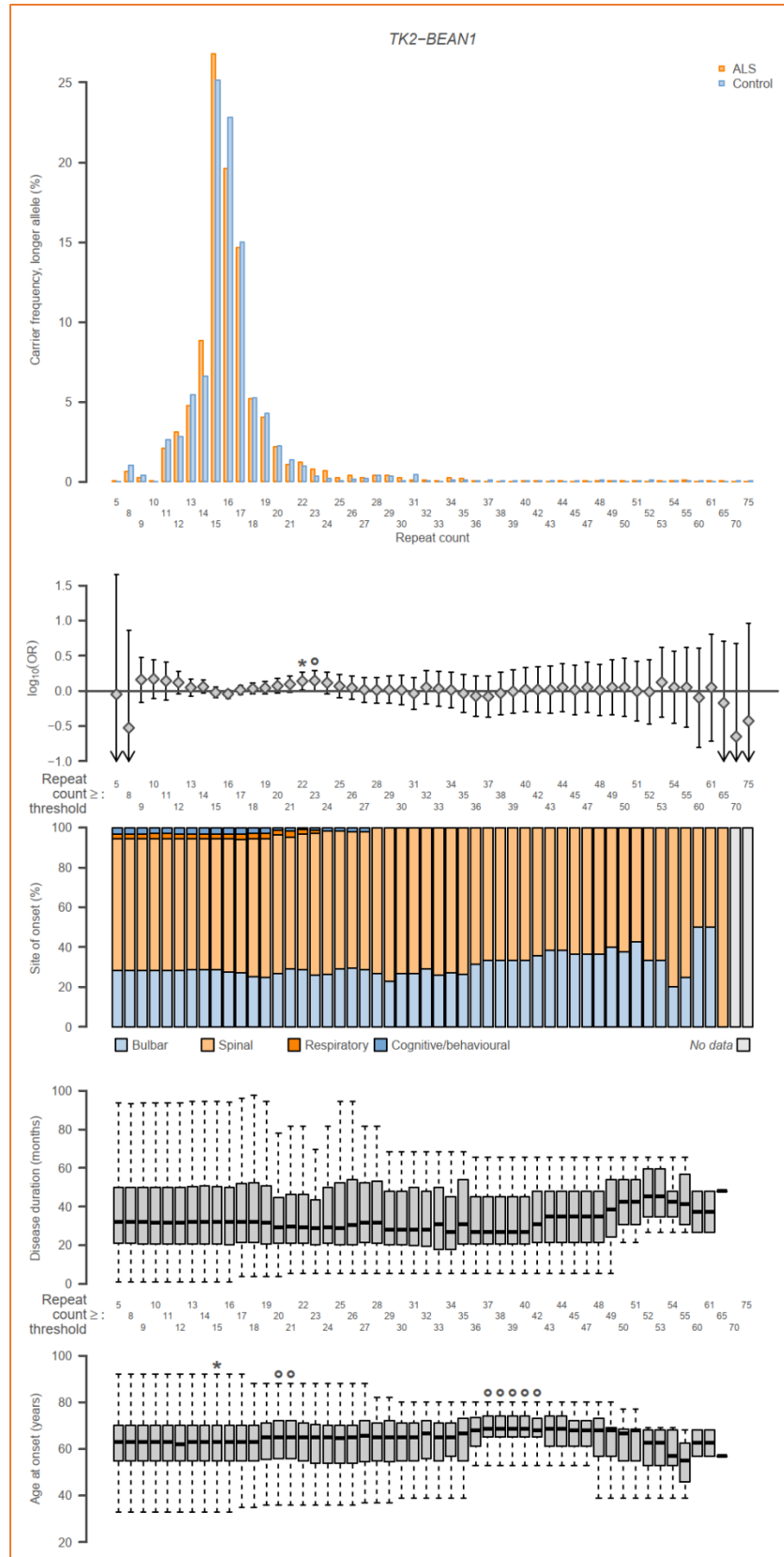


Figure 3-17 The composition of the population of ALS and Control samples for the TAAAA repeat expansion in the *TK2-BEANI* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-17 presents the results of our study on the *TK2-BEANI* TAAAA repeat expansion, which involved genotyping 1534 ALS and 1718 control individuals. The largest allele size observed in both populations was 15 repeats, with 26.79% of ALS cases and 25.15% of controls carrying this size. The second most common allele was 16 repeats, with 19.62% of ALS cases and 22.82% of controls carrying this size. Pathogenic expanded repeats (>2.5kb) were not detected by our protocol. 22 repeats or greater was significantly associated with ALS risk ($p = 0.039$; OR= 1.38, 95% CI 1.02-1.86). Site of onset and disease duration did not show significant variation across repeat sizes, with a mean of duration of approximately 31.9 months. However, when considering more than 44 repeats, there was a non-statistically significant increase in disease duration. Age at onset increased by 2 years when carrying more than 15 TAAAA repeats in the *TK2-BEANI* locus, which was statistically significant (p -value=0.0498). Results from all statistical analysis can be found in Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4.

Possible ALS risk genes

Here we define possible ALS risk genes as those in which available evidence shares features with established or novel ALS risk genes, suggesting that intermediate repeat lengths at these loci may increase ALS risk or confer a specific ALS sub-phenotype, but the overall signal is too low to be statistically significant with current sample sizes. These genes are of particular interest to the thesis, as they provide further evidence for the existence of pleiotropic effects between intermediate repeat expansions and ALS. While the results are not statistically significant, the possibility of increased risk warrants further investigation, particularly with a larger sample size and replication in other populations. Figures 3.19 to Figure 3.27 present the results of the longer repeat allele of nine genes - *ATN1*, *ATXN7*, *DMPK*, *GICP1*, *NOP56*, *NOTCH2NLC*, *NUTM2B-AS1*, *PPP2R2B* and *TBP* - all of which fall within this category.

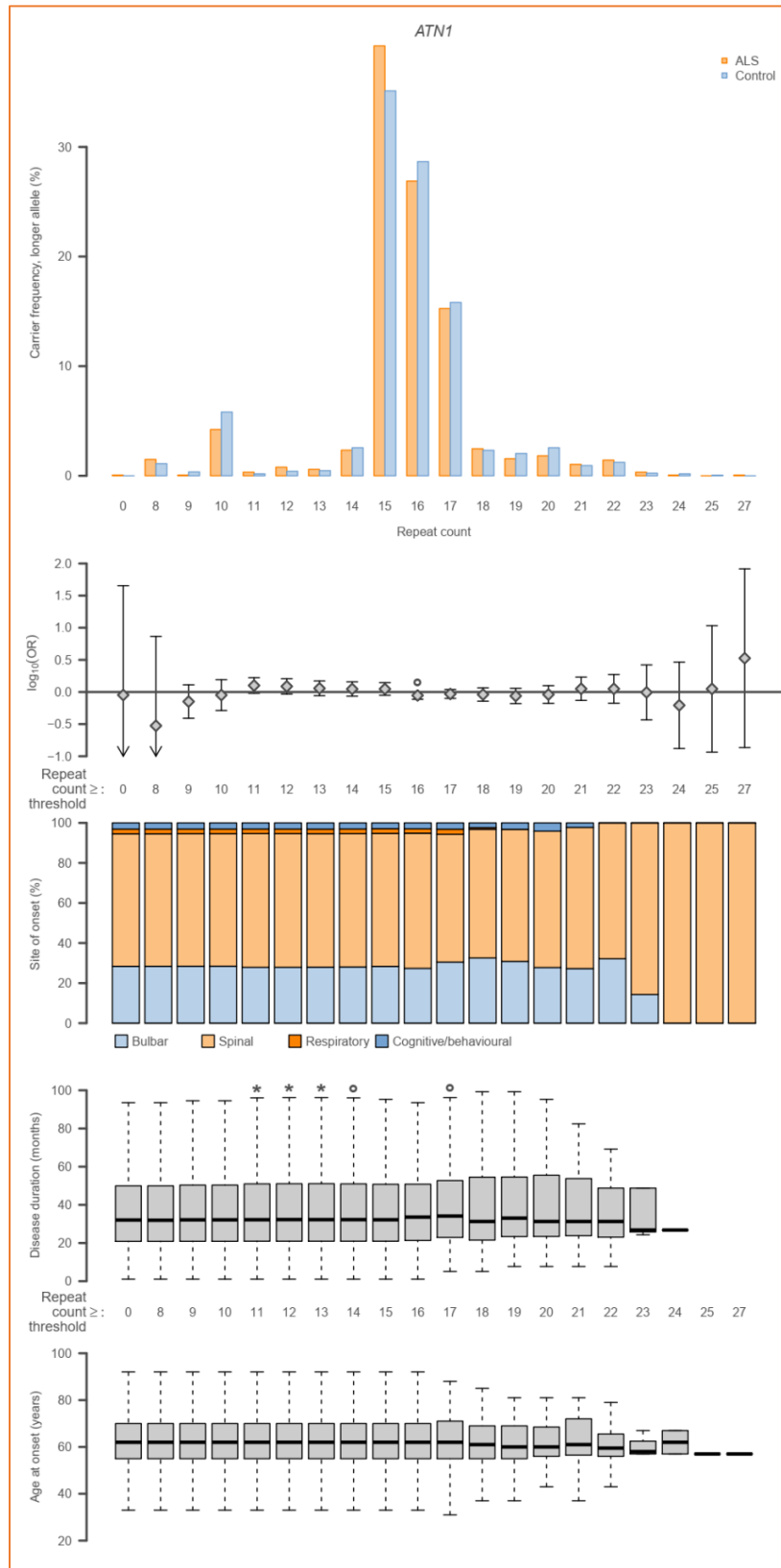


Figure 3-18 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *ATNI* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-18 presents the results of genotyping 1540 ALS patients and 1720 controls for the *ATNI* repeat expansion. The highest longer allele frequency of the CAG repeat for both populations was 15 15 (39.22% ALS and 35.12% controls), followed by 16 repeats (26.88% ALS and 28.66% controls). ALS patients had a higher percentage of carriers at 15, 18, and 22 repeats compared to control cases, although the differences were not statistically significant (p -values= 0.3583 0.4661, and 0.6923, respectively). ORs obtained from our study were not statistically significant, indicating no increased risk associated with ALS when harbouring any of these CAG intermediate repeat lengths in *ATNI*. Site of onset analysis indicated that ALS cases with CAG repeats had a higher percentage of spinal onset across all repeats followed by bulbar onset. However, when considering more than 17 and 18 CAG repeats, there was an increase of cases with bulbar onset, although not statistically significant. Disease duration was approximately 31.9 months across all repeats, with significant variations observed when harbouring ≥ 11 , ≥ 12 , and ≥ 13 CAG repeats (p -values < 0.05). Age at onset analysis revealed no significant differences across all genotyped repeats, with a median age of onset at approximately 63 years. Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

ATXN7

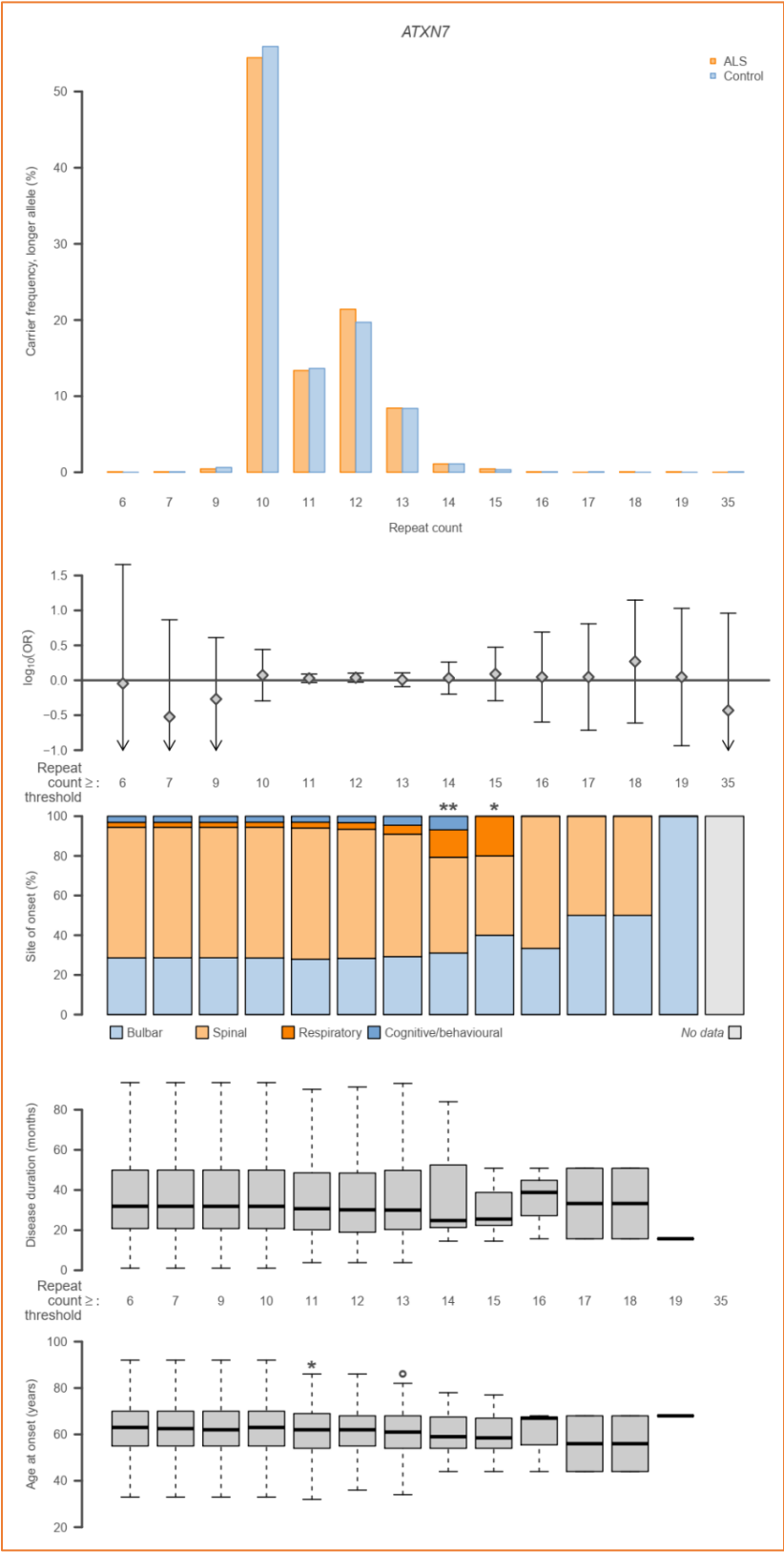


Figure 3-19 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *ATXN7* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-19, we present the results of genotyping 1541 ALS and 1715 control individuals for the *ATXN7* CAG repeat expansion. The highest longer allele frequency for both populations was found to be 10 repeats (54.45% ALS and 55.92% controls), followed by 12 repeats (21.42% ALS and 19.71% controls). Our protocol did not amplify expanded pathogenic repeats (≥ 37 repeats). ALS patients showed a higher percentage of carriers at 12 and 13 repeats compared to control cases (21.41% vs. 19.71%, 8.44% vs. 8.40%, respectively), although this was not statistically significant ($p > 0.05$). The obtained OR across the repeat counts found in our study for the *ATXN7* gene showed a slight increase in ALS risk when harbouring 18 or more CAG repeats (OR=1.86, 95% CI 0.12-10.71), which was not statistically significant (Supplementary Table S 3-2). Our site of onset bar plot indicated that ALS cases with CAG repeats had a higher percentage of spinal onset across all repeats followed by bulbar. We observed a statistically significant tendency towards respiratory ALS when considering more than 14 and 15 CAG repeats (p -values: 0.004 and 0.028, respectively). We found that disease duration was approximately 31.9 months across all repeats, but there was a slight decrease of 6.37 months in patients with more than 14 repeats, but this was not statistically significant. Age at onset decreased by 4 and 4.5 years when harbouring more than 14 and 15 repeats, respectively, but again, this was not statistically significant. Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

DMPK

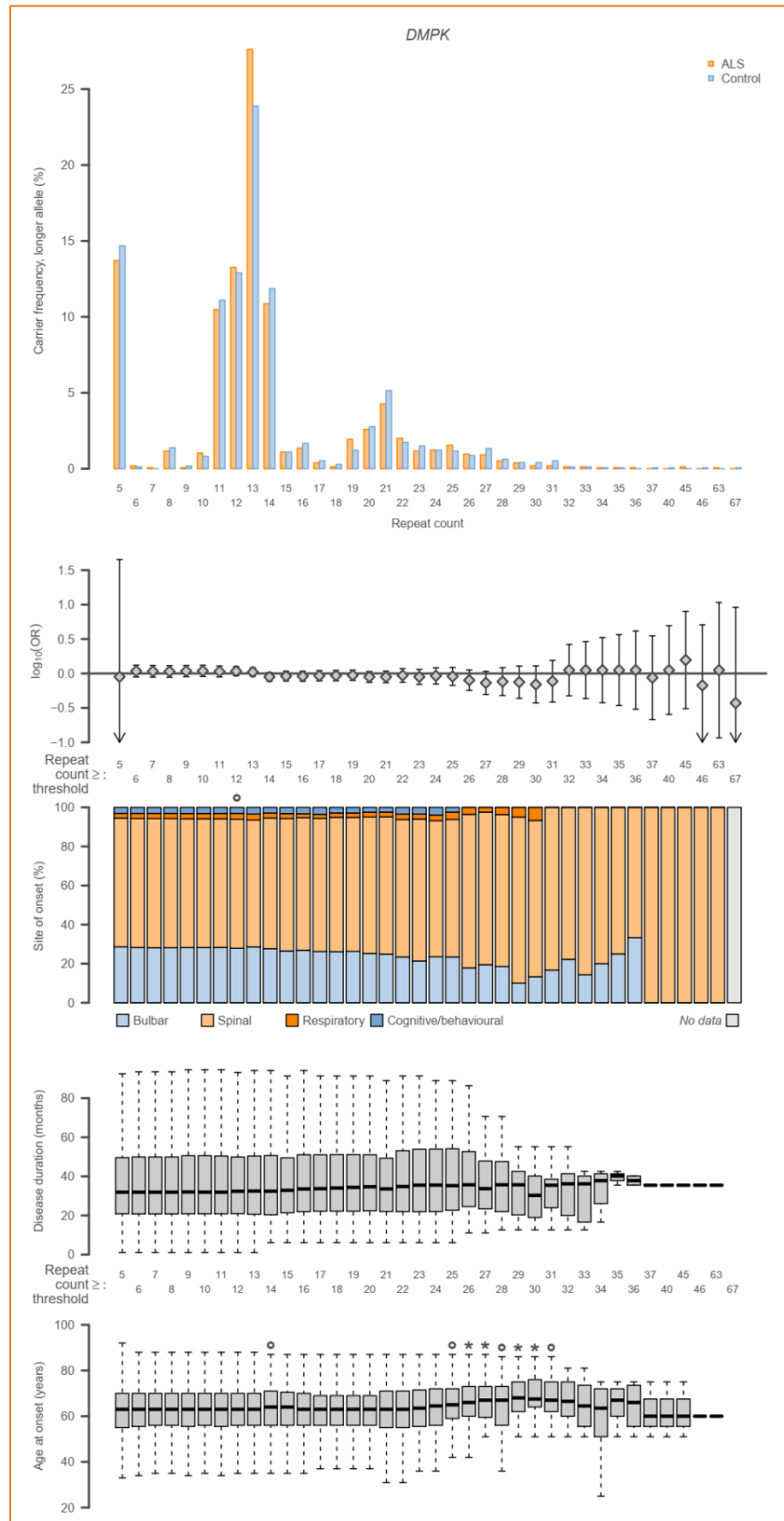


Figure 3-20 The composition of the population of ALS and Control samples for the CTG repeat expansion in the *DMPK* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-20 represents the results of 1546 ALS and 1729 control individuals genotyped for the *DMPK* CTG repeat expansion. The highest longer allele frequency of the CTG repeat was 13 in both populations, with 27.62% ALS and 23.89% controls. The obtained ORs are not statistically significant, but we observed an increased risk associated with ALS when harbouring intermediate repeat lengths in *DMPK* between 32-36 and 40-45 repeats. We found the highest increase of ALS risk when harbouring 45 or more CTG repeats (OR= 1.57, 95% CI 0.31-7.95). The site of onset bar plot suggested that ALS cases with CTG repeats had a higher percentage of spinal onset across all repeats followed by bulbar. However, when considering more than 36 CTG repeats, the cognitive/behavioural aspect in ALS patients disappeared, although it was not statistically significant. The disease duration was consistent across all repeats without significant variation (approximately 31.9 months). Age of onset increased significantly when harbouring more than 26, 27, 29, and 30 repeats by 4, 5, 6 and 5.5 years, respectively (p -values < 0.05). Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

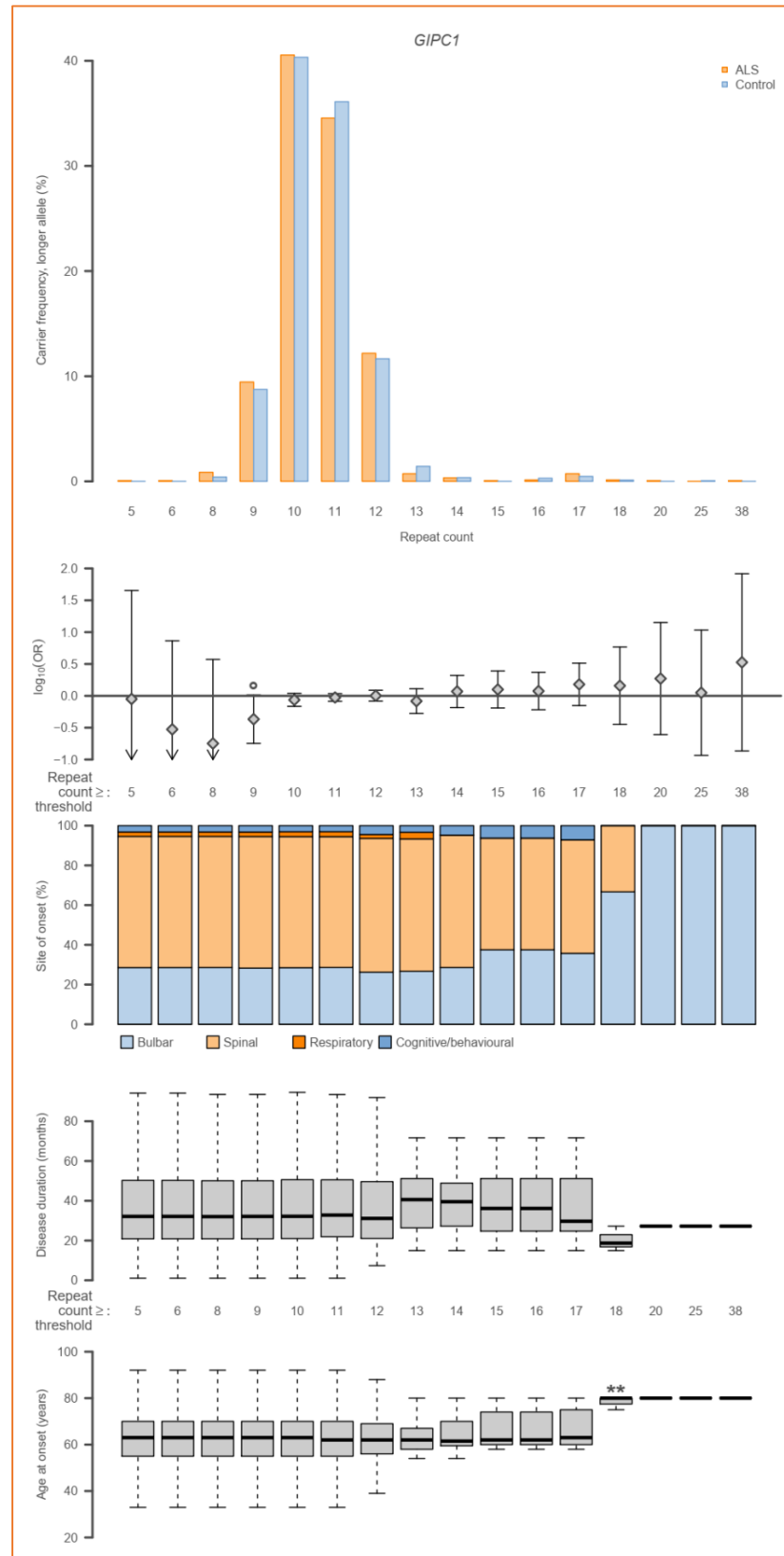


Figure 3-21 The composition of the population of ALS and Control samples for the GGC repeat expansion in the *GIPCI* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-21 provides a comprehensive analysis of the association between *GIPCI* GGC repeat expansion and ALS risk. The study included 1502 ALS patients and 1681 control individuals who were genotyped for the *GIPCI* GGC repeat expansion. The analysis revealed that both populations had the highest longer allele frequency of the trinucleotide repeat at 10 repeats, with ALS patients and control cases having similar percentages (40.55% ALS and 40.33% controls). However, ALS patients showed a higher percentage of carriers at 12 (12.18% vs. 11.66%), and 17 (0.73% vs 0.48%) than control cases, which may suggest a possible increase in ALS risk. Fisher's exact test *p*-values (all > 0.05) and ORs suggest that there is no increase in ALS risk when harbouring any of the obtained GGC repeat lengths in the *GIPCI* gene, although there is a possible increase in risk for more than 17 and 38 repeats ($OR_{\geq 17}=1.51$, 95% CI 0.70-3.26; $OR_{\geq 18}=3.36$, 95% CI 0.14-82.54). However, these results were not statistically significant. The site of onset analysis revealed that ALS cases with GGC repeats have a higher percentage of spinal onset across all repeats, followed by bulbar onset. Furthermore, there was an apparent increase in bulbar onset for GGC repeats greater than 15, although this was not statistically significant. Disease duration did not vary significantly across all repeats, with an average of approximately 31.9 months. Additionally, age at onset analysis revealed no significant differences across all genotyped repeats, with a median age of onset at approximately 63 years. Overall, the results of this study suggest that there is no significant association between *GIPCI* GGC repeat expansion and ALS risk, although there may be a possible increase in risk for specific repeat lengths. Supplementary Table S 3-2, Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

NOP56

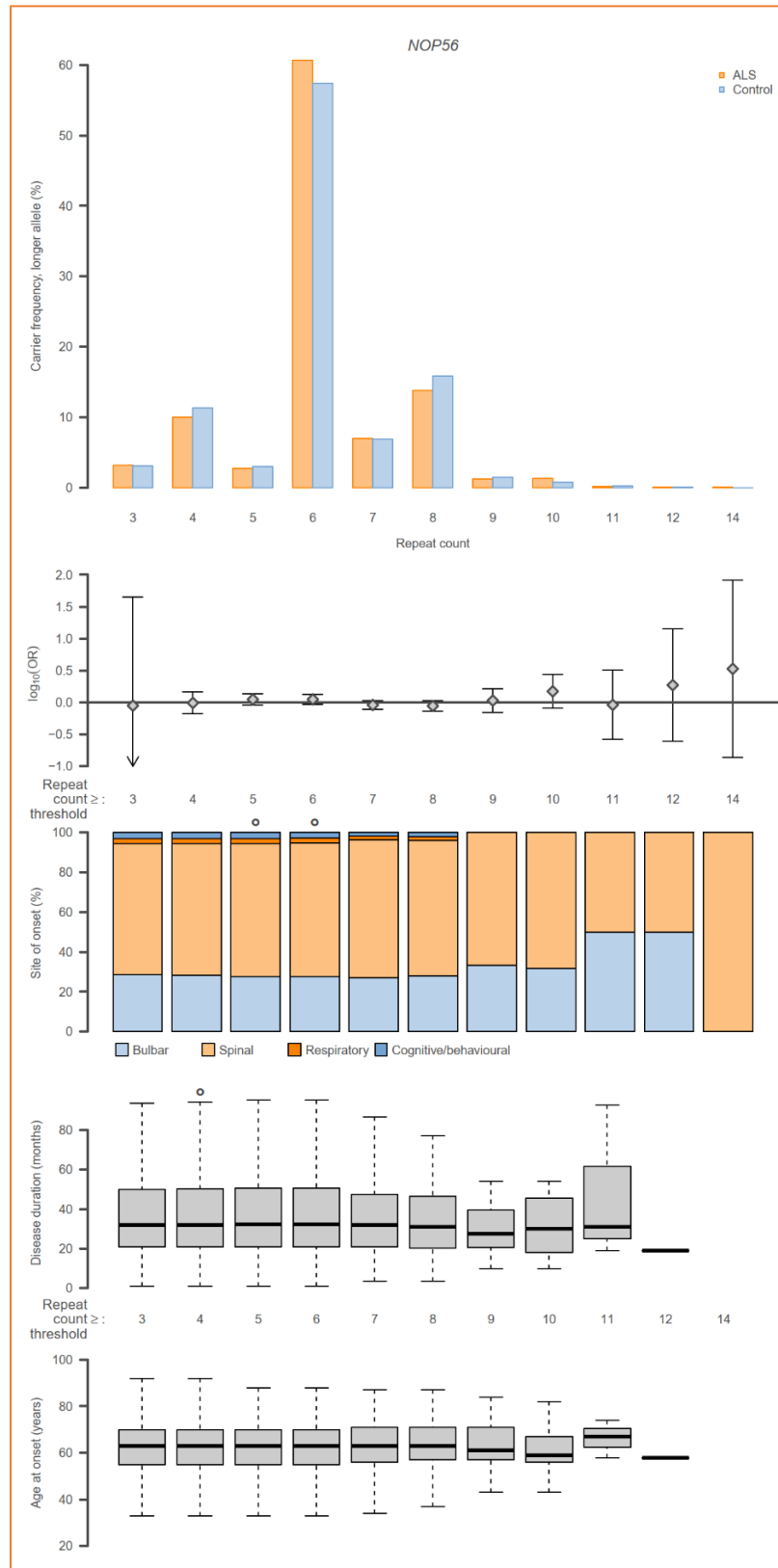


Figure 3-22 The composition of the population of ALS and Control samples for the GGCCTG repeat expansion in the *NOP56* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-22 shows results for 1539 ALS and 1725 control individuals genotyped for the *NOP56* hexanucleotide repeat expansion. For both populations, the highest longer allele frequency of the GGCCTG repeat was 6 (60.69% ALS and 57.50% controls), followed by 8 repeats (13.84% ALS and 15.83% controls). Meanwhile, ALS patients showed a higher percentage of carriers at 6, and 10 (1.30% vs 0.75%) repeats than control cases, although this was not statistically significant ($p = 0.2393$ and 0.2147 respectively). This protocol did not amplify expanded pathogenic repeats (≥ 50 repeats). The second panel shows the obtained ORs across the repeat counts found in our study for the *NOP56* gene. The overall ORs indicate no increase in ALS risk when harbouring an any of the obtained GGCCTG repeat length in the *NOP56* gene. However, when considering ≥ 10 (OR=1.49, 95% CI 0.81-2.74), ≥ 12 (OR=1.87, 95% CI 0.15-14.18) and ≥ 14 (OR=3.37, 95% CI 0.14-182.66) CCTG repeats, these results suggest a possible increase of risk in ALS, albeit not statistically significant with this sample size. Site of onset bar plot suggests that ALS cases with GGCCTG repeats have a relatively uniform distribution of site of onset across repeat count thresholds. Disease duration stays the same across all repeats without significant variation (approximately 31.9 months). Age at onset analysis revealed no significant differences across all genotyped repeats, with a median age of onset at approximately 63 years. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

NOTCH2NLC

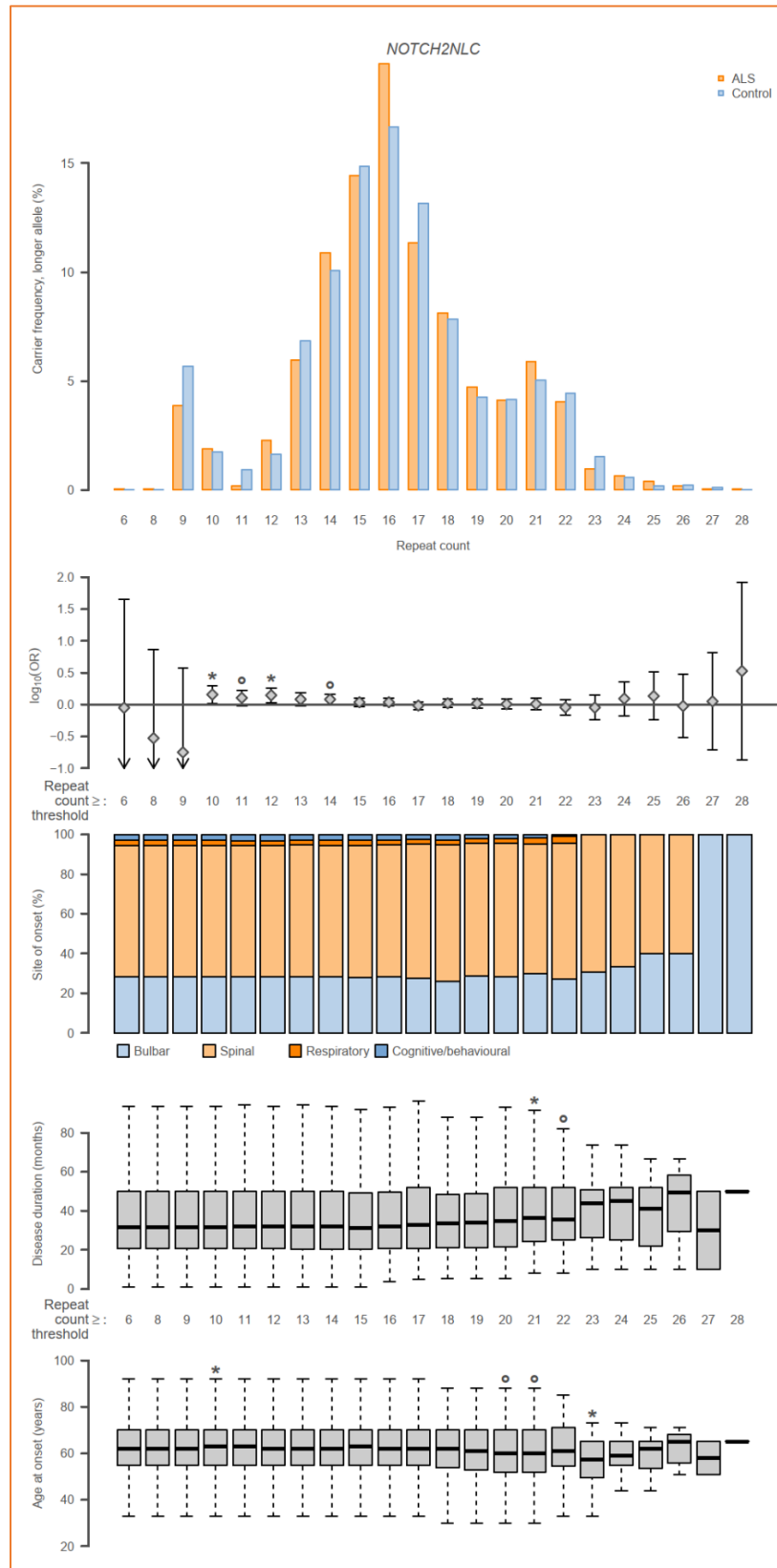


Figure 3-23 The composition of the population of ALS and Control samples for the GGC repeat expansion in the *NOTCH2NLC* gene.

See section 3.2 for a full description of the information contained within the plot.

We investigated the association between the *NOTCH2NLC* GGC repeat expansion and ALS by genotyping 1523 ALS patients and 1709 controls. The results, shown in Figure 3-23, reveal that the highest longer allele frequency of the trinucleotide repeat was 16, which was observed in 19.57% of ALS patients and 16.68% of controls, followed by 14 repeats, observed in 10.90% of ALS patients and 10.06% of controls. We observed a higher percentage of carriers at 16, 18, and 21 repeats in ALS patients compared to controls, with respective frequencies of 8.14% vs. 7.84%, and 5.91% vs. 5.03%. Fisher exact test was used to evaluate the association between ALS case/control status and the GGC repeat counts. Our analysis revealed that the GGC repeat counts were not significantly different between ALS patients and controls (p -values= 0.2370, 0.5597, and 0.8718, for carriers with 16, 18, and 21 repeats, respectively). The obtained odds ratios across the repeat counts did not indicate an increased risk of ALS. However, when considering more than 28 repeats, the results suggest a possible 3-fold increase in ALS risk, but the results were not statistically significant (OR=3.37, 95% CI 0.14-82.75). Our site of onset bar plot suggests that ALS cases with GGC repeats have a higher percentage of spinal onset, followed by bulbar onset. Disease duration was found to increase by 1.88 months when harbouring more than 21 repeats (p -value 0.031), and age at onset was found to decrease when harbouring more than 23 repeats by 5.5 years (p -value 0.028), with both results reaching statistical significance (p -value < 0.05). Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

NUTM2B-AS1

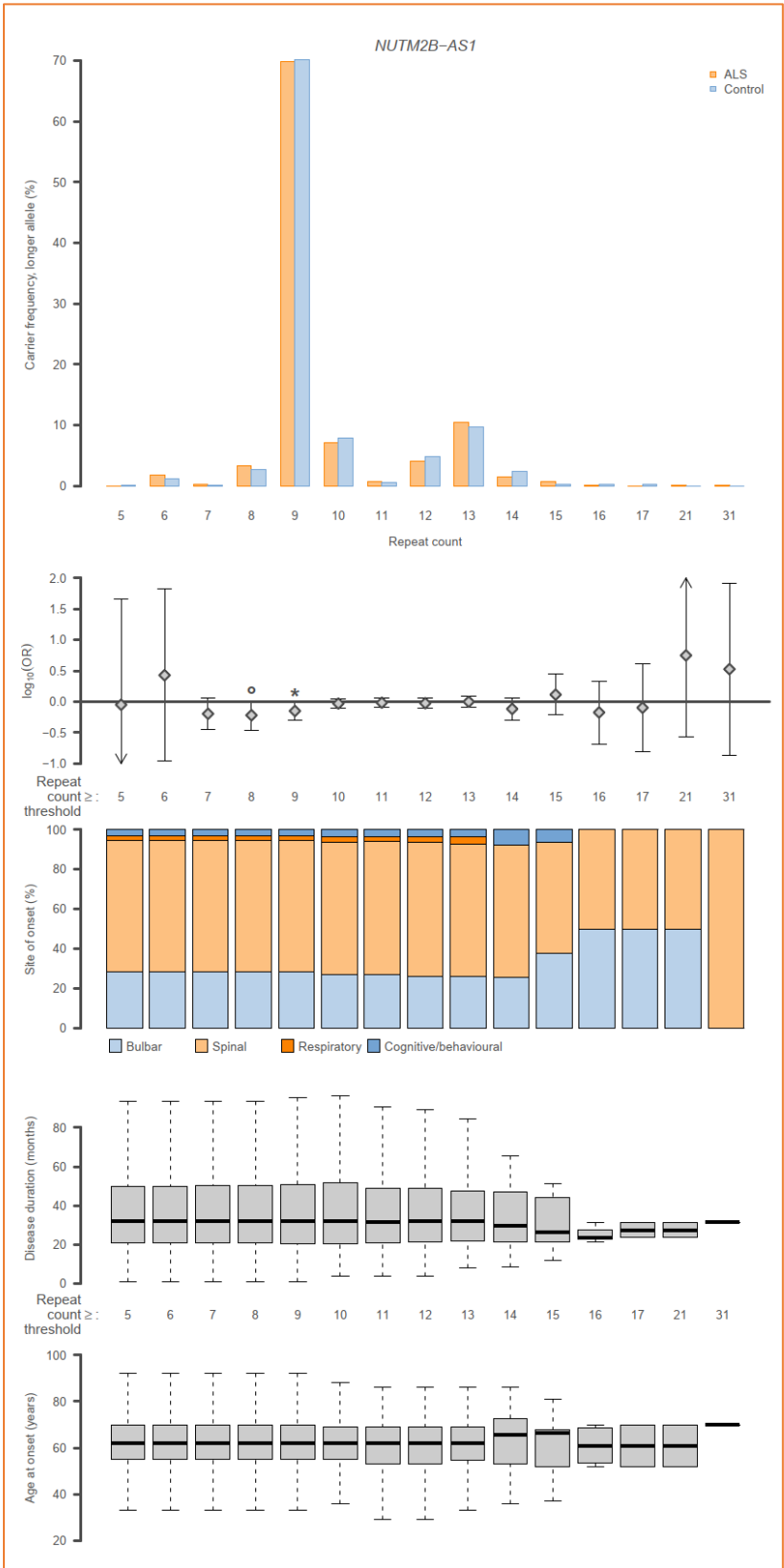


Figure 3-24 The composition of the population of ALS and Control samples for the CGG repeat expansion in the *NUTM2B-AS1* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-24 shows results for 1539 ALS patients and 1724 control individuals genotyped for the *NUTM2B-AS1* CGG repeat expansion. The highest allele frequency of the trinucleotide repeat was 9, observed in 69.92% of ALS patients and 70.13% of controls, followed by 13 repeats, which had the highest percentage of ALS patients compared to control cases. Our protocol successfully amplified much longer repeats than the normal range, with repeats greater than or equal to 16 included in our analysis. Using Fisher exact test, we estimated the association between ALS case/control status and found that for GGC carriers with 16, 17, and 21 repeats, the probability that ALS and controls were sampled from the same population was 0.5554, 1.0000, and 0.2224, respectively. We obtained odds ratios (ORs) across the repeat counts found in our study for the *NUTM2B-AS1* gene, which showed no increase in ALS risk when harbouring any of the obtained GGC repeat length in the gene. However, when considering more than 21 repeats and more than 31 repeats, results may suggest a possible increase of risk in ALS, but these results are not statistically significant ($OR_{\geq 21}=5.61$, 95% CI 0.27-116.91; $OR_{\geq 31}=3.36$, 95% CI 0.14-81.61). Site of onset, disease duration and age at onset showed no significant variation across repeat length thresholds. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

PPP2R2B

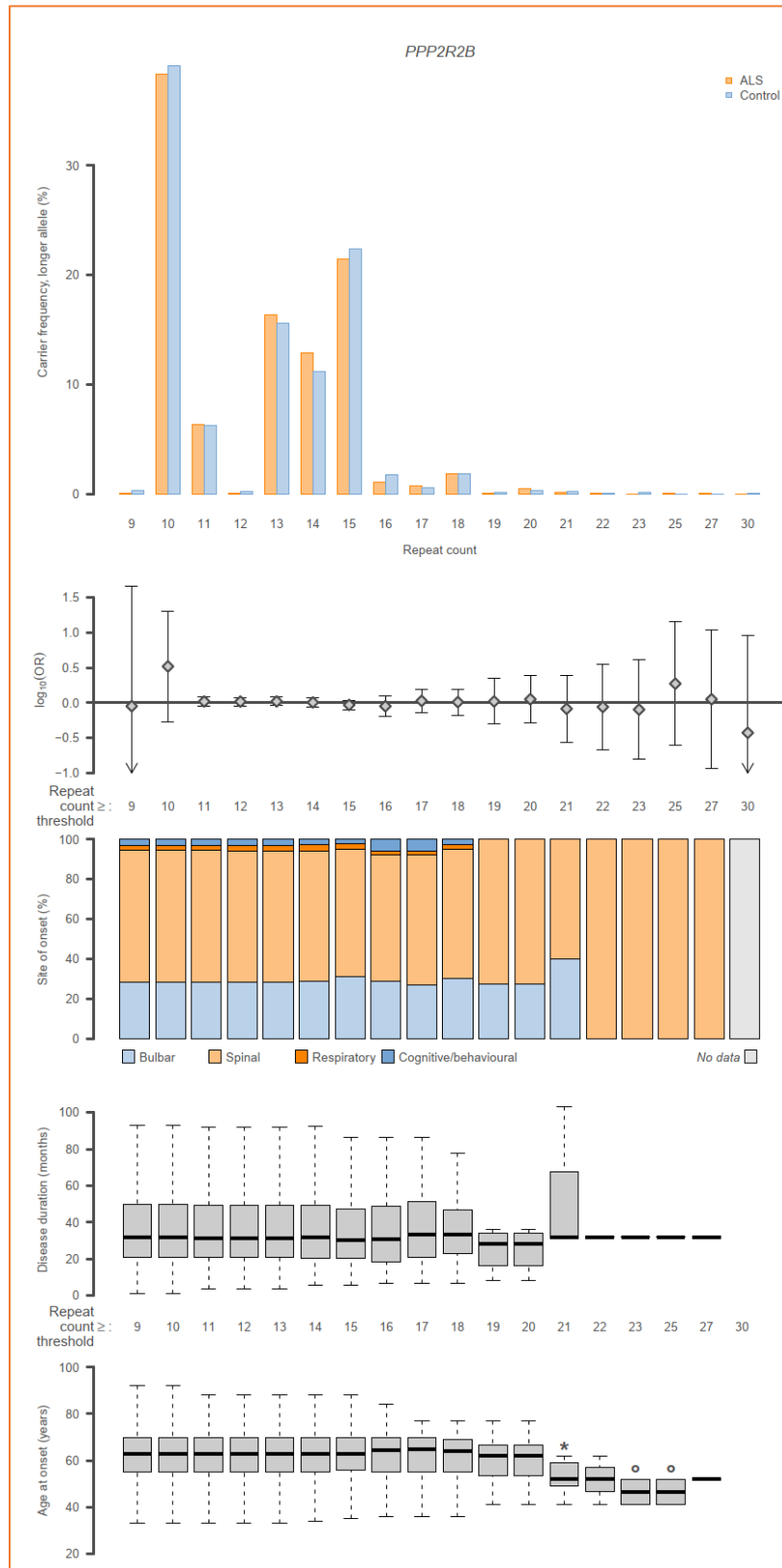


Figure 3-25 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *PPP2R2B* gene.

See section 3.2 for a full description of the information contained within the plot.

We analysed the genotypes of 1540 ALS patients and 1723 control individuals for the *PPP2R2B* CAG repeat expansion. Results in Figure 3-25 show that the highest longer allele frequency of the PolyQ repeat was 10 (38.38% ALS and 39.12% controls), followed by 15 repeats (21.49% ALS and 22.40% controls), and that ALS patients had a higher percentage of carriers at 13 and 14 repeats than control cases, although no repeat threshold was significantly different between cases and controls ($p > 0.05$). Our obtained ORs are not statistically significant, and there is no increased risk associated with ALS when harbouring any of these CAG intermediate repeat lengths in *PPP2R2B*. Site of onset was generally uniform; however, although not statistically significant, we observed an increase in ALS patients with cognitive/behavioural characteristics for 16 or more repeats. Disease duration stays the same across all repeats without significant variation (approximately 31.9 months). Age at onset decreases significantly when harbouring more than 21 repeats by 11 years (p -value 0.049). Overall, our results suggest that the *PPP2R2B* CAG repeat expansion is not a major risk factor for ALS but may affect age at onset. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

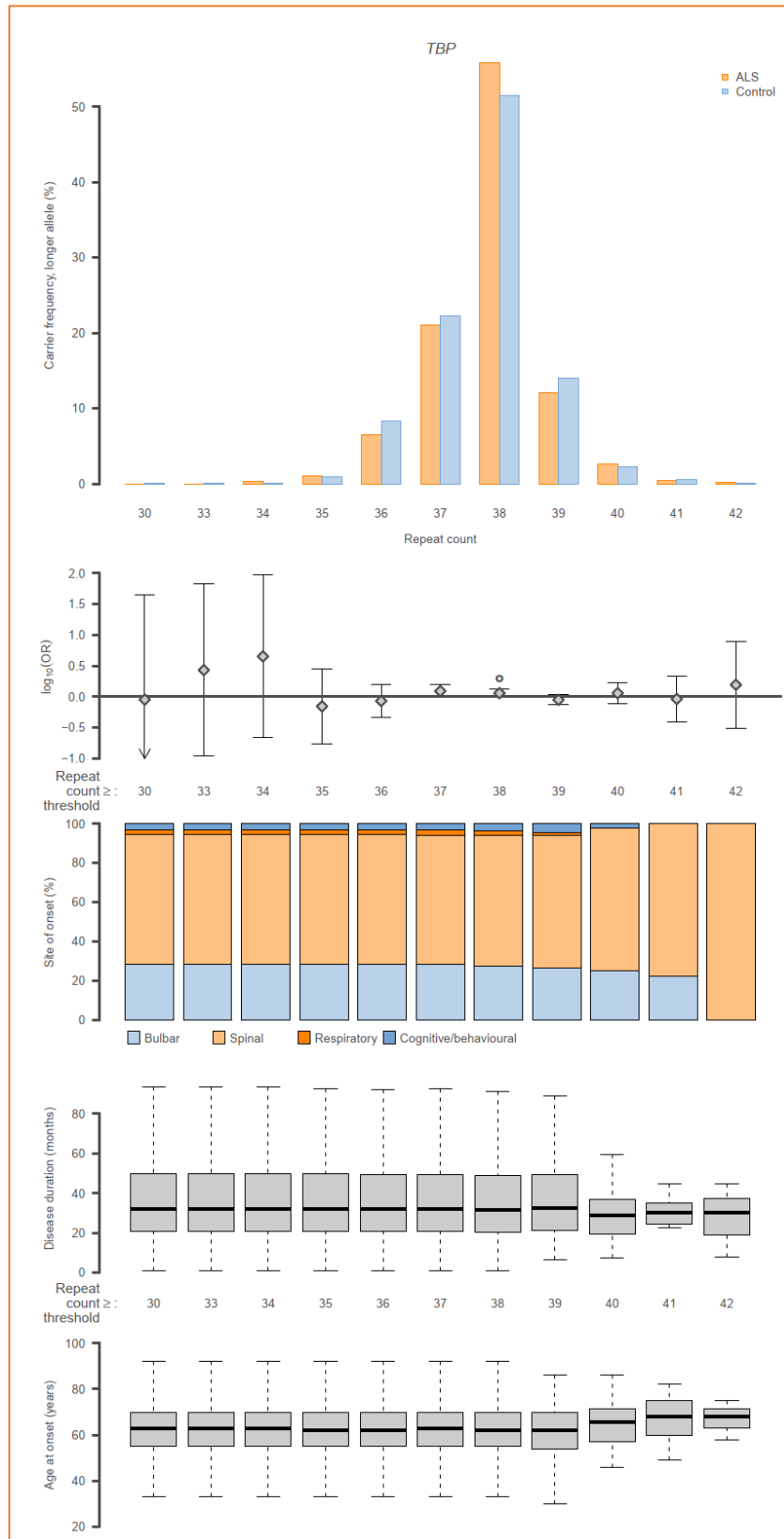


Figure 3-26 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the *TBP* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-26 represents the results of 1551 ALS and 1727 control individuals genotyped for the *TBP* CAG repeat expansion. For both populations, the highest longer allele frequency of the PolyQ repeat was 38 (55.83% ALS and 51.48% Controls), followed by 37 repeats (21.08% ALS and 22.24% Controls). This protocol did not amplify expanded pathogenic repeats (between 47-63 repeats). Case/control counts were generally not significantly different, with one borderline *p*-value (0.0867) for 38 or more repeats. The second panel shows the obtained ORs across the repeat counts in our *TBP* gene study. The obtained ORs are not statistically significant, and there is a small increase in ALS risk when harbouring ≥ 37 , ≥ 38 , ≥ 40 , and ≥ 42 repeats the latest with the highest increase of ALS risk of 56% (OR=1.56, 95% CI 0.31-7.92). Site of onset was uniform across repeat counts. When harbouring more than 40 repeats, disease duration decreases by 3.18 months, and age at onset increases by 3.5 years. Both results were not statistically significant. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

Genes not associated with ALS risk

Out of the 35 genes analysed, we did not observe a statistically significant association between unstable repeats and ALS risk in 19 of them. However, the complexity of neurodegenerative diseases suggests that further investigation of all unstable repeat regions is necessary, considering the possibility of pleiotropy and the involvement of multiple repeat expansion genes in the development of such diseases. The data presented in Figures 3.28 to 3.36 provide a comprehensive overview of the longer repeat allele distribution of the 9 genes in our sample population, accompanied by detailed explanations of the obtained results. Nonetheless, future research is required to replicate these findings in larger patient cohorts to enhance statistical power and confirm our findings. All statistical results can be found in Supplementary Table S 3-3 and Supplementary Table S 3-4.

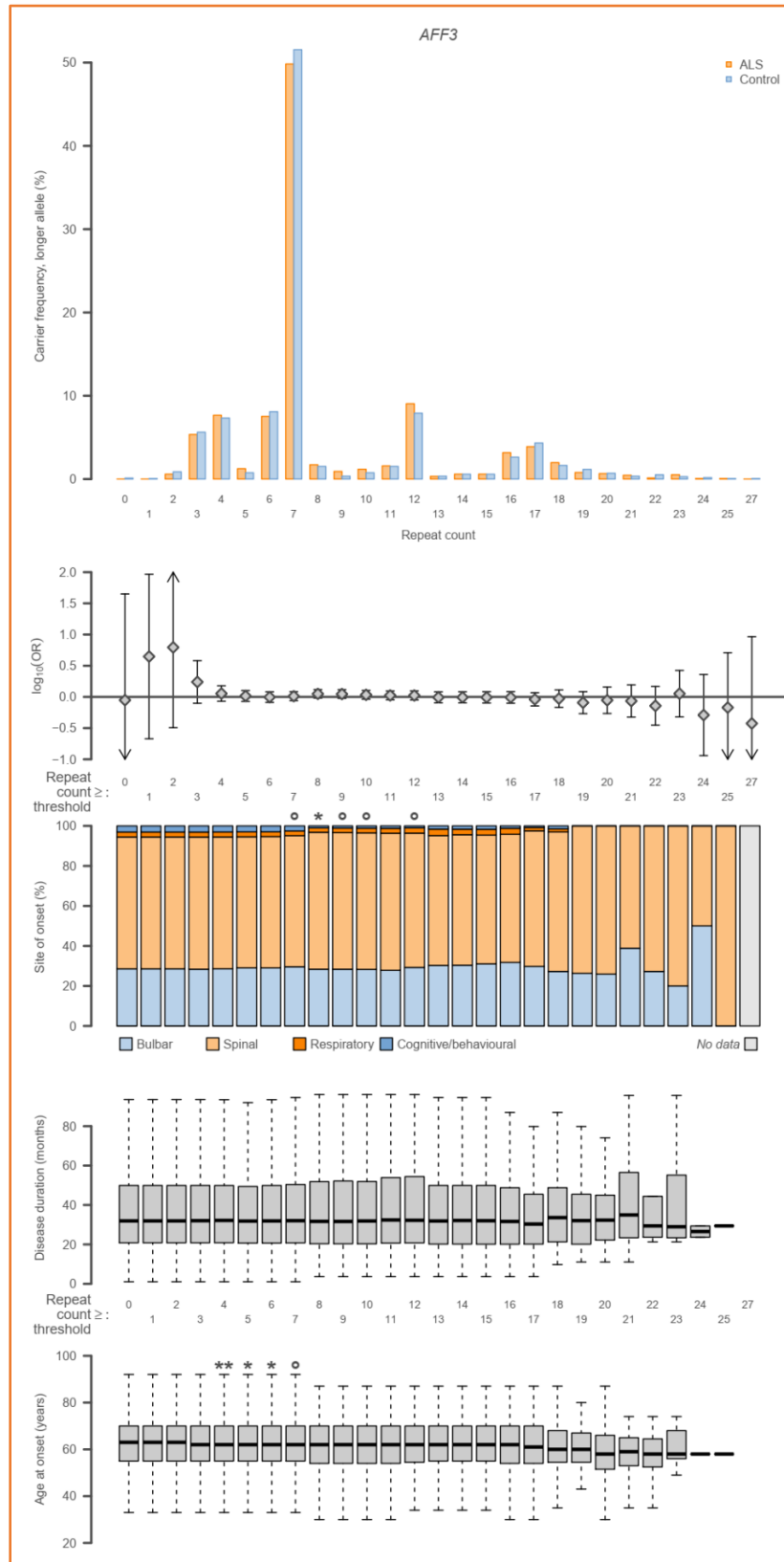


Figure 3-27 The composition of the population of ALS and Control samples for the CCG repeat expansion in the *AFF3* gene.

See section 3.2 for a full description of the information contained within the plot.

Our results, presented in Figure 3-27, show that the highest longer allele frequency of the CCG repeat count is 7 (49.83% ALS and 51.55% Controls), followed by 12 repeats (9.05% ALS and 7.92% Controls), in both the ALS and control populations. However, ALS patients showed a higher percentage of individuals with 16 and 18 repeats than control cases (3.17% vs. 2.64%, and 1.98% vs. 1.64%, respectively). Fisher exact test was used to estimate the association between ALS case/control status, and for all categories of repeats, the probability that ALS and controls are sampled from the same population is not statistically significant. The second panel of the figure shows the obtained odds ratios (ORs), which are also not statistically significant, indicating that there is no increased risk associated with ALS when harbouring any of these CCG intermediate repeat lengths in *AFF3*. The site of onset bar plot suggests that ALS cases with CCG repeats have a higher percentage of spinal onset across all repeats followed by bulbar. Interestingly, our results show that when harbouring more than 8 repeats, there is a higher risk of spinal onset than 7 or fewer repeats (p -value: 0.041). Disease duration stays the same across all repeats without significant variation, between 31-32 months. Interestingly, the age of onset boxplot shows significance when harbouring more than 4, 5, and 6 repeats, suggesting that for the few patients before the cut-off their distribution seems to be different, but this is likely due to chance variation. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

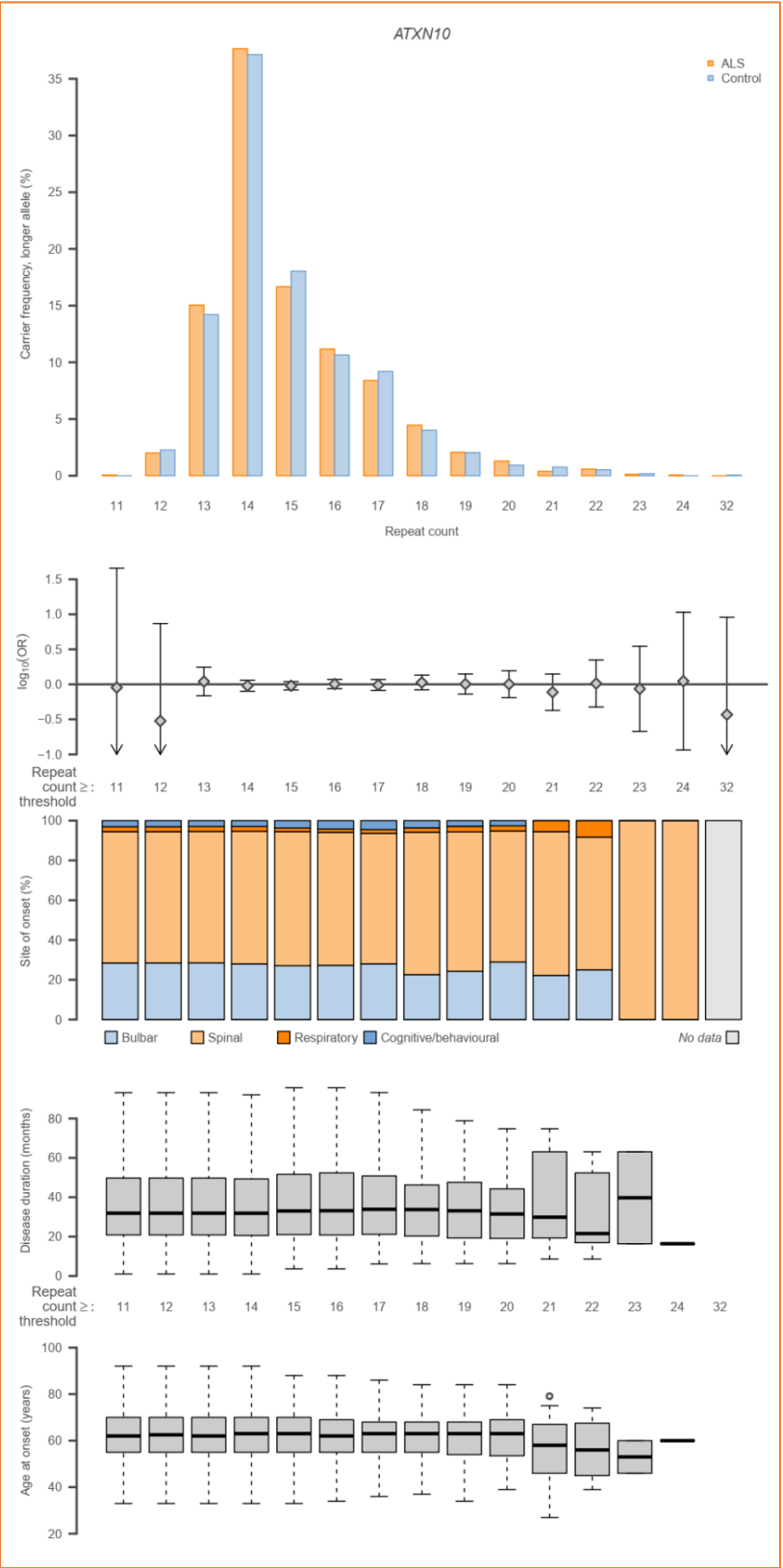


Figure 3-28 The composition of the population of ALS and Control samples for the ATTCT repeat expansion in the *ATXN10* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-28 genotyping results for the *ATXN10* pentanucleotide repeat expansion in 1548 ALS patients and 1718 controls. The highest frequency of the longer ATTCT repeat allele was observed for 14 repeats (37.66% in ALS and 37.14% in controls), followed by 15 repeats (16.67% in ALS and 18.04% in controls). The genotyping protocol used did not amplify expanded pathogenic repeats with more than 280 CAG repeats. ALS patients had a higher percentage of carriers at 13, 16, and 18 repeats compared to control cases; however, Fisher's exact test showed that ALS and control populations were not significantly different in terms of their repeat length distribution at these thresholds (p -values= 0.7198, 0.9073, and 0.6639, respectively). Odds ratios (ORs) calculated from our study did not show any statistically significant increased risk associated with ALS when harbouring intermediate ATTCT repeat lengths in *ATXN10*. Site of onset analysis indicated that spinal onset was more frequent across all repeat lengths, followed by bulbar onset. The mean disease duration did not show significant variation across all repeat lengths. However, the age at onset decreased by 5 years when harbouring more than 21 repeats in the *ATXN10* gene, which was not statistically significant (p -value: 0.089). Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

CACNA1A

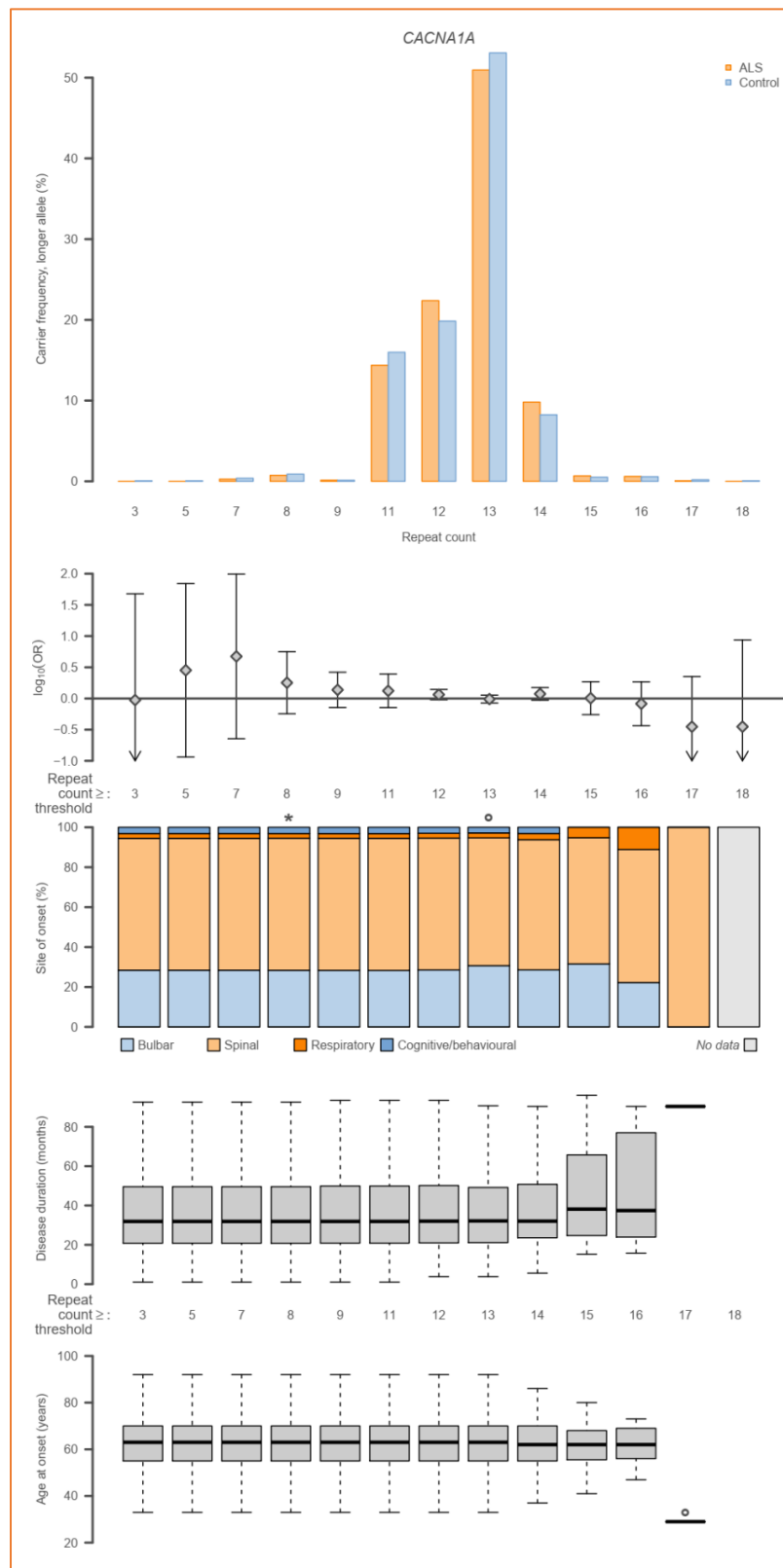


Figure 3-29 The composition of the population of ALS and Control samples for the PolyQ repeat expansion in the CACNA1A gene.

See section 3.2 for a full description of the information contained within the plot.

In Figure 3-29, we present the results of genotyping 1488 ALS patients and 1576 controls for the *CACNA1A* CAG repeat expansion. We found that the highest longer allele frequency for both populations was 13 (50.94% ALS and 53.08% Controls), followed by 12 repeats (22.38% ALS and 19.85% Controls). Interestingly, we observed that ALS patients had a higher percentage of carriers at 12 and 14 (9.81% vs. 8.24%) repeats compared to control cases, but these differences were not statistically significant (Fisher's exact test p -values= 0.1228 and 0.1548, respectively). The ORs obtained for *CACNA1A* were not statistically significant, indicating that there is no increased risk associated with ALS when harbouring any of these CAG intermediate repeat expansions. We also examined the site of onset, disease duration, and age at onset in relation to the genotyped repeats. The site of onset bar plot suggested that ALS cases with CAG repeats had a higher percentage of spinal onset across all repeats followed by bulbar. Disease duration median was consistent across all repeats (approximately 31.9 months), although when harbouring more than 15 repeats the median increases by 6.3 months, although not statistically significantly. Age at onset boxplot showed no significance across all genotyped repeats, with medians between 61 and 62 years. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

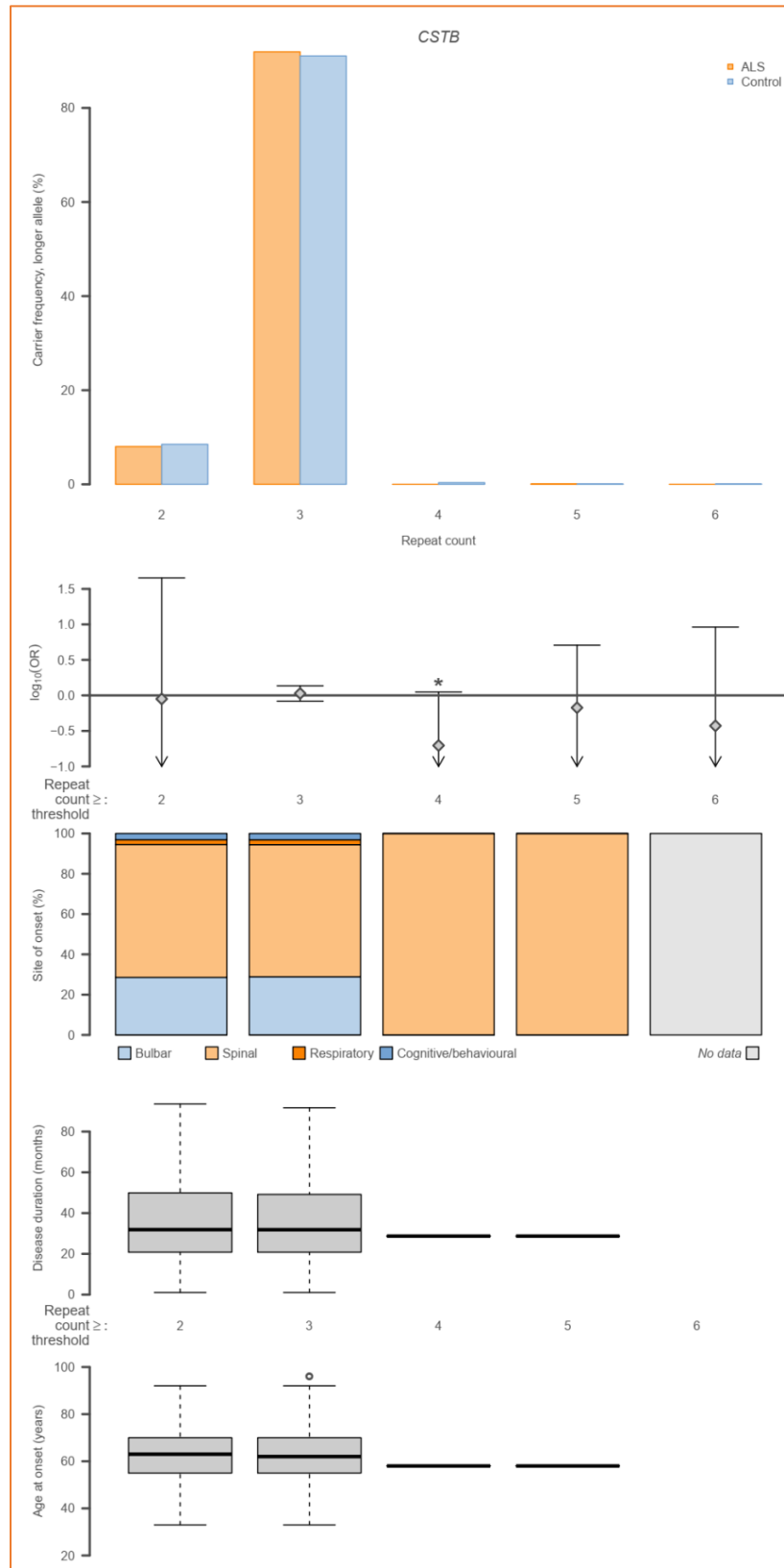


Figure 3-30 The composition of the population of ALS and Control samples for the CCCC GCCCGCG repeat expansion in the *CSTB* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-30 presents the results of genotyping 1535 ALS and 1719 control individuals for the *CSTB* dodecanucleotide repeat expansion. The analysis revealed that the highest longer allele frequency of the CCCC GCCCGCG repeat was 3 and 2 repeats in both populations, with a higher percentage of carriers in ALS patients compared to controls (91.92% vs. 91.04%, and 8.01% vs. 8.49%, respectively). No amplified expansions were detected using this protocol (between 30-125 repeats). Fisher exact test was conducted to assess the association between ALS case/control status. The test revealed that the probability of ALS and controls being sampled from the same population was 1.00 for carriers with 2 repeats and 0.66 for carriers with 3 repeats. The obtained Ors across the repeat counts in our *CSTB* gene study are not statistically significant, indicating no increased risk associated with ALS when harbouring intermediate repeat lengths in *CSTB* (≥ 30 repeats). However, our data suggests a possible protective effect against ALS of 20% (OR= 0.20, with a 95% CI 0.035-1.118) when harbouring 4 or more repeats, which was statistically significant. The site of onset bar plot suggests that ALS cases with CAG repeats have a higher percentage of spinal onset across more than 2 and 3 CCCC GCCCGCG repeats, followed by bulbar. The disease duration (≈ 28.7 to ≈ 31.9 months) and the age at onset showed no significant differences across all genotyped repeats. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

DIP2B

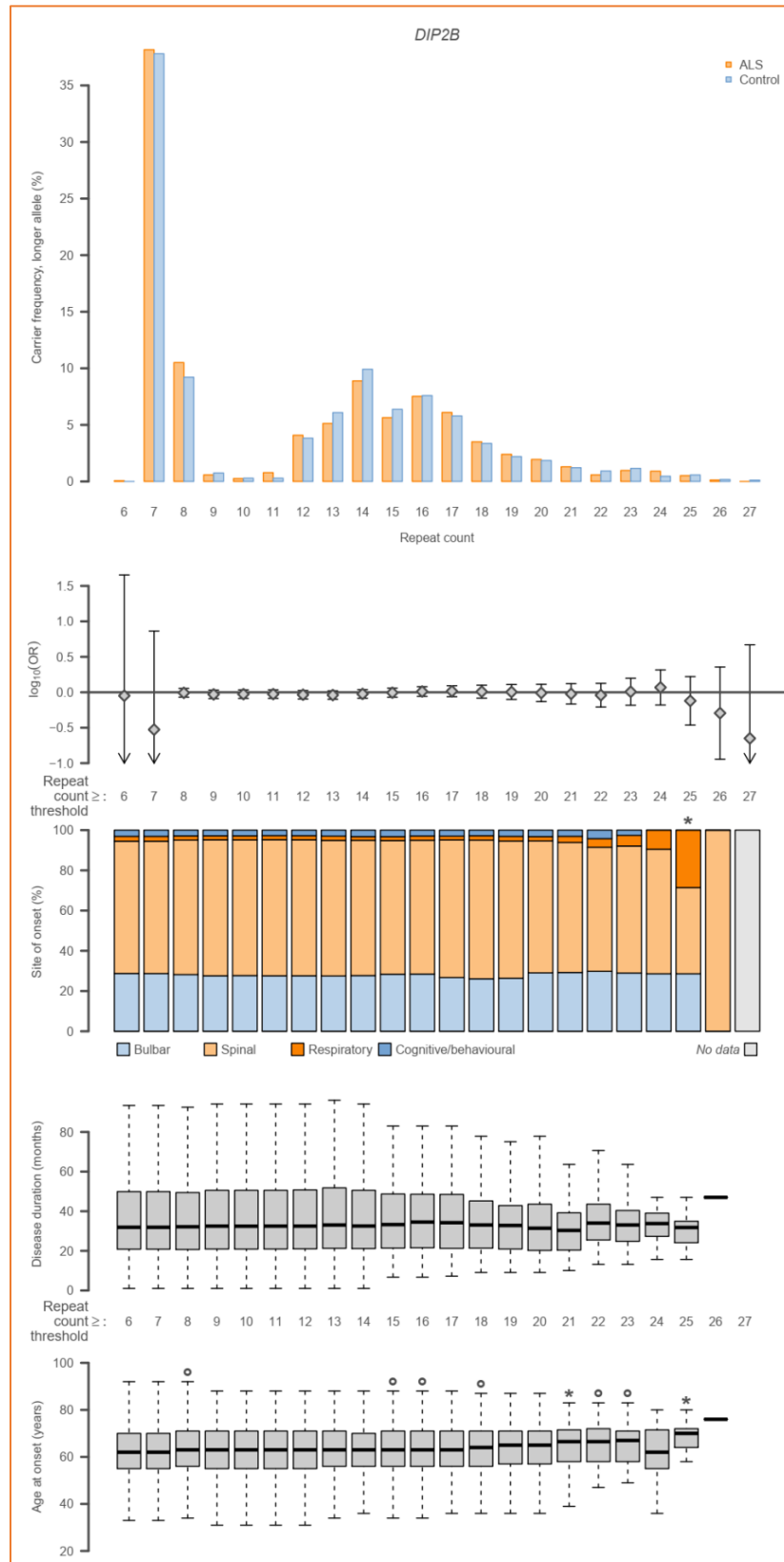


Figure 3-31 The composition of the population of ALS and Control samples for the CGG repeat expansion in the *DIP2B* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-31 presents the results of our study on the *DIP2B* CGG repeat expansion in 1541 ALS and 1725 control individuals. We found that the highest longer allele frequency of the CGG repeat was 7 and 8 in both populations, and ALS patients had a higher percentage of carriers than controls (38.16% vs. 37.80% and 10.51% vs. 9.22%, respectively; $p=0.472$ and 0.829). Our study also found that there is no increased risk associated with ALS when harbouring any of the possible CGG intermediate repeat lengths in *DIP2B* (between 23 and 27). The obtained ORs are not statistically significant. There was no significant difference in site of onset across all expansion lengths except for a higher risk of respiratory onset when harbouring more than 25 repeats, which was statistically significant (p -value: 0.022). We found no significant variation in disease duration across all repeats (median approximately 31.9 months). However, age at onset increases when harbouring more than 21 and 25 repeats by 1.6 and 0.1 years, respectively (p -values: 0.018 and 0.048). Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

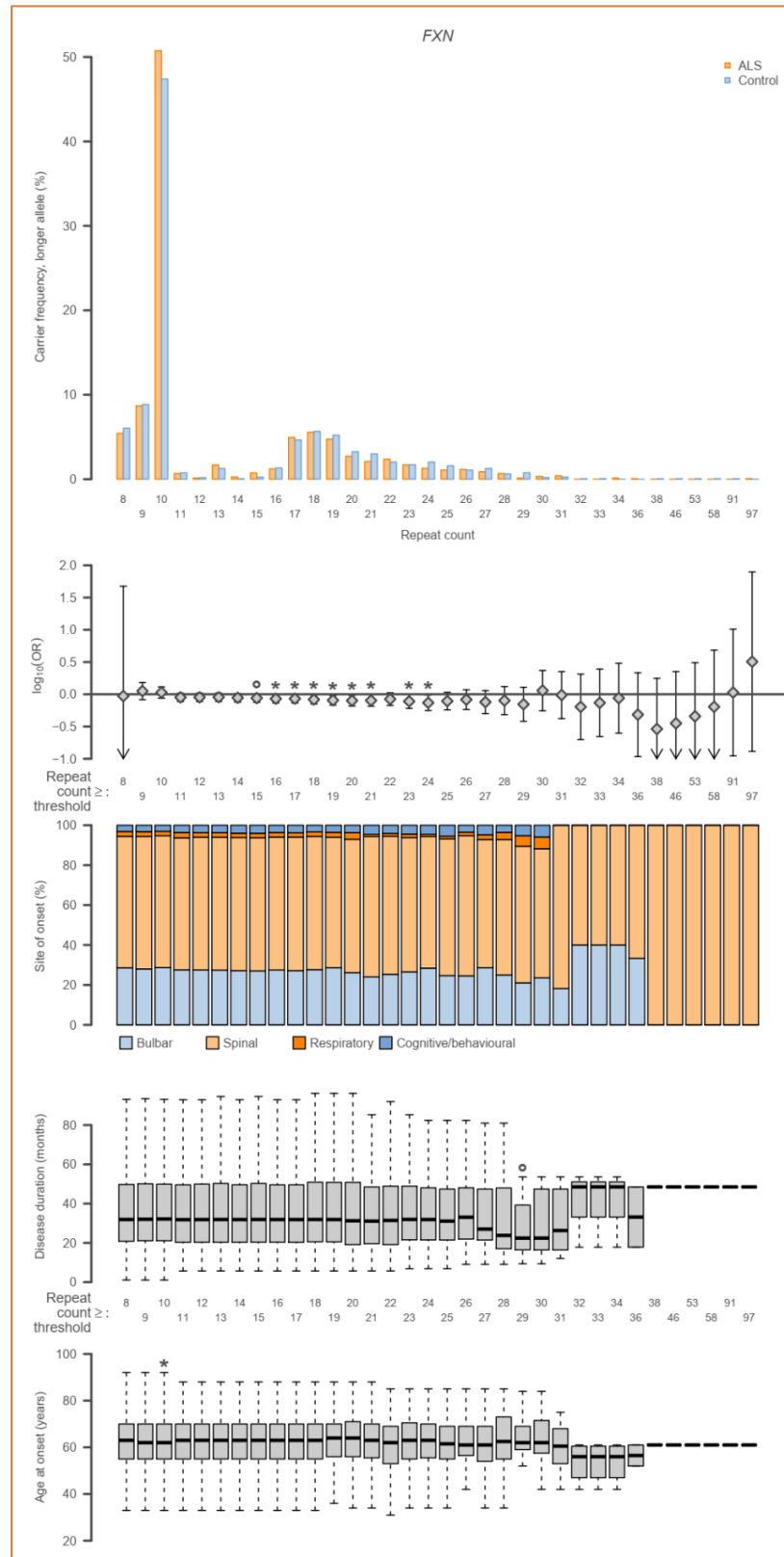


Figure 3-32 The composition of the population of ALS and Control samples for the GAA repeat expansion in the *FXN* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-32 represents the results of 1474 ALS and 1570 control individuals genotyped for the *FXN* GAA repeat expansion. For both populations, the highest longer allele frequency of the GAA repeat was 10 (50.75% ALS and 47.39% Controls). This protocol did not amplify expanded pathogenic repeats (≥ 200 repeats). Fisher exact test was used to estimate the association between ALS case/control status for all categories 10 repeats, the probability that ALS and controls are sampled from the same population is 0.5373. In other words, for *FXN* GAA repeat count, ALS and controls are not significantly different (p -value= 0.5373). The second panel in Figure 3-32 shows the obtained ORs across the repeat counts in our *FXN* gene study. The obtained ORs are not statistically significant, and there is no increased risk associated with ALS when harbouring intermediate repeat lengths in GAA (between 35-65 repeats). Although, when considering repeats between ≥ 15 - ≥ 21 , and ≥ 23 - ≥ 24 , the obtained ORs suggests a possible protective effect against ALS. These results are statistically significant. Site of onset bar plot suggests that ALS cases with GAA repeats have more percentage of spinal onset across all repeats followed by bulbar. Disease duration stays the same across all repeats without significant variation (approximately 31.9 months). Interestingly, there is a decrease in the median of 9.7 months when harbouring more than 29 repeats, but this was not statistically significant (p -value 0.075). Age at onset boxplot showed no significant differences across all genotyped repeats. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

LRP12

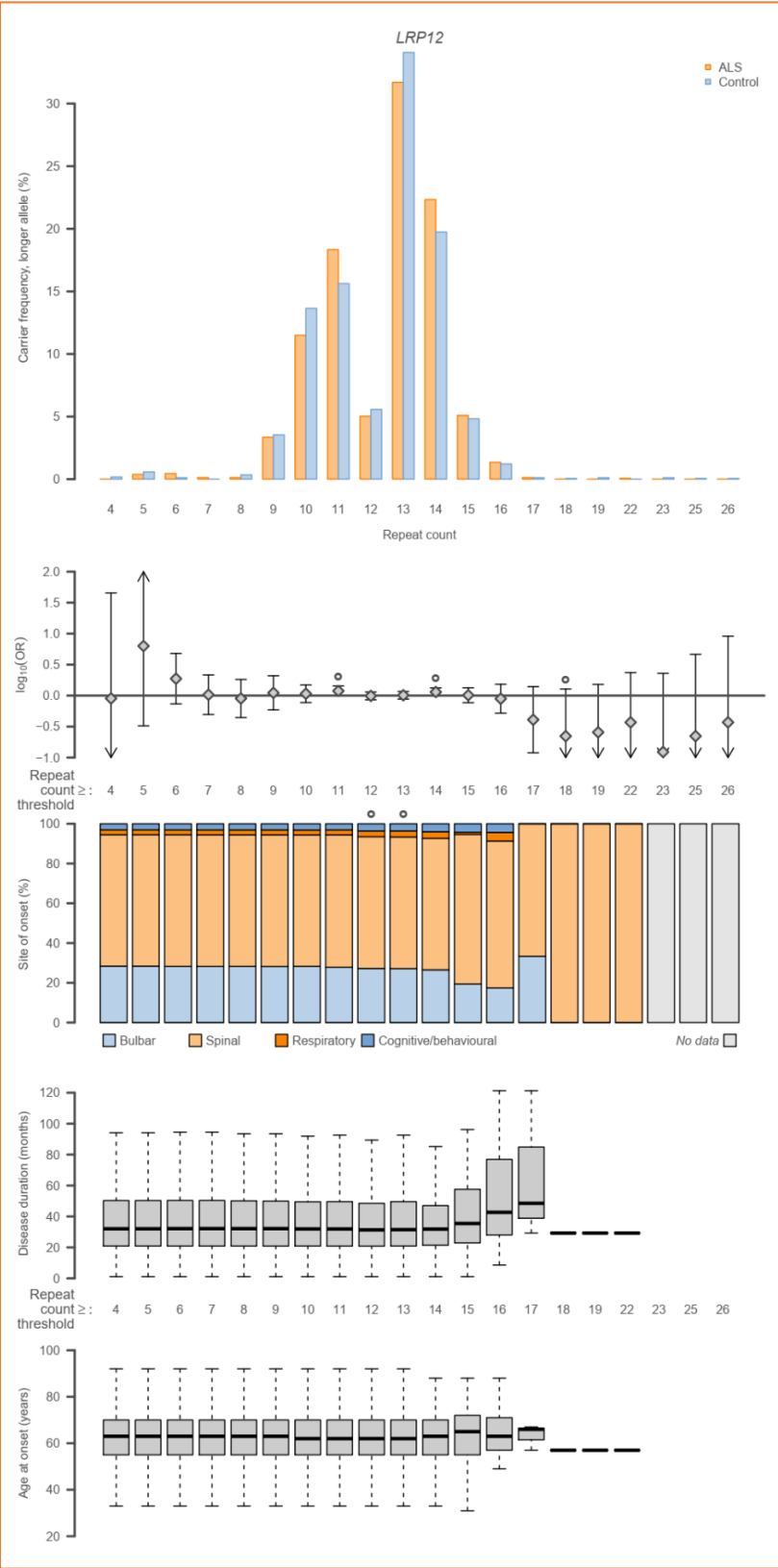


Figure 3-33 The composition of the population of ALS and Control samples for the CGG repeat expansion in the *LRP12* gene

See section 3.2 for a full description of the information contained within the plot.

Figure 3-33 illustrates the results of genotyping 1549 ALS and 1722 control individuals for the *LRP12* CGG repeat expansion. Our study found that the most frequent longer allele of the trinucleotide repeat was 12 (5.04% in ALS and 5.58% in Controls). Furthermore, we observed that ALS patients had a higher percentage of carriers than controls at 11, 14, and 15 repeats, with frequencies of 18.33%, 22.34%, and 5.10% respectively, compared to 15.62%, 17.75%, and 4.82% in controls. Fisher exact test was used to estimate the association between ALS case/control status. Our results indicate that both populations are not significantly different in terms of their repeat (p -values= 0.092, 0.944, and 0.683, respectively). The obtained odds ratios (ORs) from our study for *LRP12* gene do not indicate statistically significant increased risk associated with ALS when harbouring normal-intermediate repeat lengths (between 13-45 repeats). However, when harbouring more than 11 and 14 repeats, ORs suggest an increase in risk of ALS by 18% and 14%, respectively (OR=1.19, 95% CI 0.99 – 1.43; OR=1.14, 95% CI 0.98 - 1.33, respectively). Conversely, for ≥ 18 repeats, the data suggests a possible protective effect against ALS (OR=0.22, 95% CI 0.038 - 1.28). Site of onset bar plot showed that ALS cases with CGG repeats have a higher percentage of spinal onset across all repeats, followed by bulbar. Disease duration remained relatively stable across all repeats without significant variation, with an average of approximately 31.9 months. Nonetheless, we observed a slight increase in disease duration with more than 15, 16, and 17 repeats, which could be attributed to the small amount of data available for those repeats. Age of onset box plot showed no significant differences across all genotyped repeats, with an average of ≈ 63 years. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

TCF4

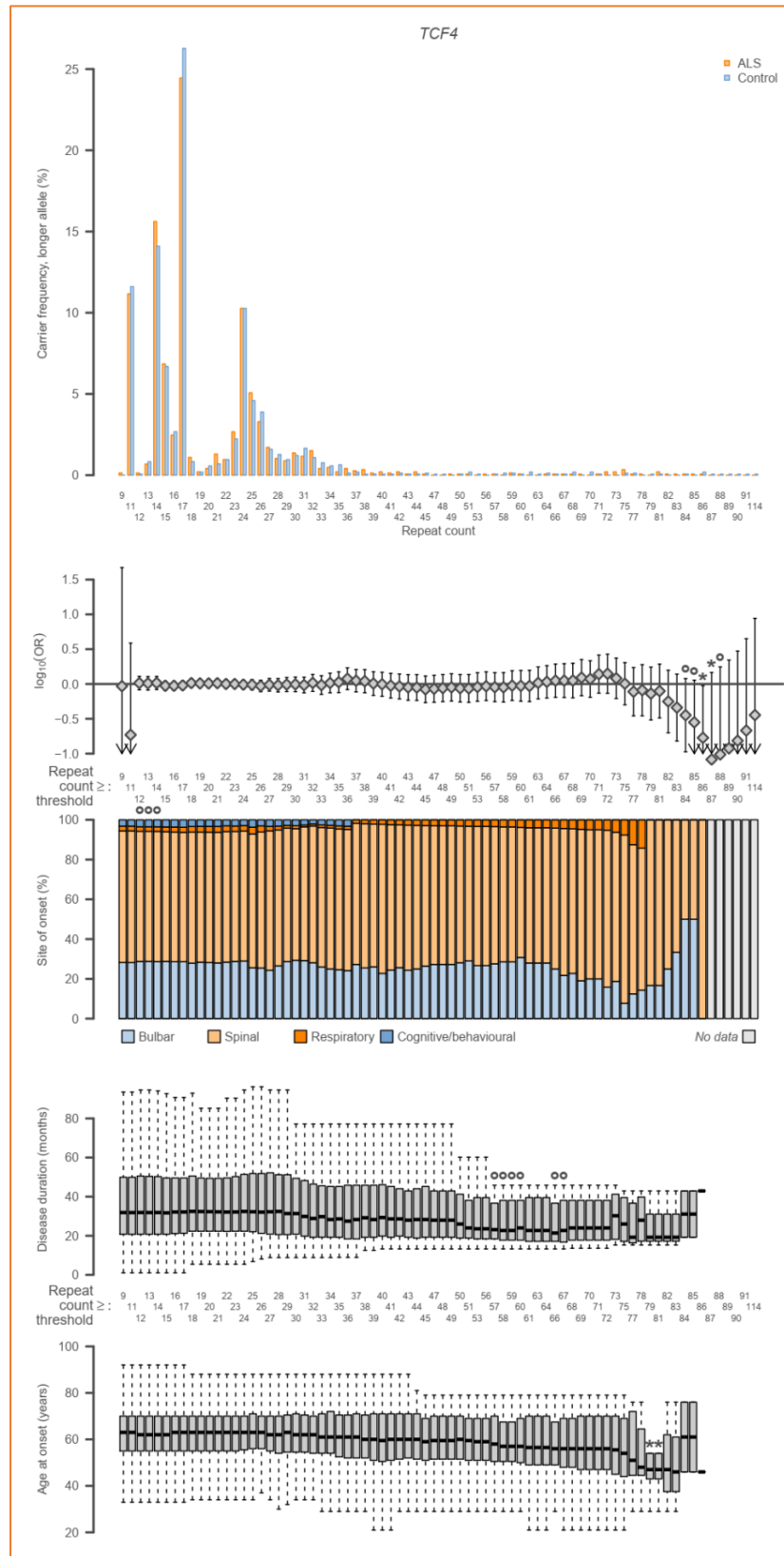


Figure 3-34 The composition of the population of ALS and Control samples for the CTG repeat expansion in the *TCF4* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-34 presents the findings of our study on the *TCF4* CTG repeat expansion in 1460 ALS and 1568 control individuals. We observed that the highest longer allele frequency of the trinucleotide repeat was 17, with a frequency of 24.45% in ALS and 26.28% in controls, followed by 14 repeats. We utilized a protocol that amplified expanded pathogenic repeats (≥ 50 repeats). No repeat expansion length was associated with ALS risk ($p > 0.05$ for all thresholds). Our analysis of the obtained odds ratios (ORs) across the repeat counts in *TCF4* indicated that there was no increased risk associated with ALS when harbouring intermediate repeat lengths in *TCF4* (between 40-47 repeats). However, the data suggests a possible protective effect against ALS when harbouring more than 86 and 87 CTG repeats, respectively ($OR_{\geq 86}=0.17$, 95% CI 0.3-0.95, and $OR_{\geq 87}=0.08$, 95% CI 0.005-1.46). Both of these results were statistically significant (Supplementary Table S 3-2). The site of onset bar plot demonstrated that ALS cases with CTG repeats had a higher percentage of spinal onset across all repeats, followed by bulbar onset. We observed that an intermediate range up to 78 CTG repeats appeared to be associated with increasing rates of respiratory onset ALS, although this was not statistically significant. We also found that disease duration remained consistent across all repeats, with no significant variation (approximately 31.9 months). The age of onset box plot showed no significant differences across all genotyped repeats, with an average of approximately 63 years. We observed a decrease in disease duration and age of onset as repeats continued to expand, but we believe that these findings were likely due to chance and the design of the test. Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

ZNF713

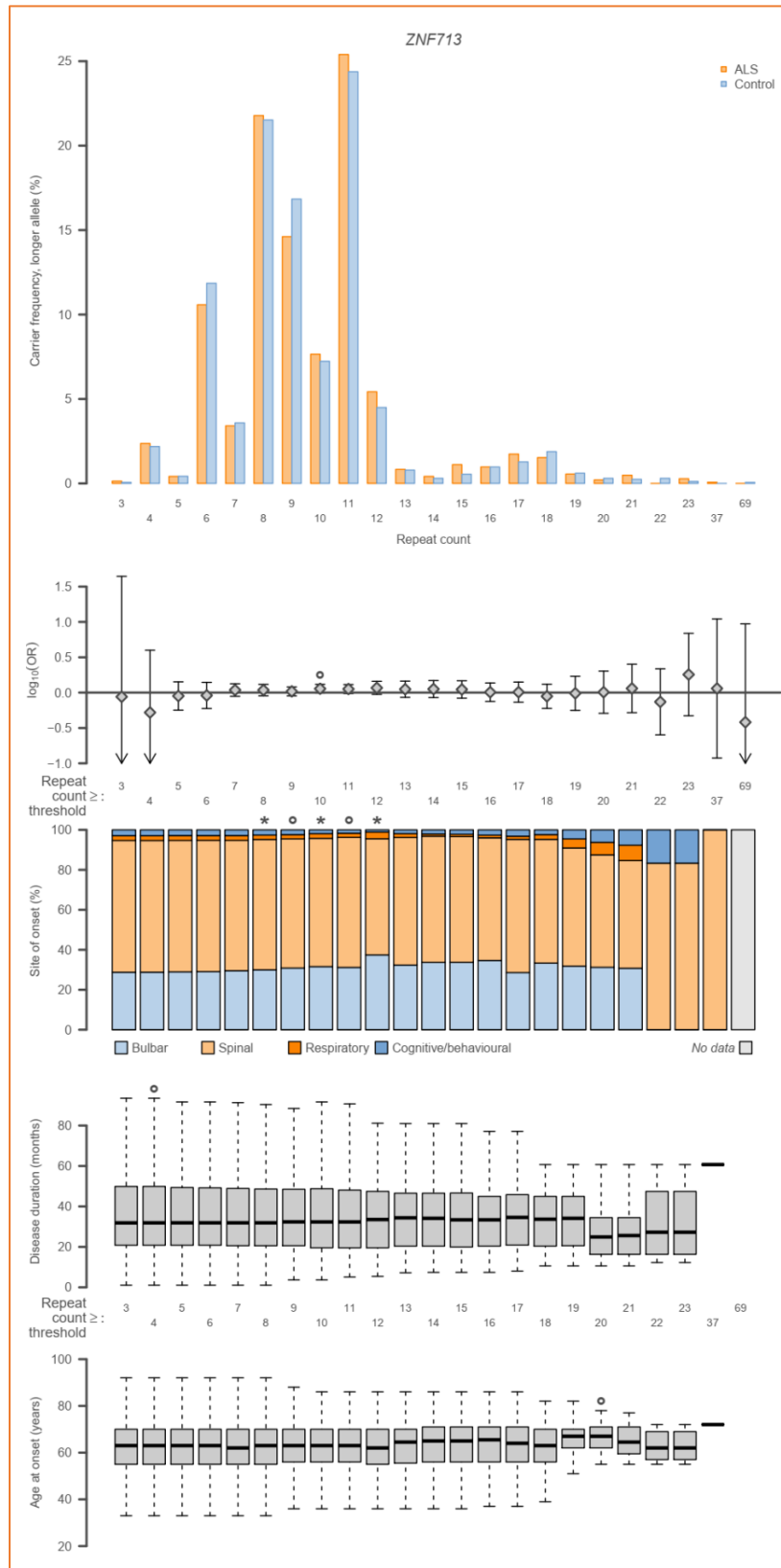


Figure 3-35 The composition of the population of ALS and Control samples for the CGG repeat expansion in the *ZNF713* gene.

See section 3.2 for a full description of the information contained within the plot.

Figure 3-35 shows results for 1437 ALS and 1645 control individuals genotyped for the *ZNF713* CGG repeat expansion. For both populations, the highest longer allele frequency of the trinucleotide repeat was 11 (25.40% ALS and 24.37% controls), followed by 8 repeats (21.78% ALS and 21.52% controls) in which ALS cases showed a higher percentage of CGG repeats. Additionally, carriers with 10 and 12 had a higher allele frequency in ALS cases than Controls (7.66% vs. 7.23%, 5.43% vs. 4.50%, respectively; $p = 0.08$ and 0.16). This protocol did not amplify expanded pathogenic repeats (≥ 85 repeats). The obtained ORs are not statistically significant, and there is no increased risk associated with ALS when harbouring any of these CGG intermediate repeat lengths in *ZNF713*. Site of onset bar plot suggests that ALS cases with CGG repeats have more percentage of spinal onset across all repeats followed by bulbar. Site of onset significantly varied when considering more than 8, 10 and 12 CGG repeats (p -values < 0.05). Disease duration stays the same across all repeats without significant variation (approximately 31.9 months). Interestingly, the mean of survival drops by 7 months, but this was not statistically significant ($p = 0.12$). Age at onset box plot showed no significant differences across all genotyped repeats, with an average of approximately 63 years. Interestingly, there is an increase in the median of 4 years when harbouring more than 20 repeats, but this was not statistically significant (p -value 0.096). Supplementary Table S 3-3 and Supplementary Table S 3-4 provides additional information on the statistical analysis of these results.

Genotyping results (FTD)

The general FTD population in the DNA bank was 44 patients, which is a small cohort of cases to assess, with low corresponding statistical power. Nevertheless, we genotyped all patients using the multiplex PCR panel to explore the contribution of these repeat expansions to FTD risk.

Figures 3.37 and 3.38 show the 7 most relevant results obtained for the longer allele for FTD cases, but given the small sample size, the results are just possible suggestions for FTD risk. Results must be replicated in a more substantial FTD population to validate any apparent association between expanded alleles and FTD risk. The plots for each gene for the shorter allele can be found in Supplementary Figure S 3-2, and for the longer alleles in Supplementary Figure S 3-3.

Among the most relevant results for the longer allele, for the *AR* gene, there is a detectable increase in the relative percentage of cases with ≥ 27 , ≥ 30 and ≥ 32 CAG repeats (Figure 3-36, Supplementary Table S 3-5). There is an overall increase in FTD risk up to 10.64 times when harbouring intermediate repeats between ≥ 27 and ≥ 32 ($OR_{\geq 32}=10.64$, 95% CI 1.71-66.33). Most FTD cases have 10 or more CGC repeats for the *ARX* gene, and FTD risk increases by 8.5 times ($OR=8.50$, 95% CI 1.66-43.50) (Figure 3-36, Supplementary Table S 3-5). For the *ATXN1* gene, there was an upsurge in FTD patients with 34 CAG repeats, with an associated increase of FTD risk by two times when harbouring 34 or more repeats ($OR=2.27$, 95% CI 0.96-5.36) (Figure 3-36, Supplementary Table S 3-5). A similar result was found for the *ATXN3* gene, harbouring more than 33 and 35 CAG repeats that appear to increase FTD risk by almost four times ($OR_{\geq 33}=3.70$, 95% CI 0.97-14.07), $OR_{\geq 35}=3.96$, 95% CI 0.72 – 21.98) (Figure 3-36, Supplementary Table S 3-5).

For the CCTG repeats in the *CNBP* gene, an increase of cases showed 23 repeats, which suggests a possible increase of FTD risk by almost nine times ($OR_{\geq 23}= 8.75$, 95% CI 1.45-52.98) (Figure 3-37, Supplementary Table S 3-5). Another relevant result was from the *C9orf72* gene, with intermediate GGGGCC repeats between ≥ 13 and ≥ 16 having an overall increase of FTD risk up to 8 times ($OR= 8.05$, 95% CI 2.00-32.39) (Figure 3-37; Supplementary Table S 3-5). Finally, in the *NUTM2B-AS1* gene there is a slight increase in FTD cases with 15 or more repeats with an

associated increase of FTD risk by almost five times (OR=4.72, 95% CI 0.85-26.39) (Figure 3-37, Supplementary Table S 3-5).

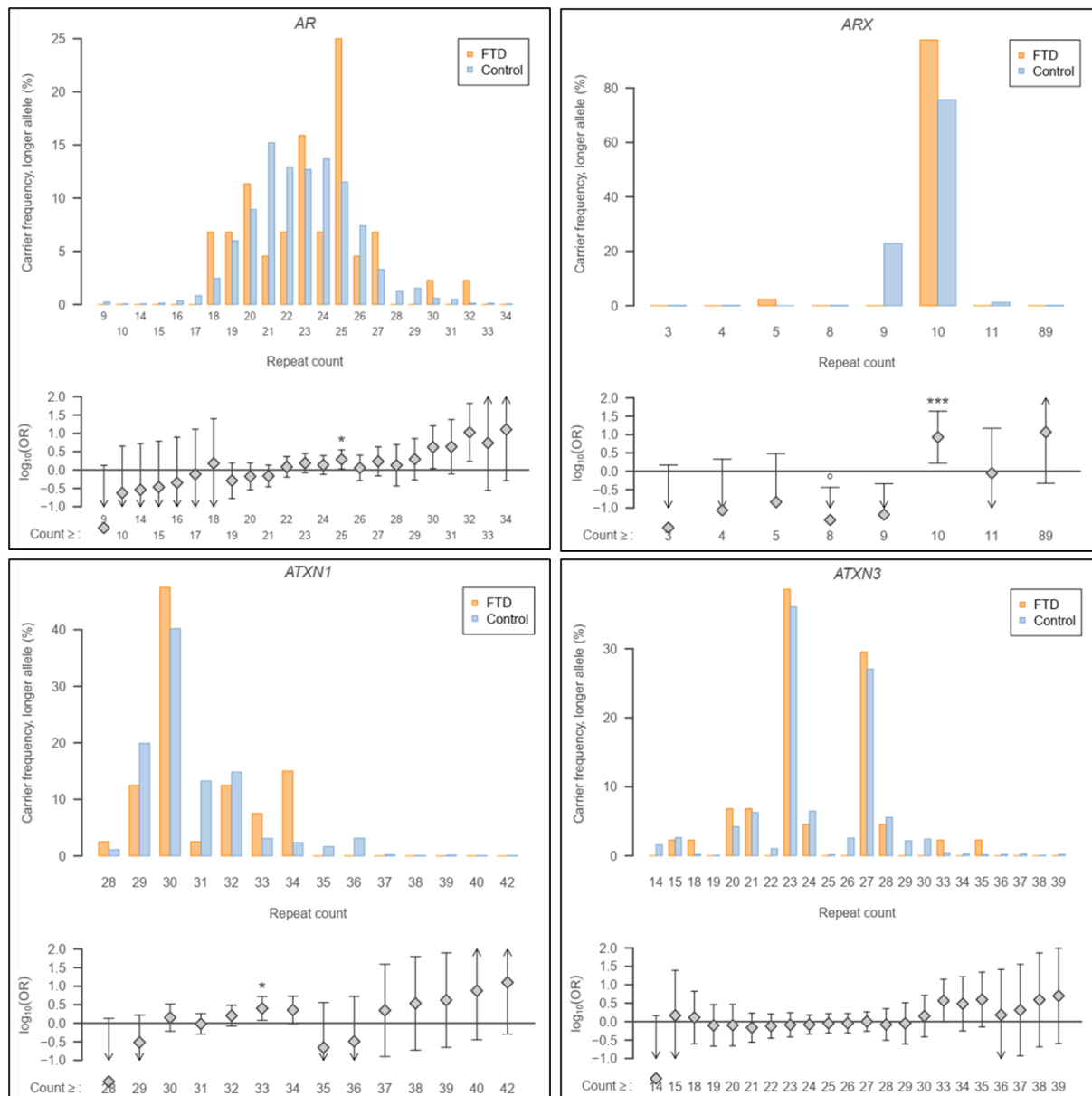


Figure 3-36 The composition of the population of FTD and control samples for the repeat expansions in the *AR*, *ARX*, *ATXN1* and *ATXN3* genes.

For each gene, top panel: allele frequencies of each allele in 44 FTD, and >1500 controls individuals genotyped for each repeat expansion in the *AR*, *ARX*, *ATXN1* and *ATXN3* genes. Bottom panel: the obtained ORs thresholded across the repeat counts found in our study for each gene.

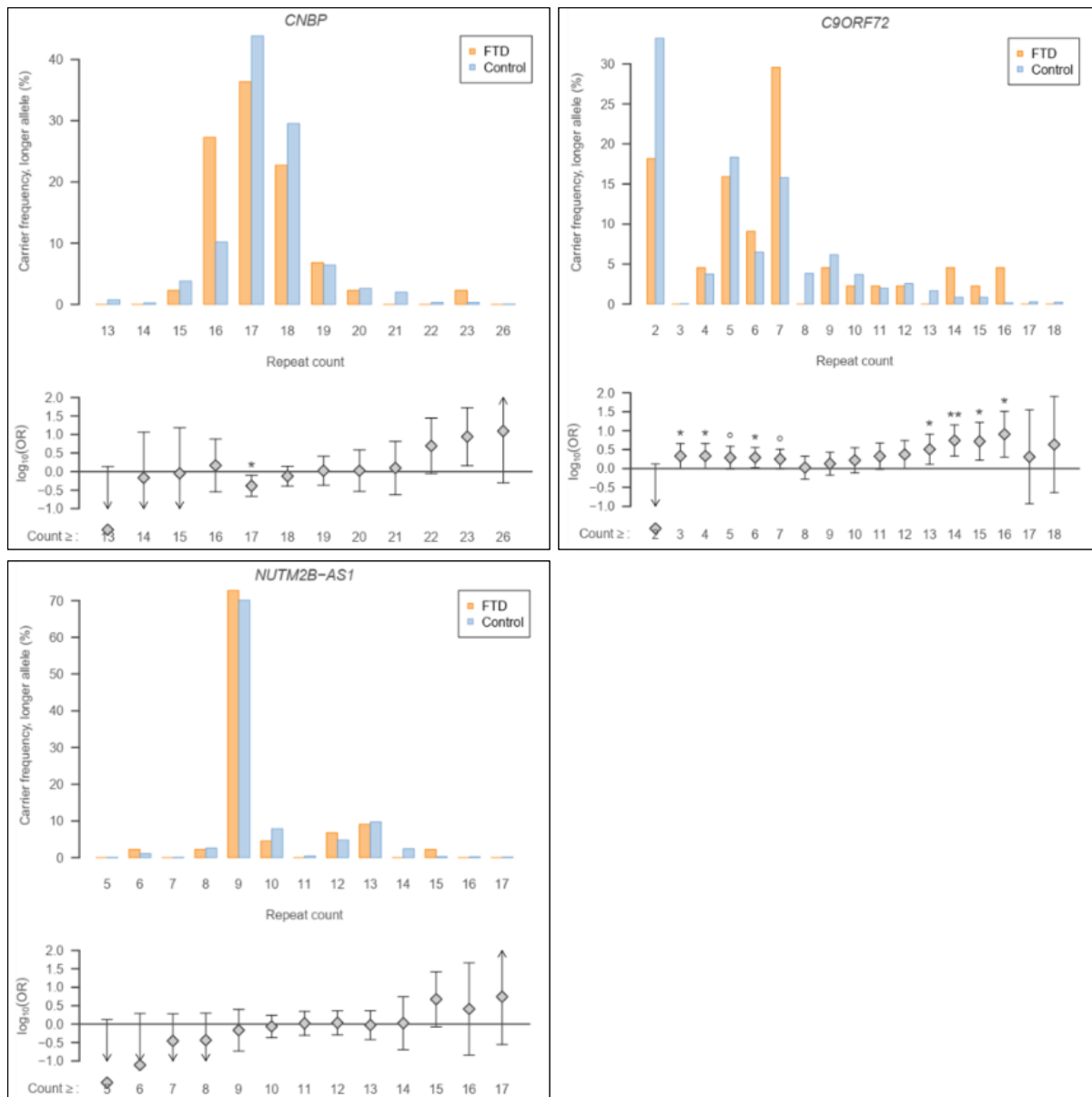


Figure 3-37 The composition of the population of FTD and control samples for the repeat expansions in the *CNBP*, *C9orf72* and *NUTM2B-AS1* genes.

For each gene, top panel: allele frequencies of each allele in 44 FTD, and >1500 controls individuals genotyped for each repeat expansion in the *CNBP*, *C9orf72* and *NUTM2B-AS1* genes. Bottom panel: the obtained ORs thresholded across the repeat counts found in our study for each gene.

3.4 Discussion

STR expansions play a significant role in human phenotypic variation and disease. STR genotyping and identifying the pathogenic REs ranges hold great promise to widen our understanding of neurological diseases, such as ALS, and potentially reduce diagnosis time and revolutionise patient treatment. In this association study of 35 RE loci in 1550 ALS and 1719 age and population-matched controls, results have been divided in five categories according to their apparent association with ALS risk. The precise intermediate repeat length cut-off for ALS may differ depending on the population.

Previously reported ALS risk genes

Firstly, this study replicates in an Irish population previously reported ALS risk genes, which includes *ATXN1*, *ATXN2*, *C9orf72*, *HTT*, and *NIPA1*.

Expanded CAG repeats in *ATXN1* gene (>39) causes SCA1, while intermediate repeat lengths (~ 32 -38) have been associated with ALS risk (Lattante et al. 2018; Tazelaar et al. 2020) (Table 3-1). The most common allele for the PolyQ repeat was 29 repeats, which corresponds with the normal range for this gene between 6-32 repeats (Orr et al. 1993). In our Irish population study, there is an increase of ALS risk between 29-38 intermediate repeats, with a higher increase of ALS by 2 times when considering ≥ 29 and ≥ 37 , although with the current sample size this was not statistically significant. Similar results were obtained in previous reports in other populations (Lattante et al. 2018; Tazelaar et al. 2020).

PolyQ expansions in *ATXN2* are associated with three clinically distinct neurological diseases, SCA2, ALS and Parkinson's disease (Table 3-1). These different situations are correlated with distinct repeat lengths and sequence configuration of the DNA repeat region. Although the precise repeat length cut-off for ALS may differ depending on the population (Elden et al. 2010), in this study the intermediate length of repeat expansion in *ATXN2* used was between 24 and 34. The highest increase of ALS risk was by 4.5 times when harbouring ≥ 30 CAG repeats in the *ATXN2* gene in the Irish population, which was statistically significant (OR=4.52, 95% CI 2.58-7.94).

These data have also contributed towards novel neuroimaging results which show that cerebellar degeneration in ALS is not driven by *ATXN2* expanded alleles. (Bede et al. 2021).

Harbouring more than 30 GGGGCC RE in *C9orf72* is a high-risk factor for FALS cases by 11 times, familial FTD by 7.1 times (Xi et al. 2012) and has been detected in other neurodegenerative diseases such as Parkinson's, Alzheimer's and Huntington's disease-like disease (Table 3-1) (Beck et al. 2013; Cacace et al. 2013; Lesage et al. 2013; Woollacott and Mead 2014).

The conventional approach to determine the expanded number of hexanucleotides repeats for *C9orf72* is by rpPCR. Nevertheless, the presented study is interested in the intermediate repeat length and its association with ALS risk and by normal PCR these repeats can be amplified. There is 7.8× increase of ALS risk in the Irish population when harbouring ≥ 19 GGGGCC repeats in the *C9orf72* gene (OR=7.84, 95% CI 0.41-151.84). These results are not statistically significant with this sample size, but are similar to previous studies showing that intermediate-length alleles of ≥ 17 and ≥ 20 repeats may increase the risk of *C9orf72*-ALS by 5 and 15 times, respectively (Kaivola et al. 2020), and parkinsonism when harbouring 20-29 repeats (Cannas et al. 2015).

The normal length of the CAG in *HTT* is below 27 repeats, and alleles with 40 or more repeats causes HD, a neurodegenerative disease (Table 3-1). However, the disease may also occur with 36-39 repeats and the age of onset is later with a slower progression (Testa and Jankovic 2019). The results obtained in the present study showed an increase of ALS risk up to 30% when harbouring intermediate CAG repeats between 29-38 repeats. Also, it was observed an overall increase of ALS risk in the pathogenic RE length (>39), with the highest risk by 5.5 times when harbouring ≥ 41 CAG repeats in *HTT* gene. Even though these results are not statistically significant for the Irish population, Dewan and collaborators in 2021 found *HTT* CAG expanded alleles (with 40-64 repeats) in 0.12% of patients previously diagnosed with FTD/ALS (Dewan et al. 2021). These results were also replicated in Italian patients with ALS, and results were similar (0.13% of patients had the pathogenic RE in *HTT* gene) (Manini et al. 2022).

NIPA1 mutations have been shown to cause spastic paraplegia type 6 (SPG6) (Svenstrup et al. 2011), but the CGC repeat region has not yet been characterised. Normal alleles for this repeat

expansion in a healthy population are between 6 to 10 repeats. Harboring 8 repeats was associated with increasing susceptibility to ALS (Table 3-1) (Tazelaar et al. 2019). In this study, results showed a statistically significant increase of ALS risk of 27% in the Irish population when harboring 9 or more CGC repeats in *NIPAI*. Similarly, there is a tendency of increase of ALS risk when harboring ≥ 10 and ≥ 15 , but results are not statistically significant.

Novel ALS risk genes

The category of novel ALS risk genes includes *ARX*, *FRA10AC1*, and *PABPN1*, in which their RE is statistically significantly associated risk with ALS in the Irish population, and for this reason are a novelty in the literature.

The *ARX* gene encodes a transcription factor belonging to the homeobox gene class (Table 3-1). The second exon of the gene is a mutational hotspot as it contains a GCG repetitive region (Jorge et al. 2013; Sato et al. 2009). Since its discovery, several mutations have been identified, including deletions, duplications, insertions, nonsense, missense and splice mutations, leading to multiple phenotype spectrums, including XLMR, DEE1 and Partington Syndrome (Bienvenu et al. 2002; Jorge et al. 2013; Strømme et al. 2002). Normal alanine tracts are 16 GCN (7-10 GCG) (Chintalaphani et al. 2021), and the pathogenic range is more than 23 for XLMR (Demos et al. 2009), and more than 20 for DEE1 and PRTS (Bienvenu et al. 2002; Chintalaphani et al. 2021; Strømme et al. 2002). Lack of functional ARX proteins disrupts cell differentiation and interrupts gene regulation required for interneuron migration (Table 3-1) (Kitamura et al. 2002). The present study found an increase in ALS risk over two times when harboring more than 17 GCN (11 pure GCG) repeats in the *ARX* gene in the Irish population. Interestingly, when harboring just one GCN less, there is a protective effect against ALS. These results prove how RE can switch from benign to pathogenic by causing risk for a disease through a change in their RNA and protein structure, function, and interaction with other molecules in the cell and DNA.

The *FRA10AC1* gene is a rare folic acid-type fragile site. No distinct phenotype has been associated with the expression of *FRA10AC1*, but some researchers proposed that *FRA10AC1* might be associated with neurological disorders, intellectual disability or tumorigenesis (von Elsner et al.

2022; Li et al. 2015) (Table 3-1). The encoded protein is a phosphoprotein with an unknown function. This gene contains a CpG island with a CGG repeat region, normally 8-14 repeats, and could expand over 200 (Table 3-1). The fragile site becomes hypermethylated when the repeat region expands, causing gene expression reduction (Li et al. 2015). In a genome-wide association study to identify genetic factors influencing the levels of cerebrospinal fluid (CSF) amyloid beta, A β ₁₋₄₂, in patients with AD, Lin and colleagues in 2015 found genome-wide significance between the CpG island at the CGG repeat region and increase of the level of CSF A β ₁₋₄₂. A β is the main component of amyloid plaques found in the brains of patients with AD. A β molecules can aggregate, forming flexible soluble oligomers, which are toxic to nerve cells (Herrera-rivero et al. 2010; Li et al. 2015). Also, A β can induce tau protein to misfold, which underlies many neurodegenerative diseases, including tauopathies in which the misfolded tau protein pathology multiplies through the brain circuits in a prion-like manner (Hallinan et al. 2019). For the Irish population, we found that harbouring intermediate CGG expansions in the *FRA10AC1* is significantly associated with ALS risk when harbouring repeats in the ≥ 17 - ≥ 24 range.

The *PABPN1* gene provides instructions for making a poly-A binding protein found throughout the body. PABPN1 protein helps add adenines to the poly-A tail of the mRNAs and regulating mRNA production, transportation out of the nucleus and breakdown of poor-quality mRNA. *PABPN1* contains a poly-(GCN) tract with normally 6 repeats (Table 3-1). Harbouring more than 7 repeats has been associated with OPMD, which frequently have weakness in the proximal muscles. *PABPN1* protein plays a key role in the formation of nuclear condensates of TDP-43 and FUS and is moderately expanded in the Irish ALS cases, increasing the risk by 4 times when harbouring 7 or more CGN repeats.

Probable ALS risk

Thirdly, nine genes are categorised as probable ALS risk. This classification tends to increase the risk of ALS as the repetitions in each allele become greater or in their intermediate repeat range, and the results are almost statistically significant. Some of the genes included in this category were associated previously with susceptibility of ALS and others are novel for ALS risk. These genes

are involved in several diseases, and some phenotypes include impaired cognitive function, spinal and bulbar atrophy, speech difficulties, dystonia, spasticity, rigidity, tremors, muscle weakness and dysarthria (Table 3-1). The genes' phenotypes and disease mechanisms lead to different types of neurodegeneration like in ALS. Therefore, the study of these genes is essential for understanding pleiotropy in ALS and its disease mechanism and must be replicated in a larger sample size and in a different population to increase power.

A CCG RE causes fragile XE syndrome (FRAXE) in *AFF2*. The most common alleles range from 6-25 repeats, and the pathogenic range for this disease is more than 200 (Knight et al. 1993). Results from this study in the Irish population suggest an increase in ALS risk when harbouring intermediate CGG repeat lengths between 31-54, with an increase of ALS risk by almost six times with 33 or more repeats. Results are not statistically significant, but this evidence of an association between ALS and *AFF2* repeat expansions may be interconnected with previously obtained results by Yuva-Aydemir and collaborators in 2019. The researchers found that the transcription factor *AFF2/FMR2* modulates the toxicity and transcription of expanded GGGGCC repeats in C9ORF72-ALS-FTD in human neurons (Yuva-Aydemir et al. 2019) (Table 3-1).

PolyQ repeat expansion in the *AR* gene causes X-linked spinal and bulbar atrophy (SBMA; Kennedy disease). The pathogenic range harbouring at least 40 or more CAG repeats, while the normal allele is between 9-35 repeats (Lieberman et al. 2018; Rohilla and Gagnon 2017; La Spada et al. 1991). Many diseases have been associated with *AR*, including partial and spinal bulbar muscular atrophy, when harbouring expanded PolyQ repeats (Monks et al. 2008) (Table 3-1). Results obtained in the Irish population suggest (although not statistically significant with this sample size) a susceptibility to ALS associated with between 32 to 52 repeats and an increase of ALS risk by 10 times when harbouring more than 32 CAG repeats in the *AR* gene. ($OR_{\geq 32}=10.32$, 95% CI 0.56-191.90).

Expanded RE in *ATXN3* cause spinocerebellar ataxia type 3 (SCA3), known as Machado-Joseph Disease (MJD). Normal range of the PolyQ expansions are between 13-36, whereas pathogenic repeats are more than 56 CAG repeats (Kawaguchi et al. 1994; McLoughlin, Moore, and Paulson 2020). The overall results obtained in this project suggest intermediate repeats in *ATXN3* may

increase ALS risk (between ≥ 31 and ≥ 41 repeats) with a higher increment by almost 3 times with 38 or more CAG repeats, although this was not statistically significant. Similar results have been previously reported for *ATXN3* gene, in which suggests an increase of ALS risk in the presence of intermediate repeats (Humphrey et al. 2021) (Table 3-1).

The opposite strand of the *ATXN8* gene (*ATXN8OS*) transcribes an mRNA that is associated with the neurodegenerative disease SCA8. CTG RE leads to a gain-of-function protein causing the disease (Table 3-1). The cut-off values for the intermediate and pathogenic lengths of repeats is debatable in the literature, for this study the consensus we decided to follow was a normal range from 16-34, and pathogenic repeats ≥ 80 . Results suggest an overall increase in ALS risk conferred by intermediate repeats between ≥ 29 and ≥ 44 with a higher risk by 3 times when harbouring ≥ 32 CTG repeats. These results were similar to a previous study by Hirano and collaborators in 2018, in which they reported intermediate CTG repeats in the *ATXN8OS* in 3 % of their ALS patients from Japan (Hirano et al. 2018).

Tetranucleotide repeat expansion in the *CNBP* gene causes DM2 via a toxic RNA gain of function mechanism (Christina L. Liquori et al. 2001) (Table 3-1). Individuals with a (CCTG)_n uninterrupted tracts have shown to be more unstable than interrupted, and this instability has been correlated with uninterrupted tracts (Bachinski et al. 2009). More than 50 repeats is considered pathogenic, whereas the range from 30 to 49 repeats are considered to be intermediate range in which the cut-off of stable and unstable allele is at 30 CCTG (Radvanszky et al. 2013). *CNBP* is considered in the present study as a probable novel ALS risk gene.

FXS and FXTAS are two clinically distinct diseases caused by CCG RE in the *FMR1* gene (Mishra et al. 2019). Researchers believe that overproduction of abnormal mRNA may bind to other proteins altering multiple protein and mRNA functions in the cell which develops multiple clinically distinctive diseases (Table 3-1). The amount of CGG length in the *FMR1* gene can determine the diagnosis of the patient (Salcedo-Arellano et al. 2020). For both diseases, the normal (5-53 repeats) allele lengths are the same but differ in their pathogenic lengths. FXS range between 55 to 200 repeats, and FXTAS from 200 or more repeats (Aziz et al. 2003; Mishra et al. 2019). Even though the partial results obtained in the study need to be replicated using rpPCR, results

may suggest an increase of ALS risk in the repeat length between 36 and 79 repeats, albeit not statistically significant with the current sample size. *FMR1* is likely a novel ALS risk gene, which may have a big impact since is an X-Linked gene and its disease mechanism may partially explain the imbalanced sex ratio in ALS.

The presence of methylation and size mosaicism has been a recurring observation in a substantial proportion of individuals bearing a full mutation, as documented in numerous studies (Genç et al. 2000; McConkie-Rosell et al. 1993; Nolin et al. 1994; Pretto et al. 2014; Tassone, Iwahashi, and Hagerman 2004). Although mosaicism can yield some level of Fragile X mental retardation protein (FMRP) production contingent on methylation status and CGG allele size, it is often associated with developmental delays. This phenomenon likely arises from the constrained availability of FMRP, attributable to diminished gene expression and the intricacies of translating residual mRNA bearing the expanded CGG repeat (Ludwig et al. 2014; Primerano et al. 2002).

Furthermore, the conventional practice of DNA testing in peripheral blood mononuclear cells (PBMCs) may not entirely reflect the mutational profile exhibited in other tissues. Mosaic patterns across diverse tissues have been the subject of scrutiny in several studies (Bonarrigo et al. 2014; Genç et al. 2000; Pretto et al. 2014), revealing instances of similarity as well as marked disparities. Consequently, predicting whether mosaic patterns observed in blood can reliably mirror those in other tissues, particularly the brain, and thereby infer clinical presentation and severity patterns on an individual basis, remains challenging. Nonetheless, these studies collectively underscore an association between mosaicism and prognosis (Genç et al. 2000; Govaerts et al. 2007; Han et al. 2006; Merenstein et al. 1994; Mueller et al. 1995; Pretto et al. 2014; Smeets et al. 1995).

Pretto and colleagues (2014) conducted a pivotal study to assess the impact of mosaicism in 18 individuals bearing a full mutation, undertaking a comparative analysis of molecular metrics in both PBMC and fibroblast cells (Pretto et al. 2014). While there were instances of similarity in CGG allele sizes and their corresponding FMR1 and FMRP expression between PBMC and fibroblast samples, notable discrepancies in CGG size allele distribution were prevalent in the majority of cases. These disparities likely contribute to the variations observed in the phenotypic presentation of Fragile X Syndrome (Pretto et al. 2014).

Given the potential influence of tissue-specific gene expression, it is imperative to consider that the number of repeats may exhibit variations depending on the cell type from which DNA is extracted. In our present study, we utilized DNA extracted from white blood cells, and it is conceivable that the expression patterns of genes specific to neurons may differ slightly. As a result, for future investigations, it is crucial to replicate this repeat expansion study in induced pluripotent stem cells (iPSCs). This approach, as explained by, has presented the ALS field with a transformative opportunity to finally model pathological mechanisms in vitro (Myszczyńska and Ferraiuolo 2016).

The *FRA11A* gene causes mental retardation; its biological region overlaps with the *C11orf80* gene on the q arm of chromosome 11 and contains CGG/CCG trinucleotide repeats. Normal alleles range between 6 and 8 CGG. Expansions up to 500 repeats suggest hypermethylation of the DNA in this region and may be associated with the expression reduction of the *C11orf80* gene product (Debacker et al. 2007) (Table 3-1). The 11q13 locus has also been associated with neuromuscular disorder SCA5, characterised by slowly progressive degeneration of the cerebellum, spinal cord and brain stem (Bártová et al. 2010; Handt, Sutherland, and Richards 2000). Results obtained in the present study suggest a susceptibility to ALS at the intermediate lengths between 10 and 64. ALS risk increases by ten times when the allele carries 23 or more repeats, which is not statistically significant. Results from the Irish population suggest a possible novel cause of ALS.

JPH3 gene causes HDL2 (Huntington's Disease-Like 2) syndrome, a neurodegenerative disease characterised by involuntary movements and abnormalities of voluntary movements, along with psychiatric symptoms (Table 3-1). Affected individuals carry alleles with more than 66 CTG repeats. The results suggest increased ALS risk when harbouring more than 18 repeats. Previous research by Figley and collaborators reported no association between ALS risk and *JPH3* RE (Figley et al. 2014); however, the marginal effects observed in our dataset with a sample size 4.6× larger than Figley *et al.* suggests that the 2014 study was substantially underpowered, and highlights the importance of large-scale population genetic analysis for such rare alleles. Indeed, our intriguing observation of greater levels of cognitive-onset ALS in expanded *JPH3* carriers derives from just two patients who nevertheless improbably ($p = 0.015$) segregate in the expanded

range; this would most likely not have been observed with a smaller sample size. The importance of this observation remains to be seen pending further replication, but it points to overlapping disease mechanisms in cognitive symptoms of ALS and HDL2 for this apparently pleiotropic locus.

SCA31 is a neurodegenerative disease caused by an insertion of a long stretch of TGGAA repeats before a stretch of TAAAA repeats in an intron shared between the *TK2* and *BEAN1* genes (Sato et al. 2009). Healthy controls do not have this TGGAA insertion, while patients have more than 500 repeats with the following conformation: [TCAC-(TGGAA)_{exp}(TAGAA)_{exp}(TAAAA TAGAA)_{exp}] (Ishiguro et al. 2021; Sato et al. 2009). For this study, we assumed that after sequencing some controls and patients, we are not measuring the long stretch of TGGAA repeats, instead the “benign” TAAAA stretch. In order for these results to be validated, a rpPCR for the long stretch should be designed and optimized to corroborate results or sequence every patient and control for this intron. The results obtained in this study suggest that there is a probable increase of ALS risk of 38% when harbouring an intermediate length of 22 or more TAAAA repeats, which was marginally statistically significant.

Possible ALS risk

Similarly, nine genes are categorised as possible ALS risk genes. The results observed for all the genes were not statistically significant; however, even with our large sample size, the power to detect significant associations with such rare alleles may be low for these genes. This group of genes is involved in several diseases, including cerebellar ataxia, dementia, myotonic dystrophy 1, Oculopharyngodistal myopathy 2, neuronal intranuclear inclusion disease (NIID), neuropathy, ataxia, parkinsonism, tremor, Alzheimer’s disease and several SCAs (SCA6, SCA7, SCA12, SCA17). Even though there is no statistically significant finding in the results, the diseases, and phenotypes of several of these neurodegenerative genes are like ALS. These genes must be replicated in a bigger sample size to increase power and properly understand their involvement in the risk of ALS.

Genes not associated with ALS risk

Finally, nine genes were observed not to have any association or increase of ALS risk when harbouring intermediate repeats. These genes have been associated with multiple diseases such as mental retardation, epilepsy, autistic spectrum, ocular and pharyngeal disorders, and several ataxias, including FRA and SCAs (SCA6 and SCA10). Even though it is known that some SCAs have pleiotropic effects, not all may be involved and/or increase ALS risk.

FTD

The results of this study provide novel insights into the potential genetic underpinnings of FTD and ALS, two complex and devastating neurodegenerative diseases. Although the sample size of FTD patients was small, the results suggest a possible association between NDREs genes and increased FTD risk. Notably, the *C9orf72* gene was found to be strongly associated with FTD risk, with intermediate repeat lengths between ≥ 13 and ≥ 16 increasing the risk of disease by eight times. Similarly, the *ATXN1* gene was found to be associated with increased FTD risk by 2 times. The study also identified potential associations between intermediate repeat lengths of other NDRE genes and increased FTD and ALS risk. Further research in larger populations is necessary to confirm these findings. The study's findings on the *ARX*, *AR*, and *CNBP* genes as probable ALS risk genes provide new insights into potential genetic factors that may influence disease onset and progression. Overall, identifying new genes associated with FTD and ALS can uncover new disease mechanisms and potential targets for diagnosis and treatment.

Site of onset, age at onset and disease duration in ALS

The site of onset is a significant factor that influences the course and symptoms of ALS, a devastating neurodegenerative disease. ALS can present in different parts of the body, and the site of onset can have a significant impact on the progression and duration of the disease, as well as on the specific symptoms experienced by individuals with the disease. Some people may experience initial symptoms in the limbs, while others may first notice symptoms in the muscles used for

speech and swallowing. In rarer cases, cognitive or behavioural changes may be the first symptoms. As mentioned earlier, cognitive/behavioural changes can occur in some people with ALS, but this is a less common subphenotype known as ALS-FTD (frontotemporal dementia). In ALS-FTD, symptoms related to changes in behaviour, personality, and/or thinking skills are observed alongside problems with movement. Understanding the site of onset in ALS is crucial in developing effective treatment plans and improving the quality of life for those living with the disease.

The duration of ALS is a critical aspect that can have a significant impact on individuals with the disease and their families. ALS is a progressive disorder that typically worsens over time and eventually leads to respiratory failure. Disease duration is influenced by several factors, including age, site of onset, and rate of progression. While some individuals may survive for several years after onset, most people with ALS survive for only a few years. Understanding disease duration is vital in providing effective care and support for individuals with ALS and their families.

The findings of this project provide compelling evidence for a link between the *JPH3* gene, CAG repeats, cognitive/behavioural site of onset, and ALS survival. Specifically, the study found that individuals with more than 19 and 20 CAG repeats were more likely to experience cognitive/behavioural site of onset in ALS. Interestingly, harbouring more than 19 repeats also increased disease duration by 13.76 months and age at onset by 7.5 years. This complex of increased apparent ALS risk for *JPH3* expansion carriers, along with increased survival and modulated cognitive involvement, highlights the importance of deep sub-phenotyping of patients and the value of exploiting pleiotropy in understanding the interconnectedness of disease pathology between related phenotypes.

Similarly, the findings of our study provide new insights into the role of the *ATXN3* gene in the pathogenesis of ALS, with results showing a slight increase in cognitive/behavioural onset in ALS patients with more than 29 CAG repeats in *ATXN3* (although the result was not statistically significant). Additionally, we observed a decrease in disease duration of 3.65 months, also not statistically significant. These findings are consistent with the role of *ATXN3* in regulating the ubiquitin-proteasome system and its potential involvement in cognitive function (McLoughlin et

al. 2020). Our results suggest that intermediate or expanded lengths of the *ATXN3* gene may be a possible genetic risk factor for cognitive/behavioural site of onset in ALS, which may have important implications for the development of therapeutic interventions.

The role of the *AFF2* gene in neurodegenerative diseases has been the subject of several studies. In our investigation, we found a significant association between intermediate CGG repeats in the *AFF2* gene and an increased risk of cognitive/behavioural site of onset in ALS patients. Our results showed that individuals with more than 21, 22, 23, and 24 CGG repeats were at higher risk of experiencing cognitive/behavioural symptoms as the site of onset in ALS. Additionally, when harbouring more than 24 repeats there is a possible decrease of 3 years in age at onset. This finding is consistent with the known function of the *AFF2* gene in transcriptional regulation, brain development and function, and its role in causing FRAXE, a form of mental retardation (Jorge et al. 2013; Yuva-Aydemir et al. 2019). These results highlight the potential role of the *AFF2* gene as a genetic risk factor for cognitive/behavioural site of onset in ALS and suggest that further research is needed to fully understand the implications of these findings and their potential for developing targeted interventions for patients with this form of ALS onset.

Recent studies have identified the *ATXN2* gene as a potential genetic risk factor for respiratory site of onset in ALS (Wang et al. 2017). Specifically, our study found that individuals with an intermediate repeat length of more than 32 in the *ATXN2* gene suggests increased risk of respiratory site of onset in ALS. Additionally, we observed an increase in disease duration of 5.34 months when harbouring more than 32 CAG repeats, and an increase of 1 year at age of onset when harbouring more than 28,29,30 repeats. This is consistent with previous research suggesting that the *ATXN2* gene plays a role in the regulation of RNA processing and cellular stress response, which may ultimately impact respiratory function (Ross et al. 2011; Wang et al. 2017). Further investigation is necessary to fully understand the role of *ATXN2* in respiratory site of onset in ALS, and to determine whether targeted interventions aimed at modulating *ATXN2* expression or activity could improve respiratory function and disease outcomes for ALS patients.

Recent studies have implicated the *ATXN7* gene in the pathogenesis of several neurodegenerative diseases, including SCA7 and a case of ALS-like phenotype (Cluse et al. 2021). In our study, we

found a significant association between intermediate CAG repeats in the *ATXN7* gene and an increased risk of respiratory onset in ALS patients. Specifically, individuals with more than 14 and 15 CAG repeats were more likely to experience respiratory symptoms as the site of onset in ALS, which was statistically significant. Although our results showed a non-significant decrease in disease duration of 31.9 months and a decrease of 4-4.5 years in age of onset in patients with more than 14 and 15 repeats, further investigation and replication is needed to validate these findings. Our results are consistent with previous reports of respiratory dysfunction in models of SCA7 (Fusco et al. 2021) and the known role of the *ATXN7* gene in transcriptional regulation and its function in regulating the activity of other genes. The respiratory system is critical for human physiology, and its regulation involves a complex network of genes, including those involved in ion channels, neurotransmitter release, and synaptic transmission (Cluse et al. 2021; Harding 1982; Niewiadomska-Cimicka and Trottier 2019).

The *DIP2B* gene has been implicated in several neurological disorders, including ALS. Our findings suggest that individuals with intermediate repeat lengths in the *DIP2B* gene may be at an increased risk for respiratory site of onset in ALS. Specifically, we found that individuals with more than 25 GGC repeats in the *DIP2B* gene were more likely to experience respiratory symptoms as the site of onset in ALS and an increase in age of onset of 0.1 years; both results were statistically significant. The *DIP2B* gene is known to play a role in the development and function of neurons, and it has been suggested that it may also be involved in the regulation of muscle function (Annear et al. 2021; Castelli et al. 2021). Our results support the hypothesis that the *DIP2B* gene may be involved in respiratory dysfunction in ALS, and further research is needed to understand the molecular mechanisms by which this gene may contribute to the disease.

The *FRA11A* gene, located in the Fragile Site A region of chromosome 11 (*FRA11A*), has been associated with mental retardation. Our study investigated the role of *FRA11A* in respiratory onset of ALS and found a significant association between intermediate CGG repeats in the *FRA11A* gene and increased risk of respiratory onset. Specifically, individuals with more than 11, 12, and 13 CGG repeats in *FRA11A* were more likely to experience respiratory onset in ALS. Furthermore, we observed a statistically significant increase in survival of ALS patients by 9.47 months when harbouring more than 11 CGG repeats and 11.9 months with more than 12 repeats. Both results

are statistically significant. Additionally, harbouring more than 11 repeats also increases age of onset in ALS patients by 2 years, although this was not statistically significant given the sample size. Expanded CGG repeats in *FRA11A* may cause DNA hypermethylation in the *C9orf80* gene, reducing its expression (Debacker et al. 2007). Our study sheds light on the potential role of *FRA11A* as a genetic risk factor for respiratory onset in ALS and may have important implications for the development of targeted interventions for patients with this specific clinical phenotype.

The length of trinucleotide repeats in various genes has been implicated in the pathogenesis of ALS, including *C9orf72*. Our study found that individuals with more than 16 repeats in the *C9orf72* gene had a decrease in disease duration of 8.87 months, which was statistically significant. Additionally, harbouring more than 15 and 16 repeats may reduce age of onset in ALS patients by 5 years. The *C9orf72* gene is one of the most commonly mutated genes in ALS and is associated with FTD. Our results suggest that the length of the trinucleotide repeats in the *C9orf72* gene may play a role in determining disease duration and age of onset in ALS patients.

Similarly, our findings showed that the length of trinucleotide repeats in the *ATNI* gene was associated with disease duration in ALS patients. Individuals with more than 11, 12, or 13 repeats in the *ATNI* gene had a statistically significant increase in survival by 31.9 months. This finding is consistent with the known function of the *ATNI* gene in transcriptional regulation and its role in neuronal development and function (Dewan et al. 2021; Malik et al. 2021).

In contrast, we found that individuals with more than 11 CGC repeats in the *ARX* gene had a statistically significant reduction in disease duration by 10.32 months and an increase in age at onset by 2 years. The *ARX* gene plays a crucial role in the development of the cerebral cortex and has been associated with various neurodevelopmental disorders (Friocourt and Parnavelas 2010; Marques et al. 2015). This finding suggests that *ARX* may contribute to ALS risk by causing alterations in RNA and protein structure, function, and interaction with other molecules in the cell and DNA. Further research is needed to determine the underlying mechanism and the potential implications of this novel gene on ALS pathogenesis. Our study highlights the importance of exploring less well-known genes in the search for novel genetic factors that contribute to ALS risk.

Our research also suggests that the length of trinucleotide repeats in the *PABPN1* gene may be a novel genetic risk factor for ALS. Individuals with more than 7 repeats in the *PABPN1* gene had a decrease in survival by 4.96 months (borderline significant, p -value: 0.071), and decreases age of onset by 3 years (not statistically significant). The *PABPN1* gene is involved in poly(A) tail binding and mRNA stabilization and has been implicated in various neuromuscular disorders. Mutations in the *PABPN1* gene have been linked to various neuromuscular disorders, including OPMD and limb-girdle muscular dystrophy type 2B (LGMD2B). In the case of OPMD, the mutant *PABPN1* protein aggregates in muscle cells, leading to cell death and muscle wasting (Richard et al. 2017). Our study suggests a potential role for *PABPN1* in ALS. One study found a rare genetic variant in the *PABPN1* gene, which could lead to increase of the aberrant *PABPN1* protein with a higher risk of developing ALS.

3.5 Summary

The discovery of more than 50 disease-causing RE that have been identified in association with highly penetrant clinical syndromes has opened a new frontier for research into non-Mendelian human disorders. Despite the long-held view of STRs as "junk" DNA, evidence has shown that these repeats may have functional consequences that could impact the structure and function of proteins, RNAs, and DNA. These effects can have significant molecular and cellular consequences, and thus, STRs variations could serve as risk alleles for human disorders. Furthermore, evolutionary genomics suggests that STRs have evolved under tight selective conditions, indicating that they may have important biological roles.

In the context of neurodegenerative disorders, many intermediate STRs have been linked to the risk of several conditions, including ALS, SCA2, SCA1, and ALS-FTD. For instance, in ALS, the CAG repeats in the *ATXN2* gene are a risk factor, and the GGGGCC repeats in *C9orf72* gene are linked to ALS-FTD. Pleiotropy, or the phenomenon where a genetic locus affects more than one trait, has been demonstrated in many disorders, including ALS. This phenomenon offers an excellent opportunity to investigate related phenotypes, such as respiratory site of onset and disease duration, to uncover additional genetic associations. The identification of these associations has significant implications for understanding the underlying mechanisms of disease and developing targeted interventions. The potential for pleiotropy and cross-phenotype associations should be further explored to identify novel genetic risk factors for neurodegenerative conditions.

Recent data suggests that STRs have a functional role in the genes where they are located and that variations in these repeats may contribute to human disease. Specifically, many identified STR disorders impact the nervous system, indicating that STRs play a specific role in these specialized cells. Due to their faster mutation rate compared to single nucleotides, STRs can act as tunable regulators of functions and gene expression, enabling them to hasten evolutionary adaptation. In this study, we sought to identify a broader role of known NDREs in ALS beyond the previously recognized risk alleles (*ATXN1*, *ATXN2*, *C9orf72*, *NIPAI*, and potentially *HTT*). Our findings revealed an overall increase in ALS risk associated with intermediate-length REs in nine probable

risk genes and three novel ALS-associated genes: *ARX*, *FRA10AC1*, and *PABNI*. Notably, the *PABNI* protein plays a critical role in forming nuclear condensates of FUS and TDP-43, while the *FRA10AC1* protein product appears to increase A β 1–42, a major component of amyloid plaques in Alzheimer's disease. These pleiotropic loci may point to shared biological mechanisms that can improve our understanding of ALS disease mechanisms, diagnosis, and prediction accuracy.

Expanded alleles of known ALS genes cause a variety of dysfunctions in the cell (Table 3-1). Theoretically, the concurrent gain-of-function roles of these genetic mechanisms could be mediated by the sequestration of RNA-binding proteins in stress granules that fail to disaggregate in senescent cells. Intermediate-length alleles suggest that these mutations are generally neurodegenerative, perhaps by failing to separate RNA-binding proteins generating stress granules in the senescent cells. This may be a promising future target for drug development. Moreover, it is possible that the combination of multiple expressed transcripts produces a polygenic effect that could be sufficient to cause neurodegeneration, ultimately dictating the specific disease. This is the subject of the next chapter, in which we investigate the overall burden of the repeat lengths of all NDRE genes with ALS risk.

In conclusion, our study highlights the potential role of various genes, including *C9orf72*, *FRA10AC1*, *ARX*, and *PABPN1*, in determining disease risk, duration and site of onset in ALS patients. Further research is needed to fully understand the implications of these findings and their potential for developing targeted interventions for ALS patients with different clinical phenotypes.

Chapter 4

Burden of NDRE genes in the increase of ALS risk in the Irish population

4.1 Introduction

Neurodegenerative diseases, such as Dementia, affects about 55 million people globally in 2019, particularly the elderly, and is estimated to increase to 139 million. Although these diseases are considered separate entities, there is evidence of overlap in their symptoms, pathology, and genetics (Health Organization 2021). Researchers have been investigating the shared genetic basis of these diseases, with large-scale genome-wide association studies identifying multiple genetic loci contributing to increased disease risk (Chen et al. 2020; Ferrari et al. 2011). Many of these genetic loci have been found to exert significant influences on multiple neurodegenerative diseases, which has important implications for clinical diagnosis and treatment, as well as for statistical analysis and prediction accuracy. While some common genetic components underlying neurodegenerative diseases have been identified, the extent of shared genetic basis remains unclear, and there is still much to learn about shared genetic loci beyond those already described (Sivakumaran et al. 2011; Solovieff et al. 2013; Watanabe et al. 2019).

In ALS, different genetic variants can have various effects on clinical symptoms. *SOD1* variants primarily cause motor degeneration, while *FUS* mutations are associated with earlier age of onset and some variants affect disease progression rate (Chia, Chiò, and Traynor 2018). *C9orf72* repeat expansions are strongly linked to cognitive and behavioural changes, and *FUS* and *TARDBP*

mutations can also lead to dementia. However, many patients with cognitive changes do not have any known genetic variants. Additionally, some mutations that increase the risk of ALS are pleiotropic, and non-motor symptoms of ALS overlap with those of other neurodegenerative diseases (Byrne et al. 2012; Cali et al. 2019; O'Brien et al. 2017). For instance, *C9orf72* repeat expansions are common in Huntington's disease phenocopies, and patients with frontotemporal dementia or ALS may have concurrent *HTT* repeat expansions without the typical features of HD (Hensman Moss et al. 2014). Furthermore, mutations in ALS risk genes have been found in patients with dementia with Lewy bodies and primary progressive multiple sclerosis, suggesting that the relationship between genotype and phenotype among genetic mutations that cause neurodegenerative disease is highly complex (Dewan et al. 2021).

There is evidence of a connection between neuropsychiatric disease and ALS among family members, particularly in regards to psychosis and suicide (Devenney et al. 2018; McHutchison et al. 2020). While *C9orf72* repeat expansions play a role in some cases, they are not more common in individuals with typical schizophrenia (Devenney et al. 2021). Family studies indicate that neuropsychiatric conditions are not evenly distributed, and instead cluster in up to 30% of kindreds of patients with ALS (O'Brien et al. 2017), suggesting that there may be a genetic pleiotropy or oligogenic inheritance at play. Additionally, there is evidence of shared polygenic risk factors between ALS and neuropsychiatric disease, with *GLT8DI* identified as a risk gene for both conditions (McLaughlin et al. 2017; Yang et al. 2018). These findings suggest that some forms of ALS may disrupt specific brain network patterns and be mediated by developmental processes that render certain brain networks more susceptible, leading to neuropsychiatric symptoms or later-onset neurodegeneration in family members. Further research is necessary to understand the complex relationship between ALS and neuropsychiatric disease, and to identify the genetic factors underlying these conditions (Goutman et al. 2022; McMackin et al. 2021).

In recent years, the genomic investigation of large patient cohorts via genome-wide association studies has revealed a substantial polygenic component to ALS risk (McLaughlin et al. 2017; van Rheenen et al. 2021), representing the additive contribution of risk alleles at common SNPs to ALS phenotypes. Although repeat expansions are now recognised as also playing a major role in contributing to ALS risk, their combined effects on determining disease liability has been

relatively overlooked in comparison to SNP-based polygenic effects, despite the substantially greater effect sizes that risk alleles of STRs confer compared to those of SNPs. Oligogenic architectures have been associated with *C9orf72*-mediated ALS and other rare mutations (Cady et al. 2015; Giannoccaro et al. 2017; McCann et al. 2021) but no study has comprehensively analysed repeat expansions en masse in a polygenic framework to explore the overall burden of STR risk alleles to ALS. Such an analysis is appealing, as it is possible that repeat expansions on different transcripts exert their molecular pathogenic effects via similar or identical mechanisms; therefore, increasing the number of risk alleles across transcripts in a patient is likely to increase their overall risk for ALS. This chapter explores this hypothesis using NDRE panel data generated in chapter 3.

Research aims

The aim of this study is to use PCR genotyping as the gold standard to determine the per-individual burden of expanded repeats in NDRE genes in the ALS Irish population. We investigate the potential for gene-gene interactions, and analyse the relationship between the burden of number repeats and developing ALS, age of onset and disease duration in the Irish population.

4.2 Methodology

PCR Genotyping

This chapter uses NDRE panel genotyping data for 1550 ALS patients and 1719 unrelated control subjects, generated in Chapter 3 using the methodology described in Chapter 2. *FMR1* genotypes were excluded from this analysis.

C9orf72 expanded genotypes

In addition to NDRE genotypes, individuals with ALS were examined to check for the existence of the long disease-causing RE using repeat-primed PCR (rpPCR, explained earlier in Chapter 2) as part of ongoing surveillance of incident Irish ALS patients. The amplified fragments were evaluated using capillary electrophoresis on an Applied Biosystems 3500 Series Genetic Analyzer and viewed with Gene Mapper v.4.0. The screening was done to detect a diminishing sawtooth pattern, which signifies the presence of a large RE. Patients having 30 hexanucleotide repeats or more and demonstrating a sawtooth rpPCR trace were determined as positive for the expansion (“expanded”, “C9 positive”). Controls were not tested for long *C9orf72* expansions using rpPCR.

Data processing, statistical analysis, and plotting

All analyses described in this chapter were performed with RStudio (version 2022.12.0+353).

To ensure an unbiased treatment of the data, not influenced by prior knowledge of the repeat expansions and their pathogenic ranges in other diseases, we applied a uniform criterion to define alleles as expanded. For each gene, an expanded/non-expanded threshold was defined from the observed distribution of the data by considering only the top percentile of data as expanded. This resulted in thresholds defined in Table 4-1. For *C9orf72*, individuals in the top percentile from the NDRE panel genotypes were combined with individuals defined as C9-positive by rpPCR. Where the 99th percentile repeat length was equal to the modal repeat value in our data for a given gene, the threshold was adjusted to be the mode plus one to avoid classifying most of the dataset as expanded for that gene.

Per individual, burden was defined as the count of expanded alleles across all NDRE genes per patient or control subject. Patient vs control burden was assessed using the chi-squared test. Analyses were also performed excluding *C9orf72*, *ATXN2* and *NIPA1* to ensure these established ALS REs were not driving the signal in our data.

Case/control burden was visualised using mosaic plots. This is a graphical representation of data that display the proportional distribution of two or more categorical variables in a single plot. In a mosaic plot, each category of one variable (in our case, burden, ie the number of repeat expansions carried) is represented by a rectangle, and these rectangles are then divided into smaller rectangles or segments according to the proportion of the other variable within each category (in our case, case/control proportions). The size of the rectangles represents the relative frequency of each category. Mosaic plots are useful for comparing the distribution of different categories across multiple variables, and can help to identify relationships and patterns between them.

Pairwise co-occurrence of expanded genes was visualised using heatmaps. To achieve this, we enumerated the co-occurrence of all pairs of genes (561 possible pairs from 34 genes) in cases and controls and calculated the proportional difference in case/control counts for each pair. We clustered the resulting matrix using the default hierarchical clustering method in the `heatmap` function in R. This analysis was conducted with and without *C9orf72*.

Finally, we also conducted an analysis to investigate the effect of the burden of repeats on the age of onset (in years) and duration of the disease (in months) in patients with ALS. To achieve this, we designed a linear regression model that compared the number of long repeats per individual with the age of onset or duration of the disease. This allowed us to examine the relationship between the burden of repeats and disease duration or age at onset, which can provide valuable insights into the underlying mechanisms of the disease.

Table 4-1 Threshold table

Gene	Common Repeats	99th percentile	Mode
<i>AFF2</i>	17	29	14
<i>AFF3</i>	8	21	7
<i>AR</i>	22	30	21
<i>ARX</i>	10	11	10
<i>ATN1</i>	15	22	15
<i>ATXN1</i>	29	36	30
<i>ATXN10</i>	14	21	14
<i>ATXN2</i>	22	31	22
<i>ATXN3</i>	14	35	23
<i>ATXN7</i>	10	14	10
<i>ATXN8</i>	25	24	14
<i>CACNA1A</i>	13	15	13
<i>CNBP</i>	20	21	17
<i>CSTB</i>	3	3	3
<i>C9orf72</i>	3	15	2
<i>DIP2B</i>	7	24	7
<i>DMPK</i>	20	30	13
<i>FRA10AC1</i>	8	26	8
<i>FRA11A</i>	8	12	9
<i>FXN</i>	6	29	10
<i>GIPC1</i>	10	16	10
<i>HTT</i>	21	33	21
<i>JPH3</i>	14	18	16
<i>LRP12</i>	9	16	13
<i>NIPA1</i>	8	10	8
<i>NOP56</i>	7	10	6
<i>NOTCH2NLC</i>	9	24	16
<i>NUTM2B-AS1</i>	6	14	9
<i>PABPN1</i>	10	7	6
<i>PPP2R2B</i>	10	18	10
<i>TBP</i>	38	40	38
<i>TCF4</i>	24	74.46	17
<i>TK2-BEANI</i>	21	36.49	15
<i>ZNF713</i>	12	20	11

4.3 Results

Burden of NDRE genes in ALS risk

The study aimed to investigate the effect of the burden of expanded repeat alleles on ALS by genotyping data of 34 NDREs using the 99-percentile method. To test the hypothesis that individuals with ALS carry more NDREs than controls, mosaic plots of the proportions of ALS cases and controls were designed. The results in Figure 4-1 revealed that the burden of expanded repeat alleles is significantly higher in ALS cases compared to controls. The proportion of ALS cases increased with the number of expansions, and the risk of developing ALS was found to be associated with the number of expanded alleles. Specifically, the proportions of ALS cases showed a significant increase to 100% when having 5 repeat expansions (p -value: 1.07×10^{-9}).

The study also aimed to investigate the effect of excluding known ALS risk genes (*C9orf72*, *ATXN2*, and *NIPAI*) on the burden of expanded repeat alleles in ALS, to explore the extent to which our results were driven by major known ALS-associated repeat expansions. Results showed that excluding *C9orf72* repeat expansions from the model led to an increase in the proportion of ALS cases when the number of repeats increased up to 4; this was statistically significant (p -value = 3.6×10^{-3} , Figure 4-2). After excluding both *C9orf72* and *ATXN2* genotypes, the burden of expanded repeat alleles in ALS cases remained similar to the previous model, and the risk of developing ALS was found to increase up to 4 repeat expansions, with a reduced but still significant p -value of 0.0226 (Figure 4-3). Finally, excluding *NIPAI* (plus *C9orf72* and *ATXN2*) repeat expansions from the model showed a similar trend as the first model, where ALS cases carrying a maximum number of 4 repeats were seen in a higher proportion than controls. However, the p -value of the test increased to 0.07384 (Figure 4-4).

Overall, the results suggest that there is an increased burden of expanded repeat alleles in ALS cases compared to controls, with up to 5 different RE genes from the NDRE panel being associated with a higher risk of developing ALS. Even after excluding known ALS risk genes, the model suggests that the burden of repeat expansions in the remaining genes tested is associated with the development of ALS. These findings could have significant implications in understanding the

pathogenesis of ALS and identifying potential biomarkers and unifying pathogenic mechanisms for the disease.

Figures 4.5-4.8 show alternative analyses through an oligo-/polygenic lens. Here, we categorised patients and controls using a thresholding approach, based on whether individuals carried two or more expansions. To achieve this, we combined the 0 and 1 expansions and 2-5 expansions categories. Our results showed that ALS cases and controls have similar proportions of individuals with 0 or 1 expansions, but the proportion of ALS cases significantly increases up to 60% when individuals harbour 2 or more expansions (p -value 2.16×10^{-09} , Figure 4-5). Excluding known ALS risk genes such as *C9orf72*, *ATXN2*, and *NIPA1*, we found that the burden of repeat expansions in ALS cases is higher than controls when individuals harbour 2 or more intermediate repeat genes (Figure 4-6, Figure 4-7, Figure 4-8).

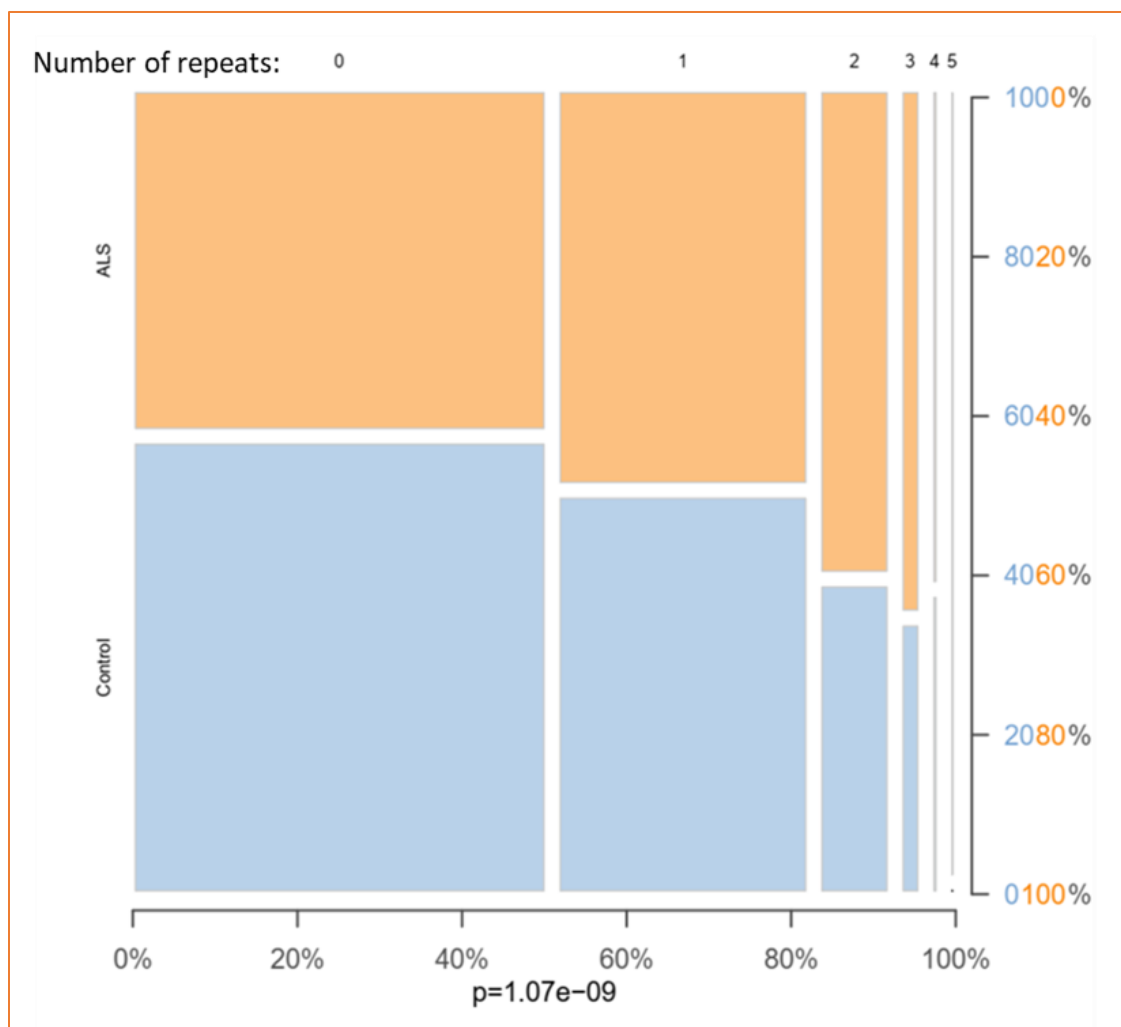


Figure 4-1 Burden of all NDREs in ALS risk using 99th-percentile model.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0, 1, 2, 3, 4 or 5 long expansions in all 34 NDRE genes. As the number of expansions increases, the proportion of ALS cases increases (p -value= 1.07×10^{-09}). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

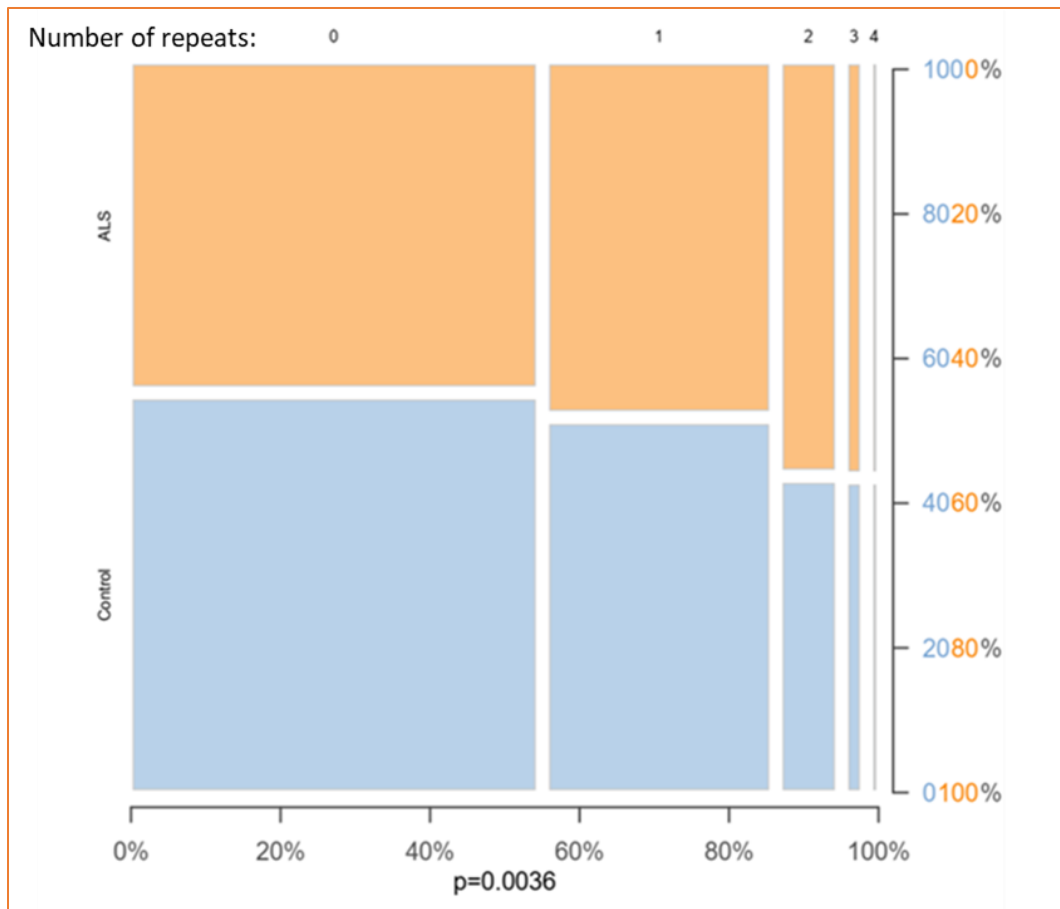


Figure 4-2 Burden of all NDREs in ALS risk using 99th-percentile model, excluding *C9orf72* gene.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0, 1, 2, 3, or 4 long expansions in all NDRE genes excluding *C9orf72*. As the number of expansions increases, the proportion of ALS cases increases ($p\text{-value}=3.6 \times 10^{-03}$). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

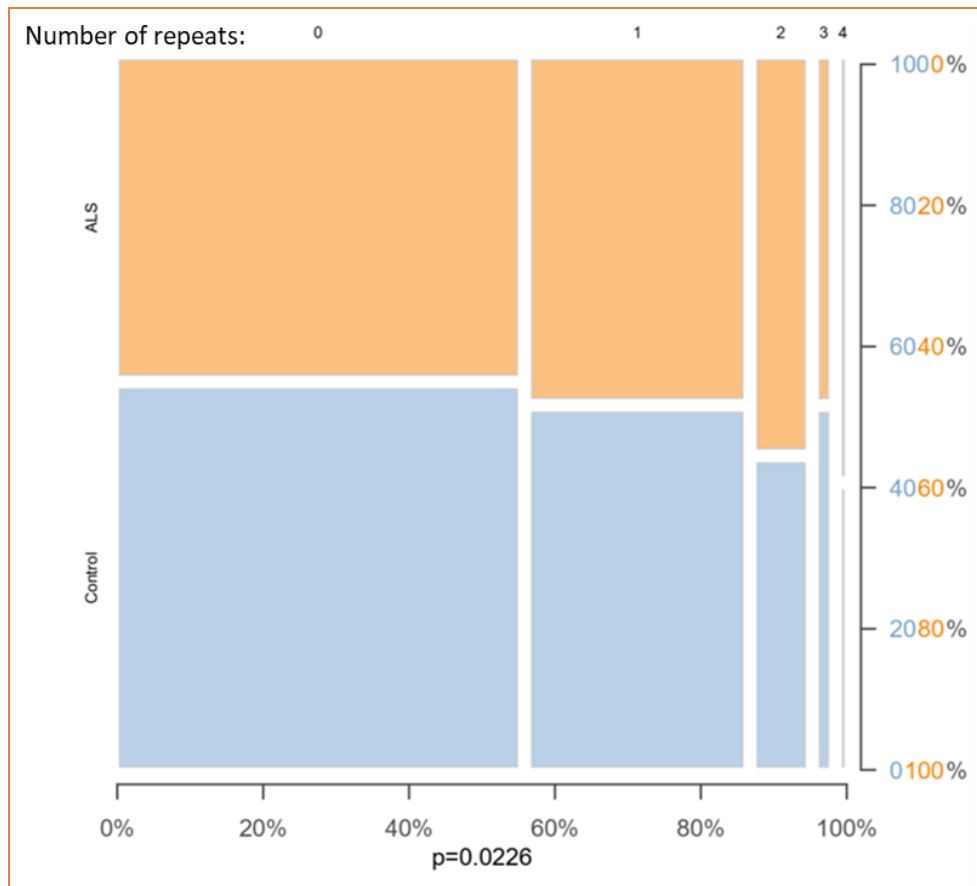


Figure 4-3 Burden of all NDREs in ALS risk using 99th-percentile model, excluding *C9orf72* and *ATXN2* genes.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0, 1, 2, 3, or 4 long expansions in all NDRE genes excluding *C9orf72* and *ATXN2*. As the number of expansions increases, the proportion of ALS cases increases (p -value=0.0226). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

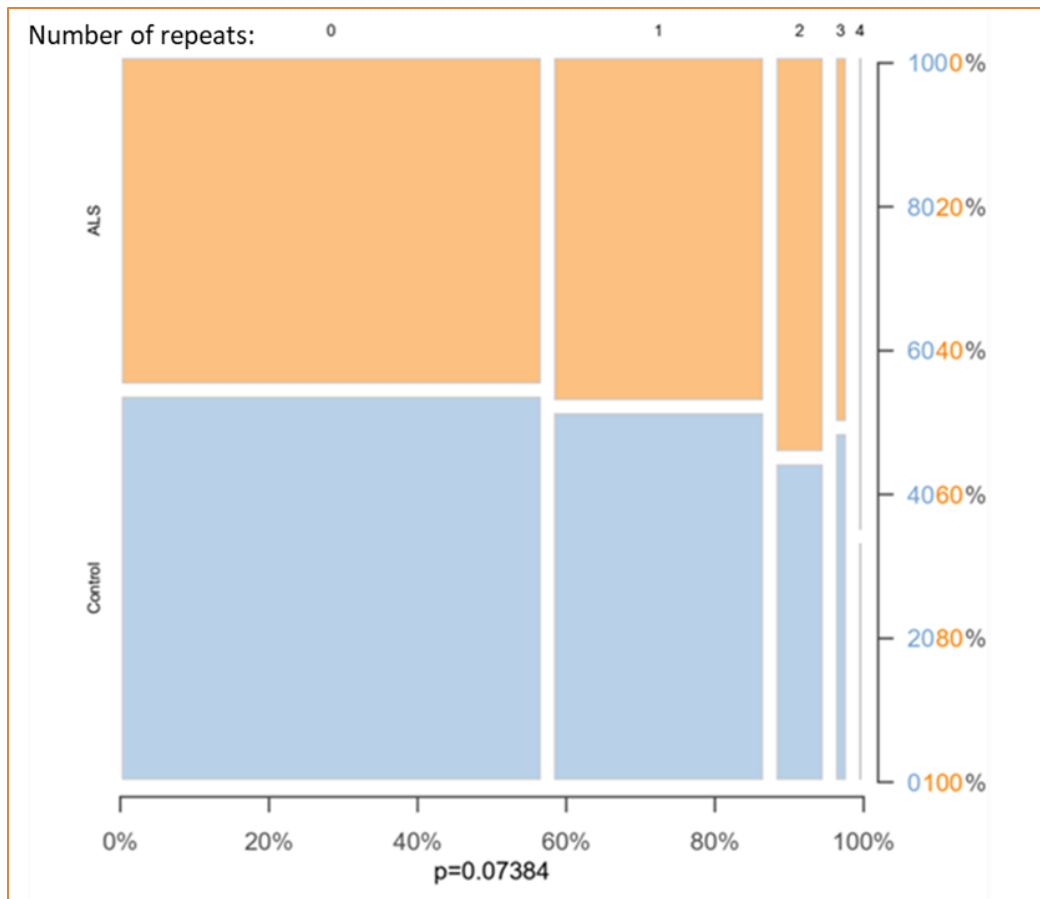


Figure 4-4 Figure 4.3 Burden of all NDREs in ALS risk using 99th-percentile model, excluding *C9orf72*, *ATXN2* and *NIPA1* genes.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0, 1, 2, 3, or 4 long expansions in all NDRE genes excluding *C9orf72*, *ATXN2* and *NIPA1*. As the number of expansions increases, the proportion of ALS cases increases, albeit less significantly than when these major genes are included in the analysis (p -value=0.07384). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

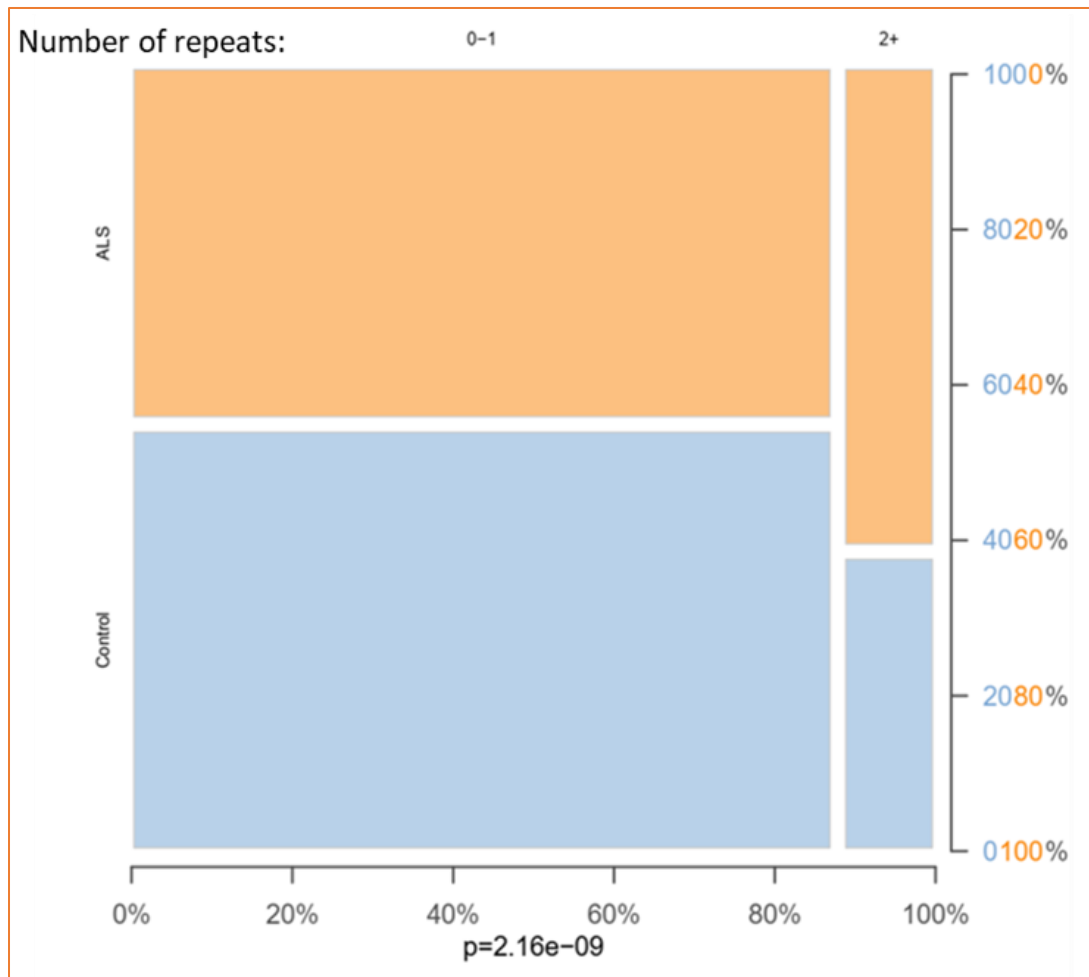


Figure 4-5 Burden of all NDREs in ALS risk using 99th-percentile model combining 0-1 and 2+.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0 or 1, and 2 or more long expansions in all 34 NDRE genes. As the number of expansions increases, the proportion of ALS cases increases (p -value= 2.16×10^{-9}). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

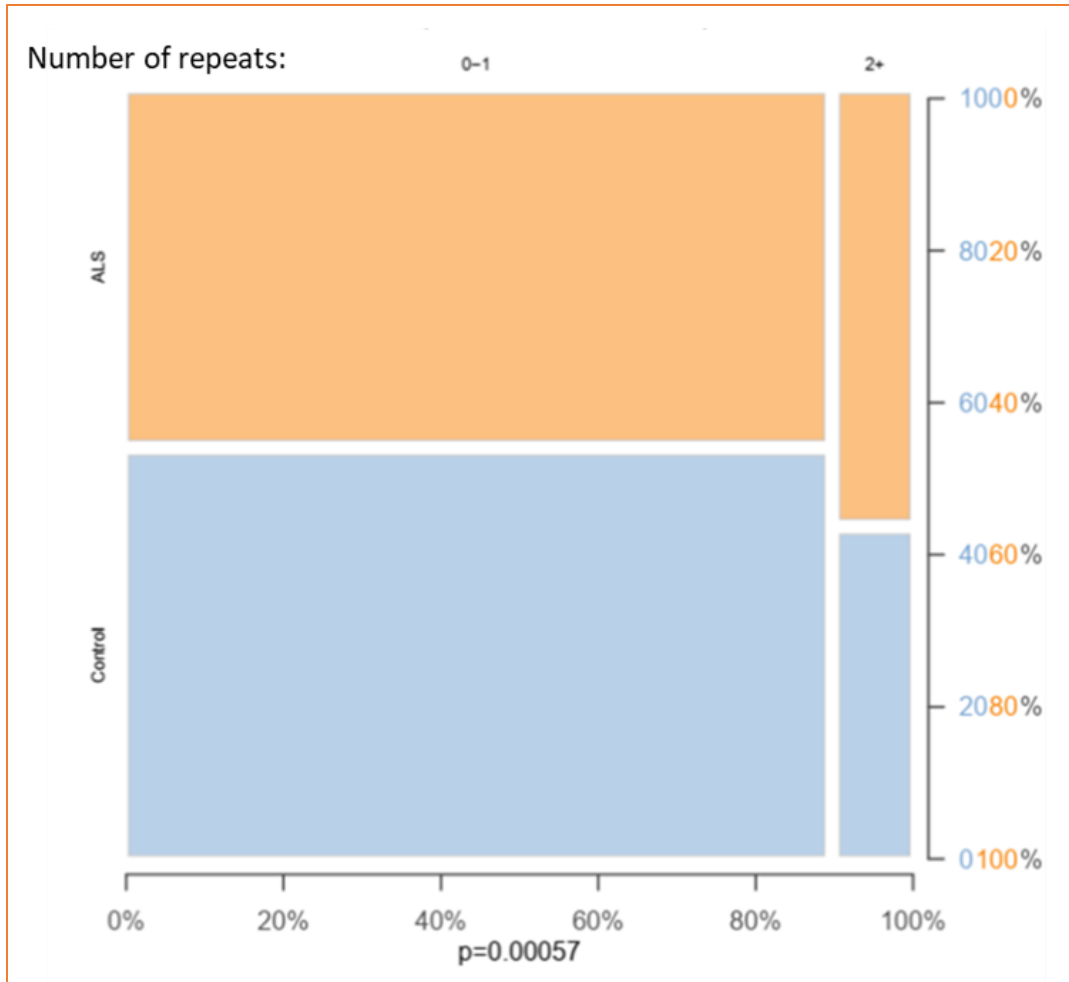


Figure 4-6 Burden of all NDREs in ALS risk using 99th-percentile model combining 0-1 and 2+, excluding *C9orf72* gene.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0 or 1, and 2 or more long expansions in all NDRE genes excluding *C9orf72*. As the number of expansions increases, the proportion of ALS cases increases ($p\text{-value}=5.7 \times 10^{-04}$). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

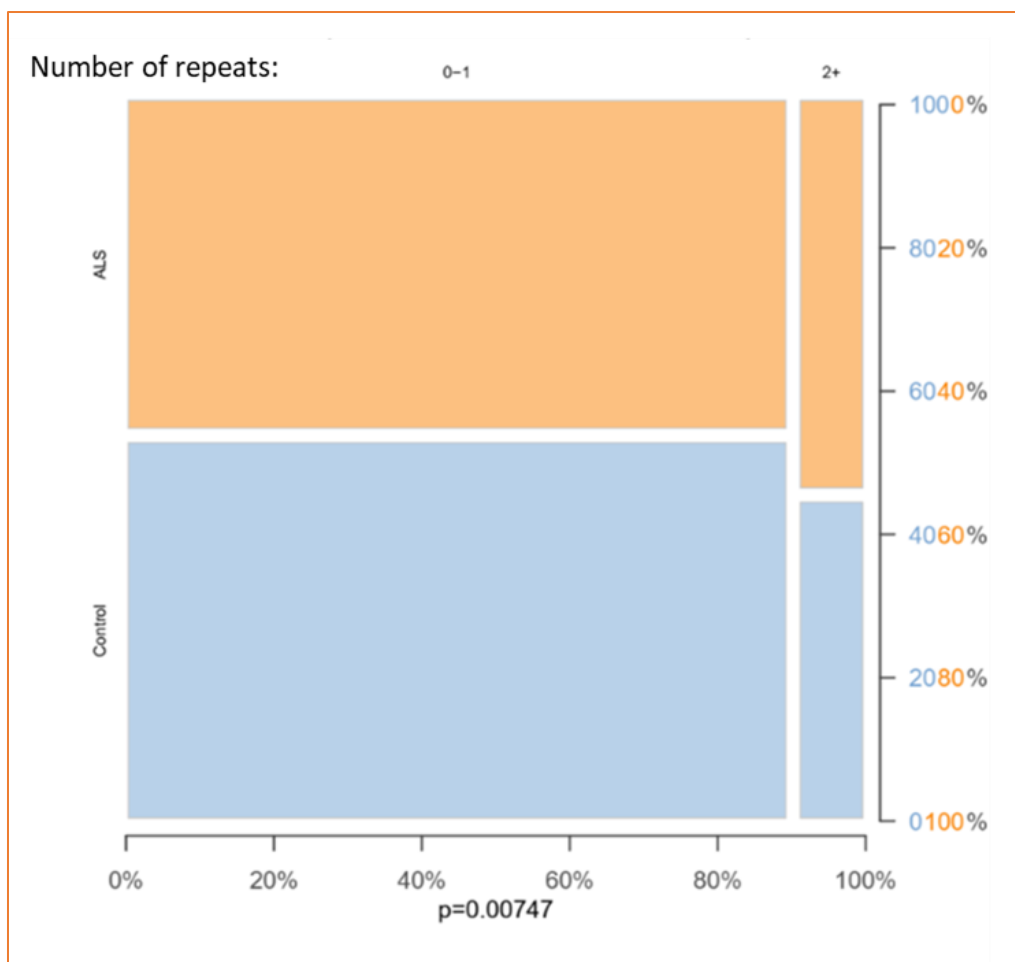


Figure 4-7 Burden of all NDREs in ALS risk using 99th-percentile model combining 0-1 and 2+, excluding *C9orf72* and *ATXN2* genes.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0 or 1, and 2 or more long expansions in all NDRE genes excluding *C9orf72* and *ATXN2*. As the number of expansions increases, the proportion of ALS cases increases ($p\text{-value} = 7.47 \times 10^{-3}$). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

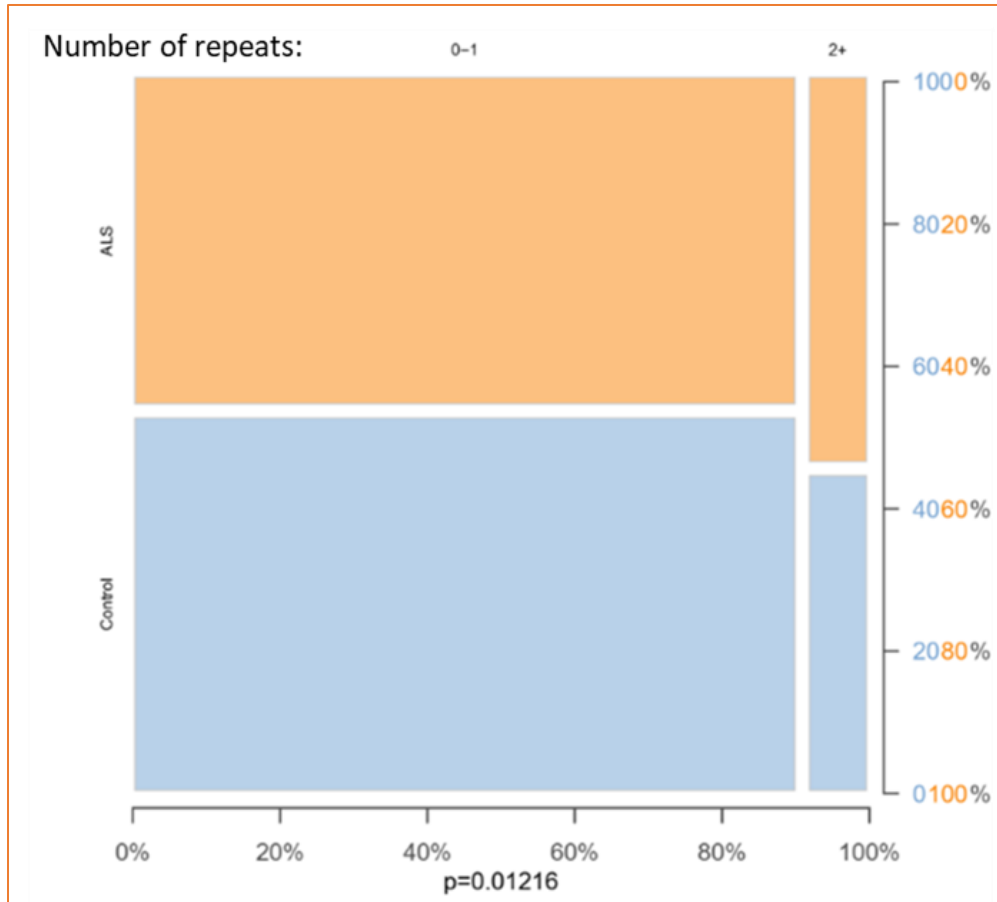


Figure 4-8 Burden of all NDREs in ALS risk using 99th-percentile model combining 0-1 and 2+, excluding *C9orf72*, *ATXN2* and *NIPA1* genes.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0 or 1, and 2 or more long expansions in all NDRE genes excluding *C9orf72*, *ATXN2* and *NIPA1* genes. As the number of expansions increases, the proportion of ALS cases increases ($p\text{-value} = 7.47 \times 10^{-3}$). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

Gene co-occurrence

The results on Figure 4-9 and Figure 4-10 show the relative co-occurrence of expansions (expressed as a proportional difference between cases and controls) in all 34 included NDRE genes (Figure 4-9) and all genes with *C9orf72* excluded (Figure 4-10). Results have been hierarchically clustered using the default method in the `heatmap` function in R to reveal clusters of mutually co-occurring expansions that are more frequent in cases than controls, including *PABPN1*, *CNBP*, *JPH3*, *ATXN2*, *FRA10AC1*, *PPP2R2B*, *NUTM2B-AS1*, *TBP*, *ATXN1*, *GIPC1*, and *ARX*. There are also other pairs of expansions that co-occur more frequently in ALS than controls but do not necessarily group together in this cluster; these are: *NIPAI* and *ATN1*, *NOP56* and *PPP2R2B*, *JPH3* and *CACNA11A*, *ATXN3* and *ATXN2*.

C9orf72 drives the majority of the signal in Figure 4-9, in which most genes have a more frequently co-occurring expansion in ALS than controls. This is expected, given our inclusion of rpPCR results which represents almost 10% of ALS cases. Therefore, excluding it from the clustering in Figure 4-10 reveals less stark co-occurrences between the remaining NDRE genes. Figure 4-10 shows the resulting heatmap with *C9orf72* excluded. A new interaction cluster emerges with more strength than before, involving *PABPN1*, *CNBP*, *JPH3*, *ATXN2*, *GIPC1*, *DIP2B*, *AR*, *AFF2*, *ATXN3*, *LRP12*, *ARX*, *ATXN1* and *CACNA1A*. This suggests that expansions in these genes, when co-occurring with each other, contributes to ALS risk, even when *C9orf72* is not present. We also see a new cluster in the right lower corner (or left upper corner), involving interactions between *TBP*, *NUTM2B-AS1*, *PPP2R2B*, *NIPAI*, *PABPN1*, *CNBP*, *JPH3* and *ATXN2*, previously scattered in Figure 4-9.

Burden of 11 NDREs genes and ALS risk

Following these results, we investigated the individual-level burden of repeat expansions in genes belonging to the previously mentioned cluster of 11 genes in Figure 4-9 (*C9orf72*, *ATXN1*, *ATXN2*, *JPH3*, *CNBP*, *PABPN1*, *PPP2R2B*, *NUTM2B-AS1*, *TBP*, *GIPC1*, and *ARX*) and its association with ALS risk. Figure 4.11 shows that carrying 1 expansion, proportions are almost 50%, with 2 expansions proportion of individuals with ALS increases up to almost 80%, and 100% of individuals with 3 expansions have ALS. As the number of expansion increases, the proportion of

ALS cases increases significantly ($p\text{-value}= 1.82 \times 10^{-17}$), indicating that carrying multiple intermediate expansions increases the risk of ALS. (Figure 4-11).

In Table 4-2, it is evident that among the 11 ALS patients harbouring three expanded genes, there exists a notable absence of identical trios of genes within our dataset. Interestingly, we observed the recurrence of six pairs of genes across the individuals. Specifically, these gene pairs comprise *C9orf72-PABPN1*, *C9orf72-TBP*, *C9orf72-GIPC1*, *ATXN2-CNBP*, *ATXN2-NUTM2B-AS1*, and *PPP2R2B-CNBP*. This observation opens up avenues for further investigation into the potential synergistic or interactive effects of these gene pairs in the context of ALS pathogenesis. Such findings may contribute to a deeper understanding of the genetic underpinnings of ALS and could potentially inform future research endeavours aimed at elucidating the complex interplay of genes in this debilitating neurodegenerative disease.

Overall, our study provides evidence that the interaction between multiple NDRE genes is associated with an increased risk of developing ALS. These findings suggest that the burden of expanded repeat alleles is a critical factor in the development of ALS and could serve as a potential biomarker for the disease.

Age of onset and disease duration

Our findings suggest that a higher burden of repeat expansions is associated with an earlier age of onset (not statistically significant) and shorter disease duration in ALS. Specifically, our results from Figure 4-12 demonstrate that for each additional repeat expansion, the age of onset decreases by 0.468 years, although this result did not reach statistical significance ($p\text{-value}=0.21$).

Moreover, we found a significant association between the burden of repeat expansions and disease duration in ALS. Figure 4-13 shows that as the number of expansions increases, the duration of the disease decreases by 3.127 months ($p\text{-value}=0.01$, Figure 4-13).

Furthermore,, we turned our focus to analysing individual-level data related to the age of onset and the duration of repeat expansions within the cluster of 11 genes presented in Figure 4-9

(comprising *C9orf72*, *ATXN1*, *ATXN2*, *JPH3*, *CNBP*, *PABPN1*, *PPP2R2B*, *NUTM2B-AS1*, *TBP*, *GIPCI*, and *ARX*). The results of this analysis unveiled intriguing insights: first, it was observed that an increased number of repeats within these genes did not exhibit a statistically significant influence on the age of onset in ALS (p -value= 0.96, Figure 4-14). However, a compelling correlation emerged when assessing the impact on lifespan. Notably, a higher number of repeats was associated with a reduction in lifespan by an average of 3.857 months (p -value=0.02, Figure 4-15Figure 4-14).

This finding underscores the importance of understanding the genetic factors contributing to the duration of the disease course in ALS, shedding light on the potential influence of repeat expansions in this cluster of genes. Such insights provide valuable information for further research efforts aiming to elucidate the intricate interplay between genetic variations and disease progression in ALS.

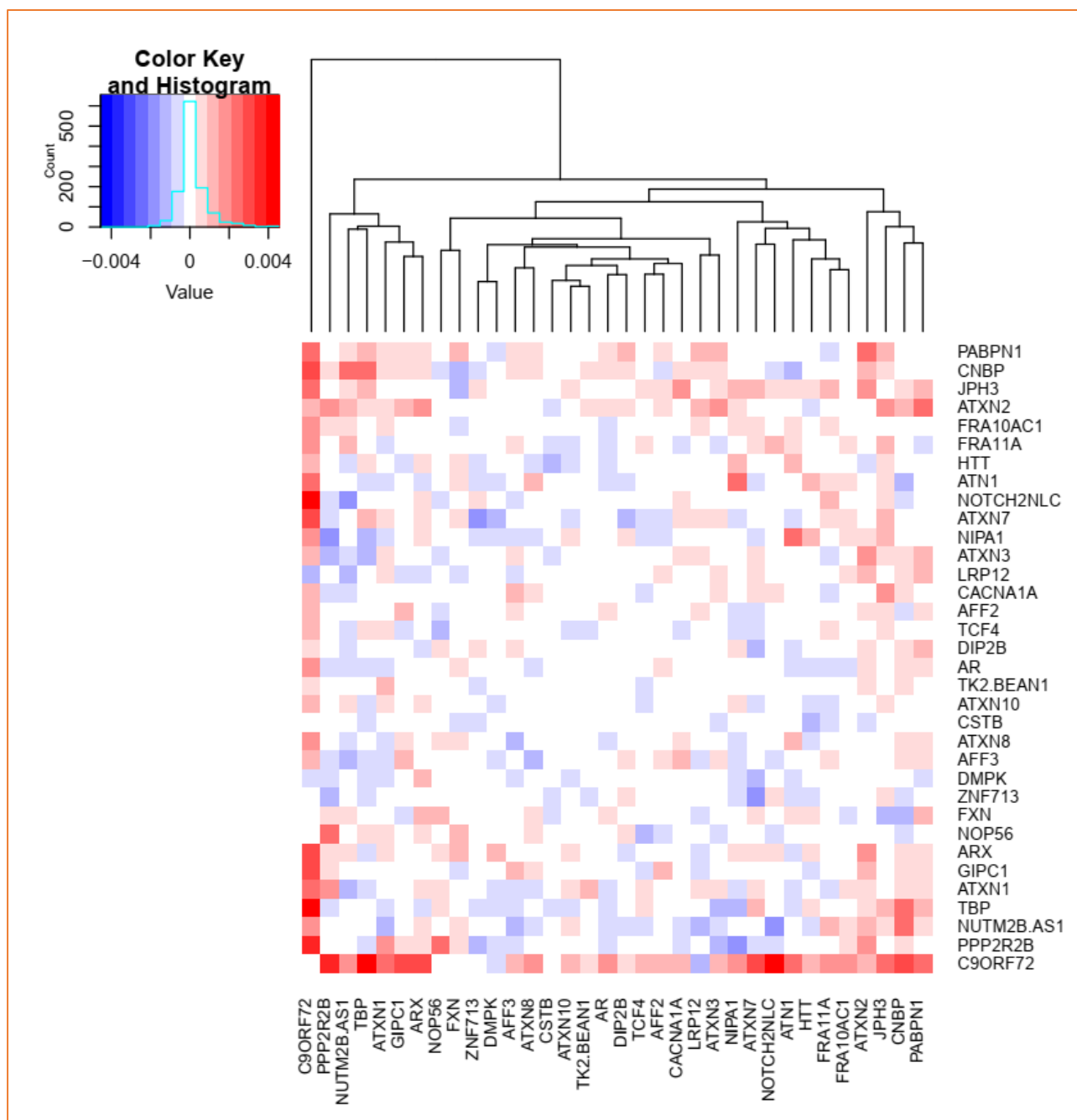


Figure 4-9 Gene-gene interaction including all NDRE genes.

Heatmap plot of the proportional difference between cases and controls of the number of individuals who had expanded alleles for every gene pair across all 34 NDRE genes. Colour key and histogram shows gene pairs where ALS proportions was higher in red and gene pairs where controls proportions were higher in blue.

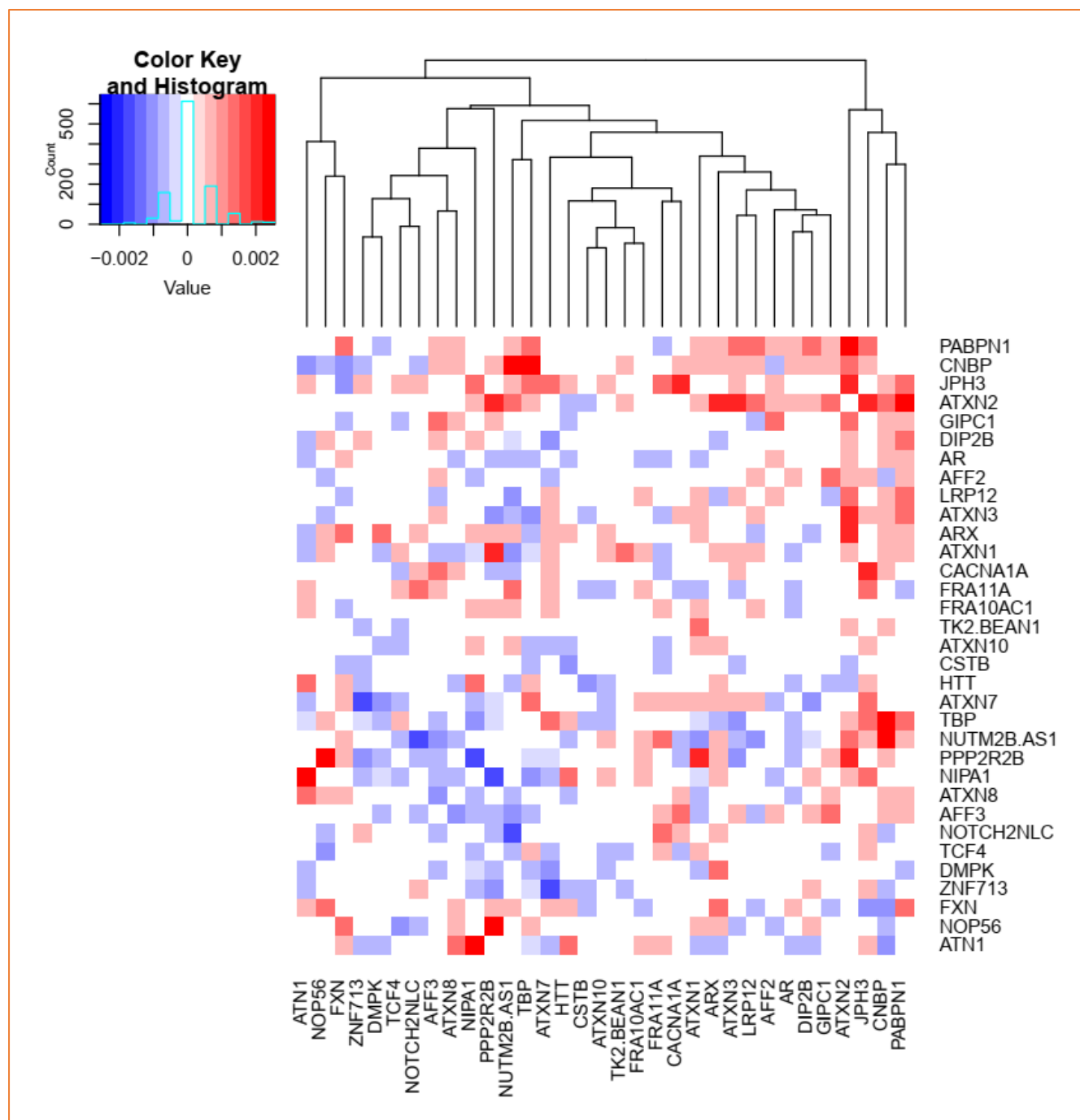


Figure 4-10 Gene-gene interaction including all NDRE genes, excluding *C9orf72*.

Heatmap plot of the proportional difference between cases and controls of the number of individuals who had expanded alleles for every gene pair across all 33 NDRE genes, excluding *C9orf72*. Colour key and histogram shows gene pairs where ALS proportions was higher in red and gene pairs where controls proportions were higher in blue.

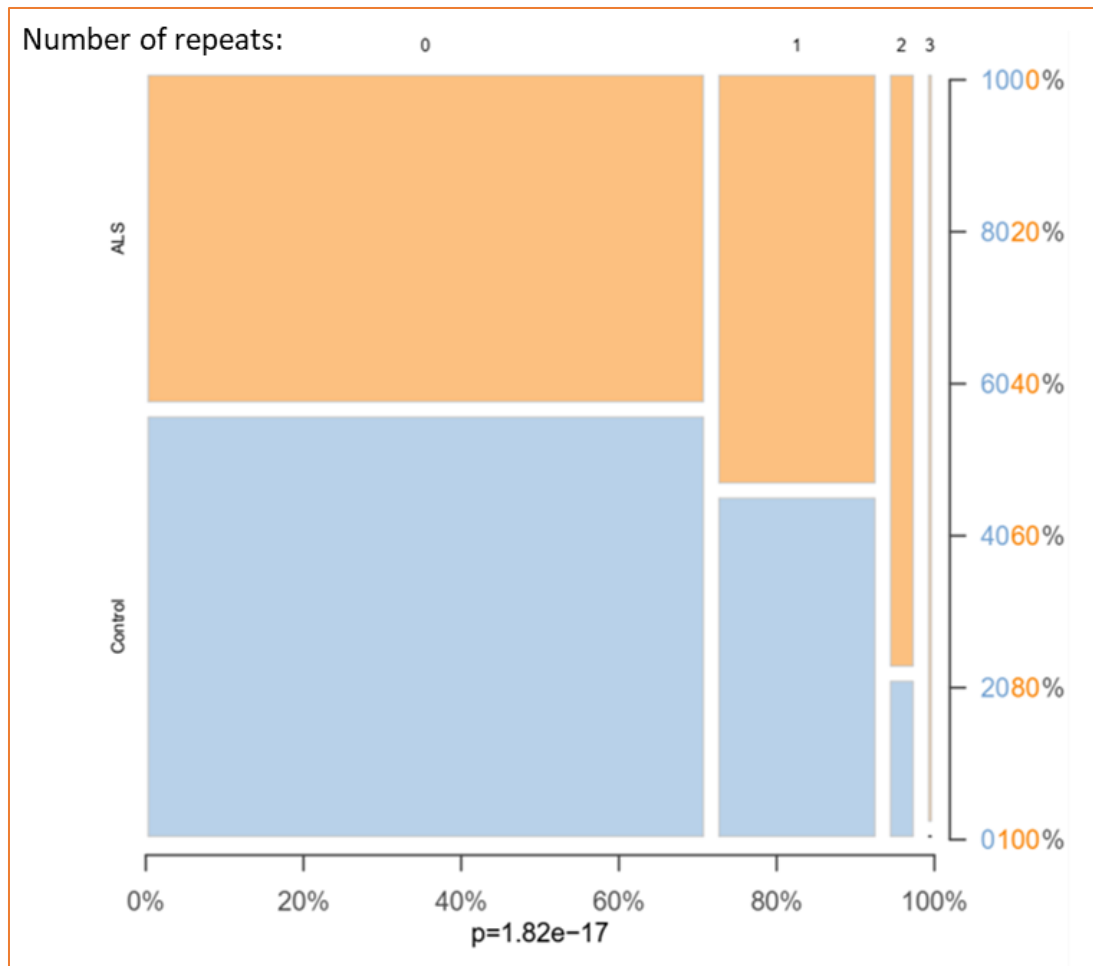


Figure 4-11 Burden from the cluster of 11 genes in ALS risk using 99th-percentile model.

Mosaic plot of proportions of cases and controls is shown for individuals carrying 0, 1, 2, or 3 long expansions in the following 11 genes *C9orf72*, *ATXN1*, *ATXN2*, *JPH3*, *CNBP*, *PABPN1*, *PPP2R2B*, *NUTM2B-AS1*, *TBP*, *GIPCI*, and *ARX*. As the number of expansions increases, the proportion of ALS cases increases ($p\text{-value}=1.82 \times 10^{-17}$). Percentages in orange correspond to ALS proportions, and in blue correspond to control proportions. Percentages in the bottom (black) correspond to the population proportions.

Table 4-2 Gene combination from ALS patients with 3 long expansions.

DNA bank ID	Disease	Number of long expansions	Gene 1	Gene 2	Gene 3
371	ALS	3	<i>C9ORF72</i>	<i>TBP</i>	<i>PABPN1</i>
707	ALS	3	<i>C9ORF72</i>	<i>CNBP</i>	<i>TBP</i>
1141	ALS	3	<i>C9ORF72</i>	<i>PABPN1</i>	<i>GIPCI</i>
248	ALS	3	<i>ATXN2</i>	<i>JPH3</i>	<i>CNBP</i>
649	ALS	3	<i>ATXN2</i>	<i>PPP2R2B</i>	<i>NUTM2B.ASI</i>
656	ALS	3	<i>C9ORF72</i>	<i>ATXN2</i>	<i>GIPCI</i>
854	ALS	3	<i>ATXN2</i>	<i>CNBP</i>	<i>ARX</i>
1232	ALS	3	<i>CNBP</i>	<i>PPP2R2B</i>	<i>GIPCI</i>
1238	ALS	3	<i>ATXN2</i>	<i>TBP</i>	<i>NUTM2B.ASI</i>
1687	ALS	3	<i>JPH3</i>	<i>PABPN1</i>	<i>ARX</i>
1719	ALS	3	<i>CNBP</i>	<i>PPP2R2B</i>	<i>ATXN1</i>

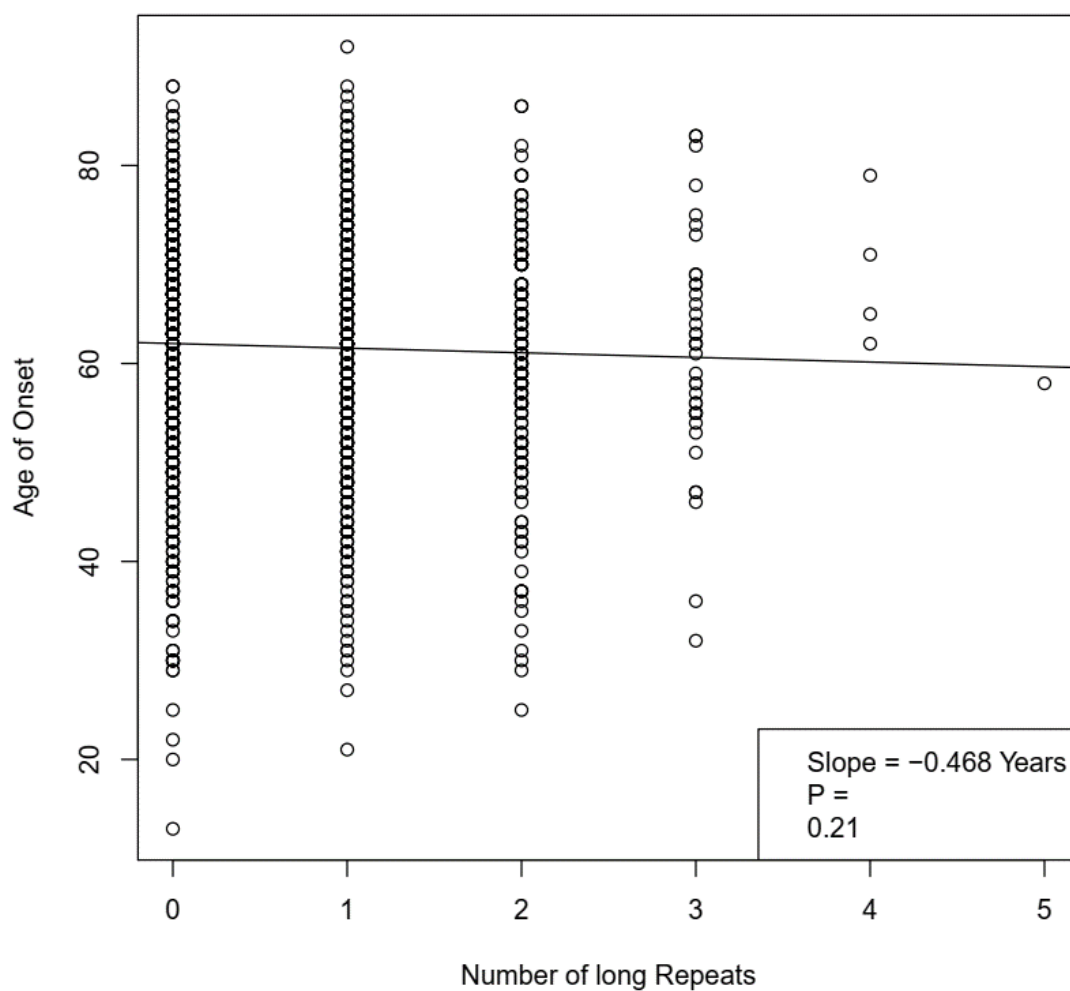


Figure 4-12 Burden of repeat expansions on age of onset.

Regression model compares the number of long repeats per individual with the age of onset. There is a negative effect between age of onset and number of long repeats (slope= -0.468 years, p -value= 0.21). As the number of long repeats increases, the age of onset decreases 0.468 years.

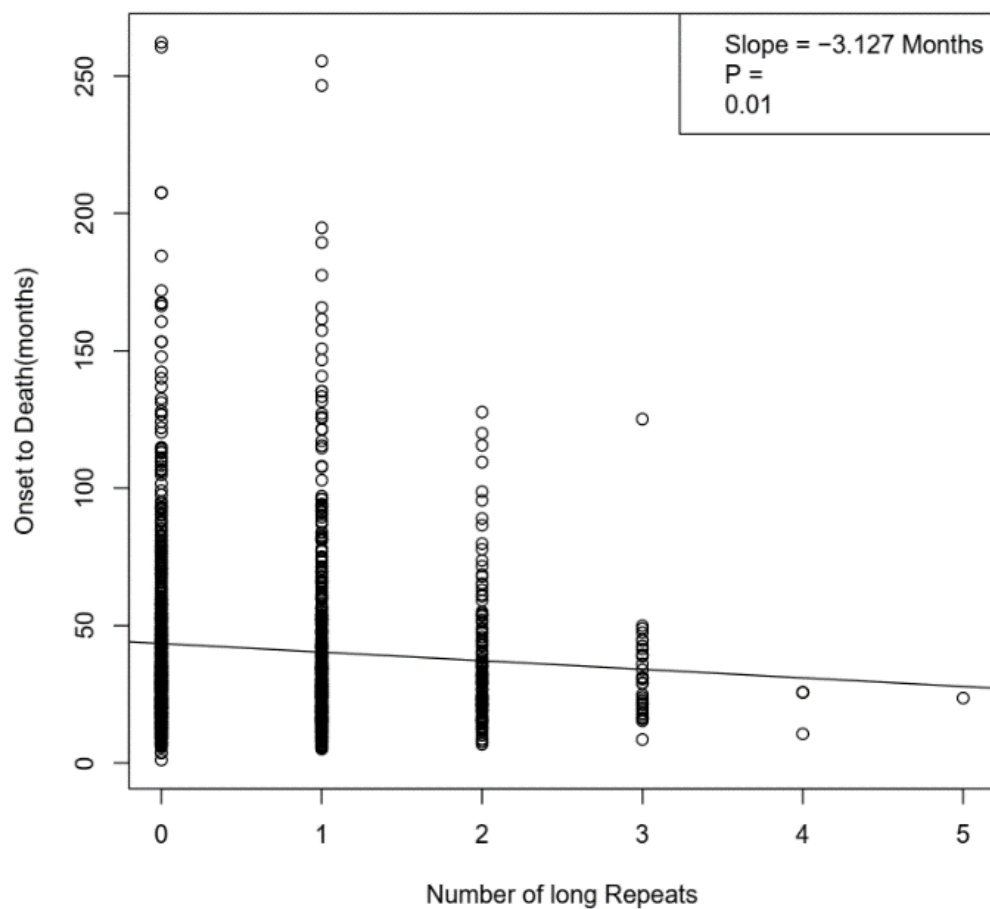


Figure 4-13 Burden of repeat expansions on disease duration (excluding one outlier).

Regression model compares the number of long repeats per individual with the disease duration. There is a statistically significant negative effect between disease duration and number of long repeats (slope= -3.127 months, p -value= 0.01). As the number of long repeats increases, the age of onset decreases 3.127 months.

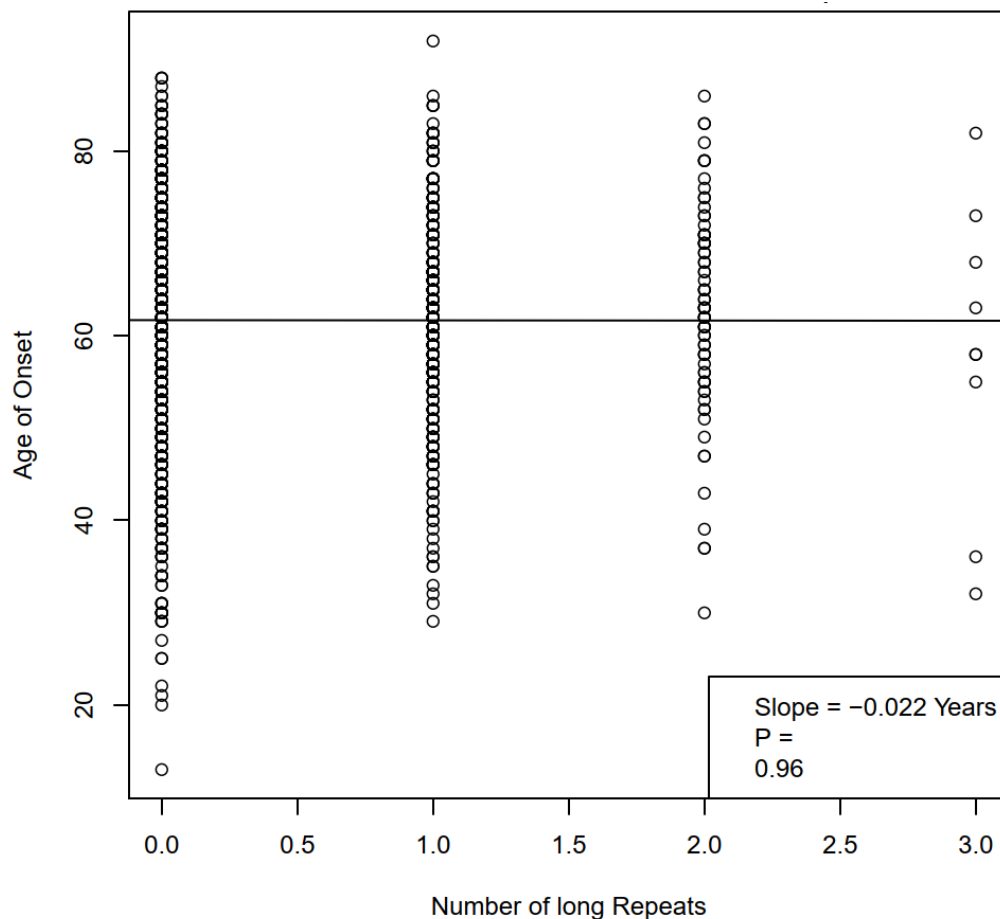


Figure 4-14 Burden of repeat expansions on age of onset from the cluster of 11 genes.

Regression model compares the number of long repeats per individual with the disease duration from the following 11 genes *C9orf72*, *ATXN1*, *ATXN2*, *JPH3*, *CNBP*, *PABPN1*, *PPP2R2B*, *NUTM2B-AS1*, *TBP*, *GIPC1*, and *ARX*. There is a negative effect between age of onset and number of long repeats (slope= -0.022 years, p -value= 0.96). As the number of long repeats increases, the age of onset decreases 0.022 years.

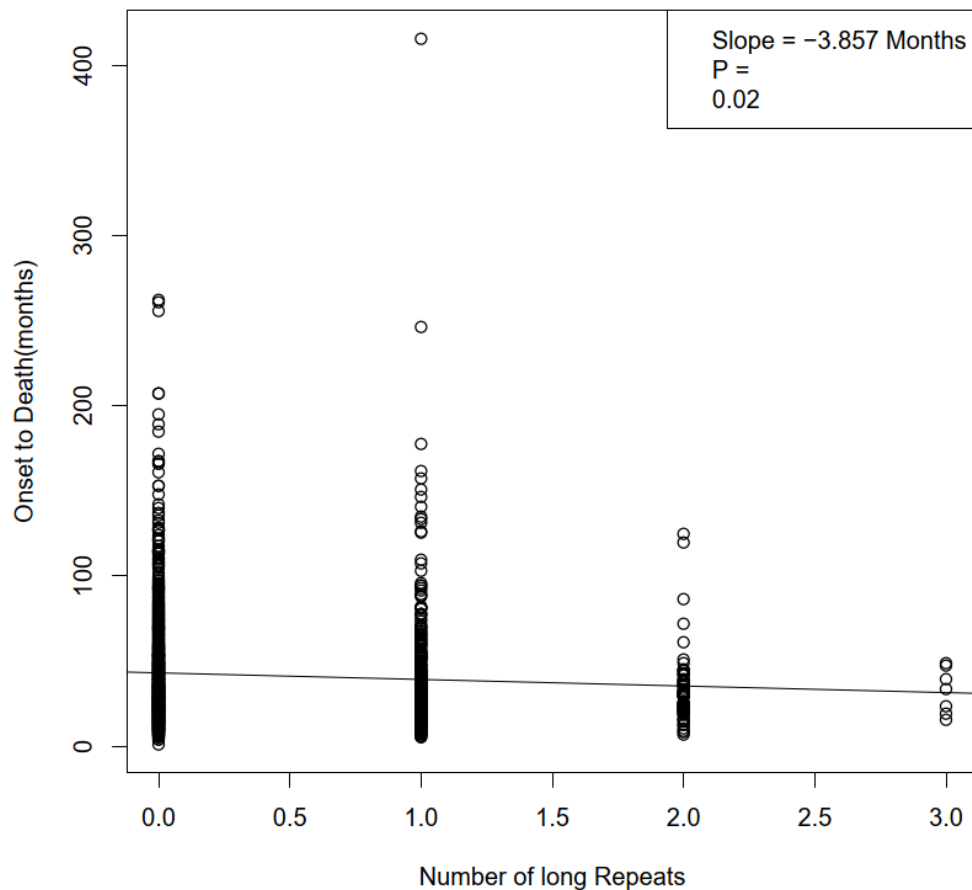


Figure 4-15 Burden of repeat expansions on disease duration from the cluster of 11 genes.

Regression model compares the number of long repeats per individual with the disease duration from the following 11 genes *C9orf72*, *ATXN1*, *ATXN2*, *JPH3*, *CNBP*, *PABPN1*, *PPP2R2B*, *NUTM2B-AS1*, *TBP*, *GIPCI*, and *ARX*. There is a statistically significant negative effect between disease duration and number of long repeats (slope= -3.857 months, p -value= 0.02). As the number of long repeats increases, the age of onset decreases 3.857 months.

4.4 Discussion

Burden of NDRE genes in ALS risk

In this study we assess the burden of 34 REs in the risk of ALS in the Irish population. Our findings reveal a significant increase in the proportion of ALS cases with increasing number of repeat expansions using two different methods. Results for the 99th-percentile method showed that the proportion of ALS cases increases to 100% when 5 repeat expansions are carried (p -value: 1.07×10^{-9} , Figure 4-1). Similarly, when using a threshold of 2 or more repeat expansions, the proportion of ALS cases is much greater (60%) above this threshold than below it (p -value 2.16×10^{-9} , Figure 4-5). This signal remains significant after the exclusion of *C9orf72*, *ATXN2* and *NIPA1* ($p=0.012$, Figure 4-8), indicating that the trend is not driven by these genes.

In our current investigation, we have embarked on a comprehensive exploration of the influence of 34 RE on the susceptibility to ALS within the Irish population. The consequential outcomes of our study provide compelling evidence of a substantial increase in the prevalence of ALS cases in direct correlation with an escalating count of genes with RE.

A particularly intriguing facet of our results lies in the potential implication of an amplified ALS risk associated with the concurrent presence of multiple NDRE genes. This ground-breaking revelation has introduced a myriad of inquiries, each with multiple plausible solutions. Particularly captivating is the statistical model's suggestion of a potential gene interaction in patients, subsequently intensifying the risk of ALS. This intriguing conjecture leads us to speculate about molecular processes where RNA and/or proteins originating from these NDRE genes might harbour an excessive number of repeats beyond the typical range. Such anomalies could potentially disrupt a spectrum of cellular interactions, thereby instigating irreparable disturbances and culminating in cellular damage, or even the formation of aggregates notorious for provoking inflammation, cellular toxicity, and eventual cell demise.

The path to unravelling the precise mechanisms underpinning this heightened ALS risk necessitates a two-fold approach. Firstly, the validation of our findings across a broader and more

diverse population is imperative. Secondly, researching into functional studies utilizing induced iPSC-derived cellular models offers a promising path to dissect the intricate mechanisms of multiple NDRE genes impact on ALS. By bridging the gap between statistical observation and mechanistic understanding, these functional studies hold the potential to illuminate the intricate workings of these genes and their contributions to the complex landscape of ALS.

Gene-Gene Interactions in NDREs and their Relationship to ALS Risk in the Irish Population.

Our study aimed to assess the potential impact of interactions between the 34 NDRE genes on ALS risk in the Irish population. Using cluster analysis visualised in the heatmaps in Figure 4-9 and Figure 4-10, we identified a distinctive cluster of 11 genes, including *C9orf72*, *ATXN1*, *ATXN2*, *JPH3*, *CNBP*, *PABPN1*, *PPP2R2B*, *NUTM2B-AS1*, *TBP*, *GIPCI*, and *ARX*, in which the proportion of ALS cases exhibiting mutual co-occurrence of expansions was significantly higher than controls. We then examined the burden of these 11 genes using the chi-square test (visualised as a mosaic plot), which revealed a statistically significant increase in ALS risk with the number of long repeats ($p\text{-value} = 1.82 \times 10^{-17}$, Figure 4-11). Interestingly, our results also demonstrated that ALS risk increases to 100% when the number of repeats reaches 3. These findings support the hypothesis that multiple repeat expansions in a single patient increase the risk of developing ALS, and suggest that the underlying mechanisms of the disease are not only associated with the risk of one gene, but also with the burden of multiple NDREs. The findings also point to the possibility that there remain to be discovered many more repeat expansions that contribute to ALS risk; as these are discovered and included in similar analyses, we expect that the proportion of ALS cases with zero expansions will decrease, and the number of possible expansions carried by individuals will increase.

Age of onset and disease duration

Our study aimed to investigate the impact of repeat expansions on age of onset and disease duration in ALS patients. While our analysis did not reveal a statistically significant association between repeat expansions and age of onset, our results showed a trend towards younger onset with

increasing repeat expansions. Additionally, we found a significant association between the burden of repeat expansions and disease duration, with a decrease in duration of 3.127 months as the number of expansions increased (p -value=0.01, Figure 4-13). These results suggest that the burden of repeat expansions in multiple genes may be a useful predictor of age of onset and disease duration in ALS, particularly as more pathogenic repeat expansions are discovered in the future. However, further studies with larger sample sizes are needed to confirm and validate these findings.

Previous studies have suggested that ALS may have both an oligogenic and polygenic basis, where multiple gene variants may act together to cause or influence the clinical manifestation of the disease (Van Blitterswijk et al. 2012; Keogh et al. 2018; Williams et al. 2013). It is believed that ALS develops through multiple steps (Al-Chalabi et al. 2014) and that oligogenic and polygenic factors may contribute to the highly variable presentation and course of the disease, even in cases with identical causal mutations (Black et al. 2017; Morgan et al. 2017). The discovery by McCann and colleagues of multiple ALS-implicated variants in 18 genes, including well-known genes such as *C9orf72*, *ATXN2*, *TARDBP*, *SOD1*, *SQSTM1* and *SETX*, supports this potential oligogenic or polygenic basis of ALS. Particularly significant is the finding that 6.82% of Australian sporadic ALS cases carry two or more variants, which implies that these variants may act in concert to cause or influence the course of the disease. These results lend support to the idea that the development of ALS is a multistep process that may involve oligogenic or polygenic factors, and highlight the need for further research to investigate the role of these variants in disease onset and progression (McCann et al. 2017).

Recently, two separate studies have revealed that intermediate repeat lengths in *ATXN2*, which were previously linked to an increased risk of ALS, may act as modifiers of the phenotype in ALS/FTD patients with the *C9orf72* expansion (van Blitterswijk et al. 2014; Elden et al. 2010; Lattante et al. 2014). The *C9orf72* expansion has been associated with various mutations in ALS and FTD patients, including *TARDBP*, *FUS*, *SOD1*, *PGRN*, *MAPT*, *PSEN-2*, and *SQSTM1*. In addition, it has been linked to variants in less common genes such as *ANG*, *UBQLN2*, *VAPB*, *SETX*,

ALS2, and others (Lattante et al. 2015). Double mutations have also been observed in patients without *C9orf72* mutations. Furthermore, patients with mutations in *TARDBP*, *FUS*, and *SOD1* were discovered to carry variants in *ANG*, which is a low-risk ALS gene (Van Blitterswijk et al. 2012; Kenna et al. 2013; Lattante et al. 2015; Luigetti et al. 2011).

Polygenic risk scores have been suggested as a way to categorize patients with ALS, especially those with SALS. These scores estimate an individual's risk of developing the disease based on the number and risk level of gene variants they carry. The complexity of ALS makes it difficult to design clinical trials and apply treatments effectively to all cases (Kiernan et al. 2011; Swinnen and Robberecht 2014; Tiryaki and Horak 2014). Using polygenic risk scores to classify patients has the potential to increase the success of therapeutic trials; however currently the predictive value of SNP-based scores is extremely low (SNP-based heritability $\approx 2.8\%$, van Rheenen *et al.*, 2021). Nevertheless, McCann and colleagues' findings indicate that patients with multiple ALS-implicated gene variants tend to have earlier onset of the disease, suggesting that polygenic risk scores could be useful in clinical settings and trial design (McCann et al. 2021). The results presented in this chapter support these conclusions, while highlighting the great importance of genotyping repeat expansions as major contributors to ALS risk and onset in addition to single nucleotide variants.

Our study provides important insights into the complexity of genetic mutations in ALS and the support of an oligogenic basis of ALS. We found that a significant number of individuals harbour more than one potentially pathogenic mutation, highlighting the need for comprehensive screening of NDRE genes rather than single-gene testing. Furthermore, despite the fact that our genotyping panel consisted of only 34 NDRE genes in a small cohort of ALS cases and controls, results showed that the burden of repeat expansions is associated with a significant increase in ALS risk, with 100% of individuals that carry 5 different expansions being affected by ALS. In addition, we found that a higher number of repeats was associated with a decrease in age of onset and disease duration. These findings underscore the need for continued efforts to identify additional genes and genetic mechanisms influencing ALS risk. Improved understanding of the genetic underpinnings of ALS may inform the development of targeted treatments for individuals at high risk of developing this devastating disease.

The etiological factors contributing to ALS susceptibility can be broadly categorized into established elements encompassing genetics, age, and sex, alongside environmental variables. While our present investigation primarily focuses on the former, it's imperative to acknowledge the absence of comprehensive data within the Irish ALS DNA bank to systematically scrutinize environmental ALS risk factors, encompassing pivotal factors like smoking, exposure to environmental toxins, and intriguingly, military service – the mechanistic basis of the latter remains elusive.

While the study's limitations preclude an exhaustive investigation of environmental factors, it's imperative to acknowledge their documented significance. Chemical toxins and smoking, recognized contributors to heightened ALS risk, necessitate dedicated inquiry. Consequently, an avenue for forthcoming investigation emerges: the establishment of a distinct study centred on the burden of environmental factors. This approach is merited by the well-established correlation between these factors and ALS risk, thereby warranting meticulous scrutiny to unravel the intricate connections within the Irish context.

Future research endeavours should aim to holistically address the environmental dimension, contingent upon data accessibility. An autonomous investigation dedicated to parsing the complexities of environmental influences stands as an imperative pursuit, given their demonstrable impact on ALS susceptibility. In essence, the interplay between genetic and environmental determinants demands concerted investigation for a comprehensive understanding of ALS susceptibility. By affording both facets the requisite analytical attention, we can inch closer to a more holistic comprehension of ALS and its multifaceted underpinnings.

4.5 Summary

The study found that the risk of developing ALS increases with the number of long repeats in certain genes, with up to 5 expansions associated with higher risk (p -value: 1.07×10^{-9}). Our study investigated the relationship between long repeat expansions in 11 specific genes and the risk of

developing ALS. Our results showed that the risk of developing ALS increases with the number of long repeat expansions, and not just with specific genes like *C9orf72*, *ATXN2*, or *NIPA1*. We also found an association between the number of repeat expansions and the age of onset, with a decrease of onset with increasing repeat expansions (-0.468 years). In addition, our study showed that the burden of repeat expansions is significantly associated with disease duration, with a decrease of 2.523 months in duration as the number of expansions increased. Overall, our findings highlight the importance of considering repeat expansions in multiple genes when evaluating ALS risk, and suggest potential implications for disease management and treatment.

Chapter 5

Exploring the utility of targeted third-generation sequencing as a method for repeat expansion genotyping

5.1 Introduction

Next generation sequencing (NGS) is a viable alternative for genotyping STRs, as it allows for detection of STR expansions across the entire genome using short-read NGS platforms like Illumina. A number of bioinformatics tools like ExpansionHunter, LobSTR, RepeatSeq, HipSTR and GangSTR have been developed for this purpose, which enable identification of novel expansions and linkage of STR regions in affected family members (Dolzhenko et al. 2017; Halman and Oshlack 2020; Kumar, Cowley, and Davis 2019; Mousavi et al. 2019). Whole-genome sequencing has the added advantage of its ability to profile all STRs in the genome simultaneously, as well as non-STR mutations implicated in disease. Although NGS remains relatively expensive, it can be cost-effective by avoiding repeated molecular testing on multiple targets, and its cost is expected to fall further over time (Chintalaphani et al. 2021).

However, the utility of short-read NGS for repeat expansion diagnosis is limited by several factors. Firstly, highly repetitive, and GC-rich genome regions are difficult to profile due to inadequate coverage caused by difficulties in library preparation, PCR amplification, and sequencing. PCR amplification during library preparation can introduce stutter errors, which can be mitigated

through the use of PCR-free library preparations (Smith et al. 1994). Secondly, the repetitive nature of STR regions can cause ambiguous alignment or misalignment of short NGS reads to the reference genome. Thirdly, the short-read length (~100-150 bp) of established NGS technologies is insufficient to span large STR expansions, making it impossible to accurately determine their length. Lastly, standard NGS cannot detect epigenetic modifications, such as 5-methylcytosine, which are diagnostically important in some cases (Meienberg et al. 2016; Russ et al. 2015). Despite the usefulness of NGS for discovery of new disease-related repeat expansions, these limitations have hindered its widespread adoption for clinical diagnosis and replacement of low-throughput molecular tests like Southern blotting (Chintalaphani et al. 2021).

Recently, a number of new genomic sequencing technologies have emerged, collectively termed third-generation sequencing (3GS). One family of 3GS devices, manufactured by Oxford Nanopore Technologies (ONT), measures the displacement of ionic current as a DNA strand passes through a biological nanopore, which is then translated into DNA sequence information. Unlike short-read NGS, ONT sequencing has no upper limit on read length, with an average read length of >10 kb considered standard for genomic DNA sequencing and some examples achieving maximum read lengths in excess of 1 Mb (Mantere, Kersten, and Hoischen 2019). This enables individual ONT reads to span the entire length of large pathogenic repeat expansions, making it possible to directly measure STR lengths (Ebbert et al. 2018). Also, ONT has developed a portable, real-time DNA sequencing platform called the MinION. The MinION is a small, USB-powered device that can sequence long DNA fragments with a read length of up to several hundred kilobases. Despite its relative novelty, the MinION has already been used for a diverse array of applications, from sequencing the Ebola virus during the 2015 outbreak in West Africa (Hoenen et al. 2016) to monitoring the gut microbiome of astronauts on the International Space Station (Stahl-Rommel et al. 2021).

Although ONT sequencing currently exhibits relatively high sequencing error rates compared to NGS, accurate consensus sequence determination can be achieved with sufficient coverage (Jain et al. 2018). Several studies have demonstrated the accurate genotyping of repeat expansions with ONT (Sone et al. 2019). Furthermore, analysis of ONT signal data allows the methylation status of a given locus to be determined in parallel, which provides an additional marker for the diagnosis

of relevant repeat expansion disorders such as FXS (Giesselmann et al. 2019). Overall, ONT sequencing offers a promising approach towards accurately detecting and characterizing repeat expansions, which has important implications for understanding the genetic basis of various diseases (Chintalaphani et al. 2021).

An alternative 3GS technology, manufactured by Pacific Biosciences and termed Single Molecule Real-Time (SMRT) sequencing, uses fluorescent signals to detect nucleotide incorporation in real-time during DNA sequencing. The base-level accuracy of SMRT sequencing is greater than 99%, achieved through circular consensus sequencing (CCS) that involves ligating large DNA strands to form circular DNA molecules, allowing high coverage with a long average read length of 13.5 kb (Rhoads and Au 2015; Wenger et al. 2019). The long and highly accurate reads produced by SMRT sequencing have the potential to accurately detect the length and sequence of STRs and to identify variants in surrounding regions. Despite the recent advancements in this field, accurately genotyping STRs using long-read sequencing data remains an active area of development (Rhoads and Au 2015). Various software packages have been developed for this purpose, but benchmarking is required to ensure accuracy and to address issues such as the detection of novel pathogenic STR expansions and the resolution of challenging cases such as mixed conformations and allelic differences. While the potential of ONT and PacBio SMRT technologies to replace current clinical molecular diagnostics is high, more research and development is necessary to address the computational analysis of long-read sequencing data (Mitsuhashi et al. 2019; Rhoads and Au 2015).

Nevertheless, the utilization of long-read sequencing technologies from ONT and PacBio for whole-genome analysis is now feasible and has the potential to lead to the discovery of numerous novel disease-related STR expansions in the near future. For instance, recent studies have identified ‘GGC’ and ‘TTTCA’ repeat expansions in the *NOTCH2NLC* and *SAMD12* genes, respectively, using long-read sequencing combined with bioinformatics tools and other methodologies (Sone et al. 2019). However, due to the high costs and large data volumes associated with whole-genome sequencing, targeted sequencing of candidate genes offers a more cost-effective and practical approach for clinical implementation. This approach requires the

development of reliable methods for amplification-free enrichment and sequencing of long DNA fragments spanning STR regions (Ishiura et al. 2018; Zeng et al. 2019).

CRISPR-Cas9 guide-ribonucleoprotein (RNP) is a promising method for the selective cleavage of target loci in the identification of STR expansions in genes such as *HTT*, *C9orf72*, *ATXN10*, and *NOTCH2NLC*, followed by the ligation of a magnetic adaptor to allow for the isolation of target molecules before PacBio SMRT sequencing (Sone et al. 2019; Tsai et al. 2017). A similar approach can be applied to ONT sequencing by directly ligating ONT sequencing adapters to Cas9 cleavage sites for selective sequencing (Giesselmann et al. 2019; Gilpatrick et al. 2020). Giesselmann *et al.* successfully established this approach and generated more than 40-fold coverage over the expanded 'GGGGCC' region in *C9orf72*, sufficient for accurate determination of repeat length (Giesselmann et al. 2019). Furthermore, the Cas9-mediated target enrichment technique is amenable to multiplexing, making it feasible to target multiple disease alleles simultaneously, thus improving efficiency and cost-effectiveness in diagnosis. This approach was successfully demonstrated by Tsai *et al.*, who achieved 150-2000 fold coverage depth with SMRT sequencing on multiple targets in a single assay, including *C9orf72*, *HTT*, *FMR1*, and *ATXN10*. This capability is advantageous from a diagnostic perspective, avoiding the need to order multiple tests as in standard molecular diagnostics (Tsai et al. 2017).

Research aims

The objective of this chapter is to explore the utility of Nanopore Technology in generating long-read sequences for genotyping repeat expansions. To achieve this, we will utilize a Cas9-enrichment protocol that does not rely on PCR amplification. The resulting genotypes will be compared to those obtained via PCR in Chapter 3, in order to evaluate the performance of this alternative approach.

5.2 Methodology

Study Participants

For this study, we used four different ALS patients and one control previously included in the multiplex PCR genotyping protocol. gDNA and concentrations were provided by the Irish ALS DNA bank. PCR genotypes were obtained from the results of Chapter 3.

Cas9-mediated enrichment coupled to ONT sequencing

For this project, we used a Cas9 mediated PCR free enrichment to sequence long fragments of DNA that includes the RE expansion in question using the MinION sequencer. The protocol consists of two phases. First, a tiling approach which requires probes that would be prepared separately (RNP complex formation, cleavage and dA-tailing and adapter ligation performed in separate tubes) and then pooled during the final AMPure XP bead purification (Figure 5-1).

We designed crRNAs using two online tools, first CHOPCHOP (<https://chopchop.cbu.uib.no/>) following ONT's recommendations (Nanopore Community 2018), making sure that the excised fragment was at least 3 kbp in length. For the regions that CHOPCHOP was not able to provide a valid crRNA, we used the Predesigned Alt-R™ CRISPR-Cas9 (https://www.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE) guide RNA. Candidate crRNAs were manually checked for unique mapping by aligning them to the human genome (hg38) using Snap gene viewer (Version 6.0.5). The final crRNAs were prepared by Integrated DNA Technologies and pooled together in a single tube; sequences are in Supplementary Table S 4-1 and Supplementary Table S 4-2. *HTT* probes were ordered separately from the pooled probes to test the protocol as ONT recommended. The reaction mixture comprised 2 µL of each crRNA (0.75 µL of *HTT* probes + 1.25 µL pooled probes), 1.25 µL of transactivation crRNA (tracrRNA) and 0.8 µL (62 µM) Alt-R S.p. HiFi Cas9 Nuclease v3 in 1× CutSmart Buffer for the generation of ribonucleoprotein (RNP) complexes (all components from Integrated DNA Technologies) modified from the ONT protocol (version ENR_9084_v109_revT_04Dec2018).

The dephosphorylation, Cas9-mediated digestion, and dA-tailing of 210/μl input genomic DNA were carried out according to the ONT protocol. The genomic DNA was incubated with the RNPs for 10 min at 37°C and then 2 min at 80°C for enzyme inactivation. Sequencing adapters (AMX) were ligated to the cleaved and dA-tailed target ends for 10 min at room temperature before stopping the reaction by adding one volume of 10 mM Tris–EDTA (pH 8.0). Short fragments (<3 kbp) and residual enzymes were removed by adding 0.3× AMPure XP beads (Beckman-Coulter, Brea, CA) and washing twice in long fragment buffer (ONT). The DNA was eluted by incubating for 10 min at room temperature in elution buffer (ONT). The purified library was mixed with 37.5 μl sequencing buffer (ONT) and 25.5 μl of library loading beads (ONT). The library was loaded onto a FLO-MIN106D (R9.4.1) flow cell in a MinION and sequenced using MinKNOW v22.05.5 (ONT) for 72 hours.

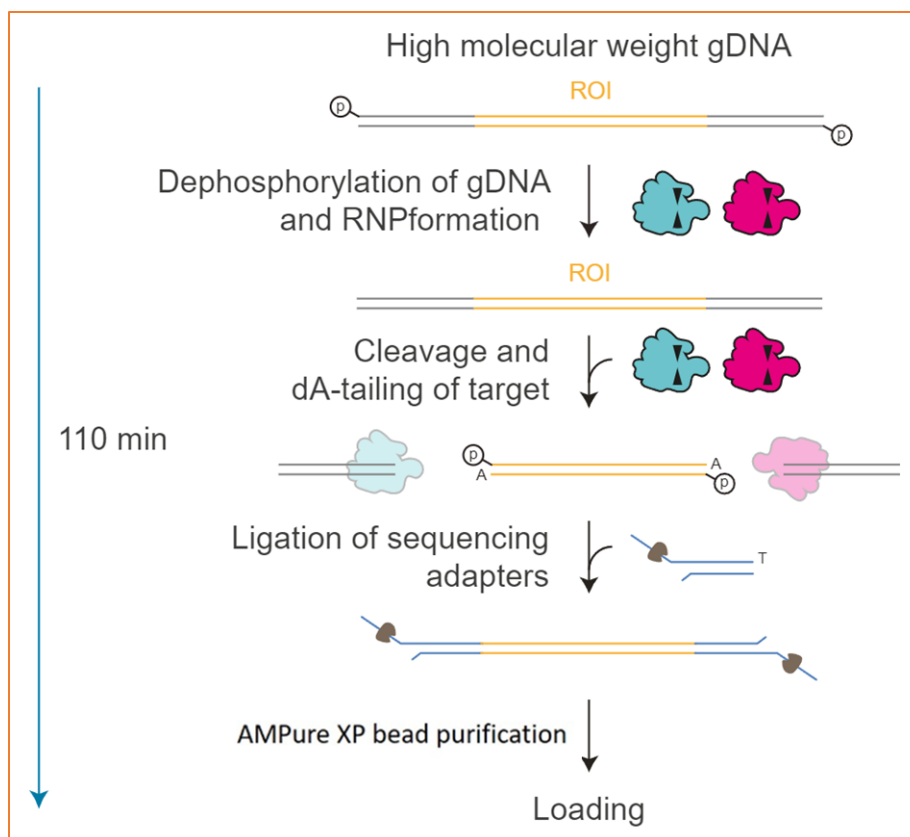


Figure 5-1 Cas9-mediates PCR-free enrichment preparation for sequencing.

ROI: region of interest (Taken from Oxford Nanopore Technologies: https://community.nanoporetech.com/docs/prepare/library_prep_protocols/cas9-targeted-sequencing/v/enr_9084_v109_rev_04dec2018).

Data processing and analysis

Raw data were collected in fast5 format from the Nanopore sequencer and processed with a variety of software tools described in the following sections in a Linux command-line environment (Ubuntu 22.04.2). Statistical analysis and plotting was performed using R version 3.6.3.

Basecalling using Guppy

Guppy is a bioinformatics toolkit that allows for immediate identification of DNA bases and a variety of subsequent processing functions on sequencing platforms manufactured by Oxford Nanopore Technologies™. Although real-time basecalling is possible with ONT's MinKNOW

software, for this study the basecalling process was deferred until after completion of each sequencing run. We used the following command line for basecalling:

```
guppy_basecaller -input_path reads -save_path output_folder/basecall --config  
dna_r9.4.1_450bps_hac.cfg
```

ONT sequence data analysis

Basic Aligners

In this study, the initial alignment of large Oxford Nanopore sequencing reads to the reference genome was performed using **bwa**, which is a widely used industry-standard aligner. However, the results of the bwa alignment were compared to those obtained using **ngmlr**, which is a newer long-read mapper specifically designed to align PacBio or Oxford Nanopore reads to large reference genomes. The primary objective of ngmlr is to accurately align reads, including those that span complex structural variations, in a fast and efficient manner. To achieve this, ngmlr uses an SV (structural variation) aware k-mer search to identify approximate mapping locations for a read, followed by a banded Smith-Waterman alignment algorithm to compute the final alignment (Sedlazeck et al. 2018).

One of the unique features of ngmlr is its use of a convex gap cost model that assigns lower penalty scores to gap extensions for longer gaps compared to shorter ones, enabling the tool to compute highly precise alignments for long structural variants. This gap model allows ngmlr to simultaneously account for both sequencing errors and real genomic variations, making it particularly effective at accurately identifying breakpoints stemming from structural variations. The k-mer search also enables ngmlr to detect and split reads that cannot be aligned linearly, enabling the mapper to reliably align reads to a diverse range of different structural variations, including nested SVs such as inversions flanked by deletions (Sedlazeck *et al.*, 2018).

We used the following terminal code for aligning with *ngmlr*:

```
~/ngmlr -t 4 -r GRCh38_latest_genomic.fna -q input_file_name.fastq -o
output_file_name.sam -x ont
```

where `input_file_name.fastq` and `output_file_name.sam` were specific to each sequenced sample. Data were then compressed to BAM format and indexed using `samtools v1.13`.

Following alignment, we used the Integrative Genomics Viewer (IGV) to visualize all alignments obtained from `bwa` and `ngmlr` (version 2.12.2).

RE genotyping

RepeatHMM

RepeatHMM software aims to accurately detect microsatellites, including trinucleotide repeats, in subjects of interest using long-read sequencing data. The tool offers efficient estimation of repeat counts and allows for the analysis of all types of simple repeats. RepeatHMM includes pre-defined models for over 10 well-known trinucleotide repeats, such as *ATXN3* and *FMRI*, commonly associated with trinucleotide repeat disorders (Liu et al. 2017).

To accomplish its goal, RepeatHMM involves a three-step process. First, the tool uses a split-and-align strategy to align long-reads to a reference genome, which enhances alignment quality by splitting reads at repeat regions and aligning them separately. Second, error correction is performed to remove sequencing errors that may have arisen during the sequencing process. Finally, RepeatHMM utilizes an HMM and a peak calling algorithm based on Gaussian mixture model to accurately estimate repeat counts (Liu et al. 2017).

RepeatHMM was called for each gene in the PCR panel from Chapter 3 for which a predefined model (stored in `hg38.predefined.pa`) was available (*AR*, *ATNI*, *ATXN1*, *ATXN10*, *ATXN2*, *ATXN3*, *ATXN7*, *ATXN8*, *CACNA1A*, *DMPK*, *FMRI*, *FXN*, *HTT* and *PPP2R2B*; stored in a file called `rhmmgenes`) using the following command line script:

```
for person in 1021 1052 1143 1675 Control_01
do echo $person >> rhmmresults
cat rhmmgenes | while read gene
do python repeatHMM.py BAMinput -hgfile GRCh38.fna \\\
```

```

--Onebamfile <bam data> --repeatName $gene \\
--Patternfile hg38.predefined.pa \\
--SeqTech Nanopore \\
--MinSup 0 | grep $gene | tail -3 | head -2 >> rhmmresults
done
done

```

where *<bam data>* represents the name of the bam-format aligned Nanopore sequence dataset for each individual identified using the variable *\$person*.

After processing all sequences using RepeatHMM software, results were compared with the PCR genotyping in Chapter 3. This comparison was plotted to visualize how well or how differently RepeatHMM was calling repeats in genes for which both PCR and Nanopore data were available. For this analysis, PCR data were considered to represent the ground truth for each genotype. Additionally, to explore similarities and differences between PCR and RepeatHMM genotypes, and to better understand the information available to RepeatHMM to make its genotype calls, the distribution of apparent allele lengths across all reads at each locus was plotted to visualise the depth of coverage of sequenced alleles used to support the final RepeatHMM genotypes.

Sniffles2

An important omission from RepeatHMM's predefined models is the hexanucleotide repeat expansion in *C9orf72*, which causes a substantial fraction of ALS and FTD cases. We therefore explored other possibilities for genotyping REs using the available Nanopore data. Sniffles2 is a structural variant (SV) caller that is optimized for fast and accurate detection of SVs in long-read sequencing data, including germline, somatic, and population-level data from both PacBio and Oxford Nanopore platforms. It achieves this by performing successive scans of the alignments to identify different types of SVs. (Sedlazeck *et al.*, 2018).

To evaluate candidate SVs, Sniffles2 uses an SV-scoring scheme that considers various factors such as their size, position, type, coverage, and breakpoint consistency, which helps to overcome the high insertion/deletion (indel) error rates commonly observed in long-read sequencing. One of the most innovative features of Sniffles2 is its ability to detect nested SVs, such as inverted tandem duplications (INVDUPS) and inversions flanked by indels (INVDELs) (Sedlazeck *et al.*, 2018).

For each Nanopore sequence dataset, we attempted SV calling using Sniffles2 using the following command line script:

```
for file in 1021 1052 1143 1675 Control_01
do echo $file
sniffles -I "$file".bam -v "$file"_sniffles.vcf \
--tandem-repeats human_GRCh38_no_alt_analysis_set.trf.bed
```

where the file `human_GRCh38_no_alt_analysis_set.trf.bed` contains information that allows Sniffles2 to more accurately genotype SVs in regions containing repetitive sequence. Following genotyping, resulting VCF format data were compressed using `bgzip v1.13+ds` and indexed using `tabix v1.13+ds`.

[SquiggleKit](#)

To better understand irregularities with the genotyping available from Sniffles2, the raw signal from the Nanopore sequencer was analysed using SquiggleKit. This is a versatile toolkit designed for the extraction, processing, and visualization of raw nanopore sequencing signal data. Apart from its data mining applications, SquiggleKit can be used to manage the vast amount of data files generated during nanopore sequencing, and serves as a starting point for the development of new tools based on nanopore signal data (Ferguson and Smith 2019). Raw signal was extracted from fast5-format Nanopore data using SquiggleKit's `SquigglePull.py` script with the following command line:

```
python3.8 SquigglePull.py -e -p . > <output.tsv>
```

where `<output.tsv>` held the extracted read-level current signal for all reads in the fast5 file. Relevant reads were then retrieved by read ID using the Linux `grep` command and subsequently visualised and analysed using R version 3.6.3.

5.3 Results

After careful evaluation, it was found that ngmlr provided more sensitive and accurate alignments of the large Oxford Nanopore sequencing reads to the reference genome than bwa. Figure 5-2 shows a comparison of the two aligners at the *AFF3* RE location in the control sample; bwa results are on the top panel and ngmlr on the bottom. Importantly, ngmlr managed to align 6 different sequences to the reference genome while bwa was capable to do only two. Secondly, the deletion of 12 bases is aligned at the end of the GCC region in ngmlr, while bwa located it at the beginning. Furthermore, ngmlr alignments result in similar insertions between sequence reads, whereas with bwa alignments expanded alleles appear at different locations and different lengths between reads. These differences are due to the more suitable alignment approach of ngmlr software while processing long-read sequencing data and repeat expansions.

RepeatHMM

This study aimed to evaluate the accuracy of RepeatHMM in calling the number of repeats. To achieve this, we called genotypes for all samples (where sufficient data were available) and compared the results obtained by RepeatHMM to the PCR genotype of the same sample. Figure 5-3 presents the results of the comparison of RepeatHMM calls for the number of repeats for each allele per gene.

Overall, we observed reasonable concordance between the PCR genotypes and the RepeatHMM results, excluding the *FMRI* genotypes. The variance in the allele calls by RepeatHMM was well explained by PCR alleles ($R^2=0.73$, p -value: 1.07×10^{-31}). However, there were some discrepancies between the two methods, indicating that RepeatHMM may not always provide identical results to PCR genotyping.

Specifically, we found that RepeatHMM accurately called genotypes for allele 1 and 2 in the *PPP2R2B* gene, whereas in contrast, for the *ATXN10* gene, RepeatHMM called one more repeat for allele 2 than the PCR results, while allele 1 was accurately genotyped. Similarly, we observed mismatches in some repeats for the *HTT*, *ATXN1*, and *ATXN2* genes.

For completeness, we include the results from *FMRI* in Figure 5-3; for this gene there is a clear difference in the number of repeats called by RepeatHMM and PCR genotyping. This discrepancy reinforces our previous assertion that the PCR results from this gene may not be very reliable. Considering these results, the decision to withhold the use of PCR genotyping results until further validation seems justified. This underscores the importance of careful validation and cross-checking of results from different genotyping methods.

Figure 5-4 provides a more detailed view of the information used by RepeatHMM to make statistically accurate genotype calls, including coverage, depth, allele length, and PCR and RepeatHMM results for four patients and one control across 10 genes. The results demonstrate that even with high-coverage data, the lengths of alleles can vary within individual sequence reads, potentially due to the high error rate of Nanopore sequencing. Additionally, RepeatHMM sometimes had very little sequence data available to make its calls. These observations help explain the discordance seen in Figure 5-3, which may be attributed to low sequence coverage and the Nanopore error rate.

For example, sample 1675 in Figure 5-4 had high coverage for the *HTT* gene, but RepeatHMM called it as a homozygote for the repeat instead of two alleles as detected by PCR. For *ATN1*, RepeatHMM produced a reasonably accurate call, but the results were skewed to the right in comparison to the PCR results, as seen in Figure 5-3. For *FMRI*, although there was some coverage for RepeatHMM around 20-22 repeats, the aligner did not make any calls for a second allele. Conversely, for *ATXN2*, where the coverage was low, RepeatHMM made calls at 21 and 24 repeats despite the limited data. Similar results were observed for other genes and samples. Mismatches in repeat length were common, indicating the need for additional calibration in the aligner prior to data processing.

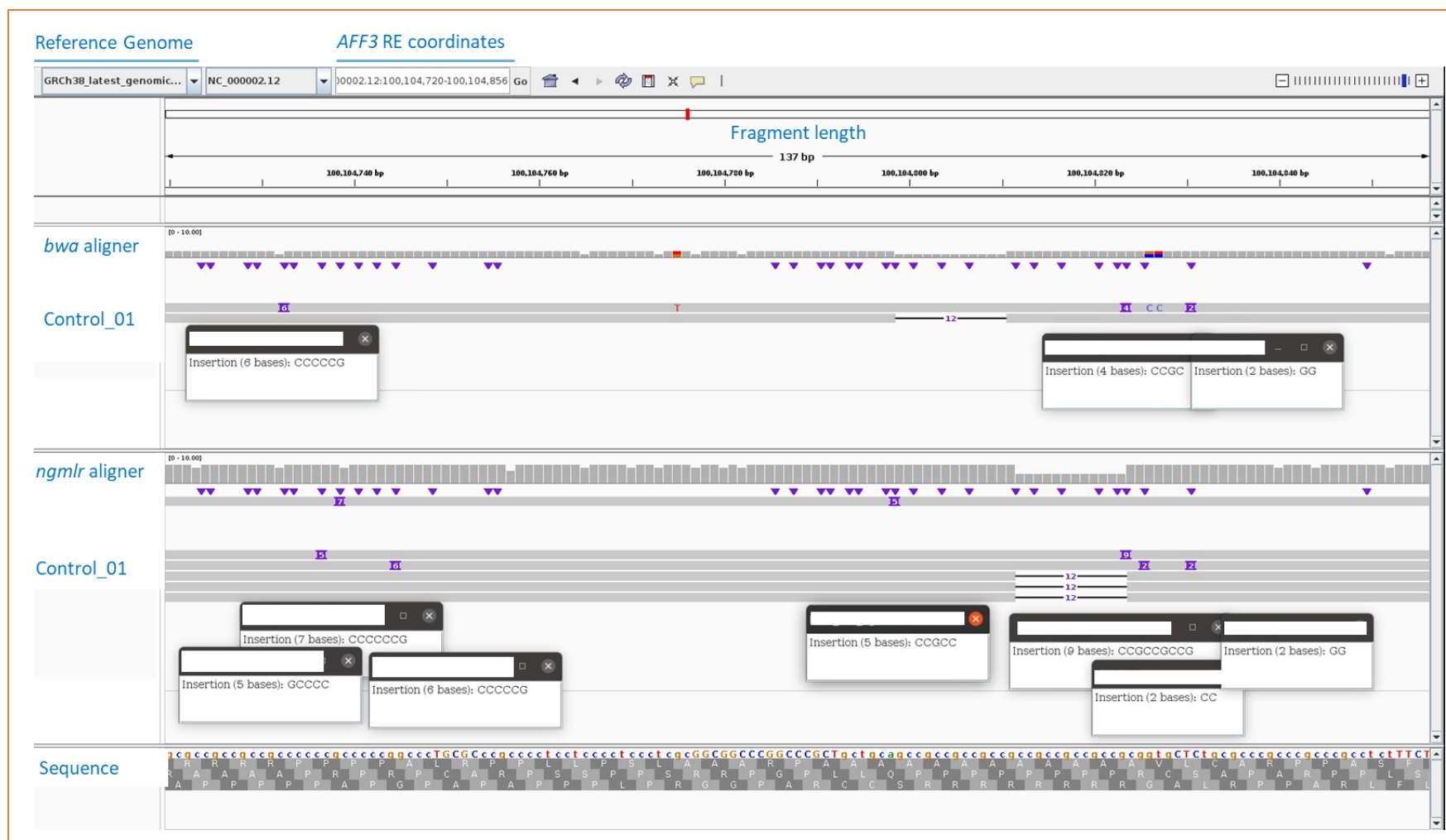


Figure 5-2 IGV visualization of bwa and ngmlr alignments for the *AFF3* gene in control 1.

IGV screenshot comparing bwa (on top) and ngmlr (bottom) aligner using the same reads from control 1. The sequencing alignment of the GCC repeat region of the *AFF3* gene. Top panel shows the bwa alignment, in which the software only covered 3X coverage for the repeat expansion. One of the alleles had a 12 bp deletion at the beginning of the RE and 4 and 2 bp insertions. In contrast, ngmlr software aligned 6 reads to the repeat region and is possible to distinguish the two alleles. Allele one has 12 bp deletion located at the end of the RE sequence, while allele 2 seems to have 9 bp insertion for one read, and scattered insertions in the other reads. Reference Genome: HG38.

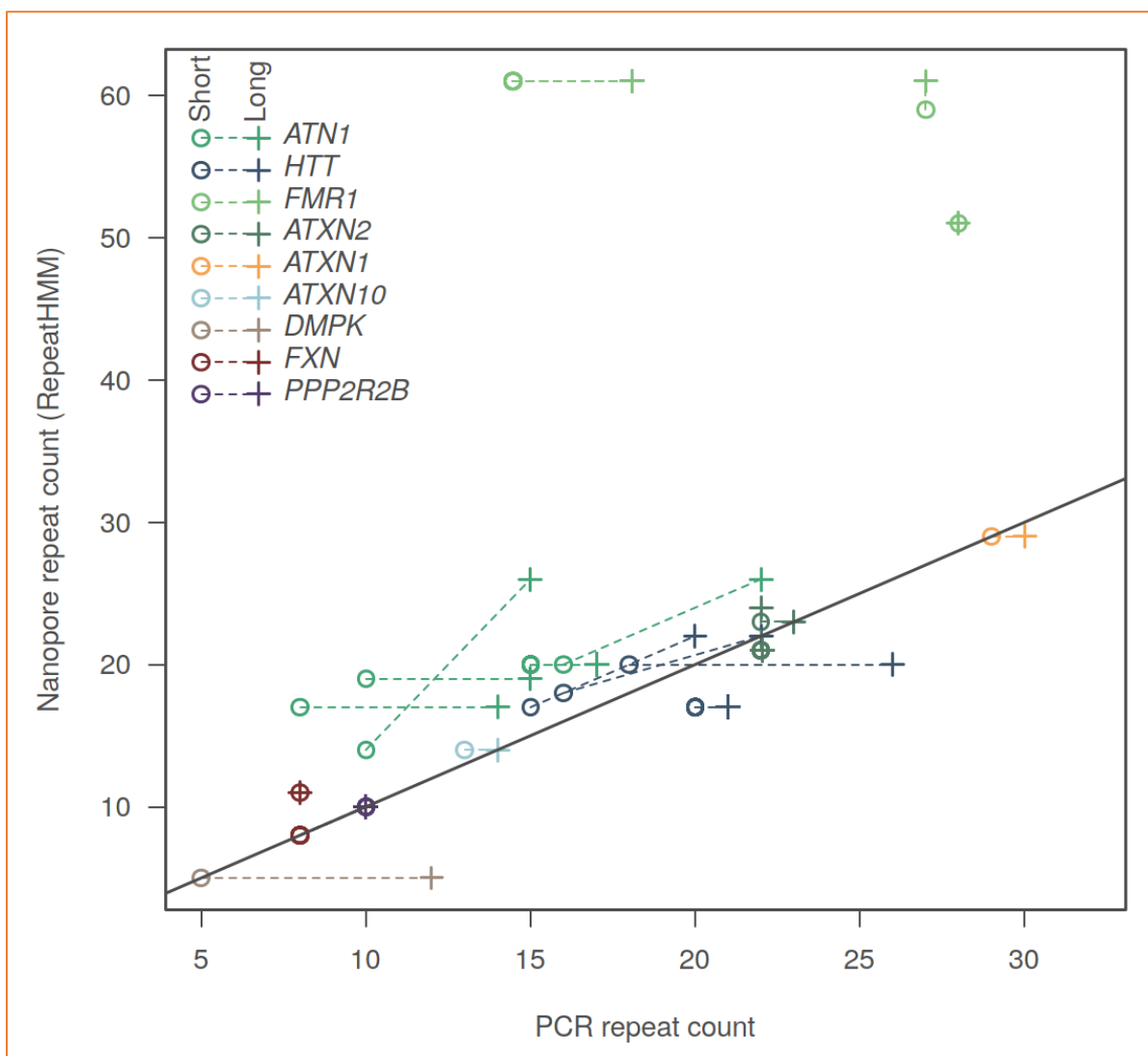


Figure 5-3 Graphic representation of the genotype results from RepeatHMM compared to PCR results.

Plot shows how well RepeatHMM results match PCR genotypes obtained in chapter 3. The solid line represents the expected 1:1 ratio in repeat count if all alleles matched between both genotyping methods. This plot contains results from 9 different genes indicated with an assigned colour. O: short allele; +: long allele; ---: connects both alleles of the same sample. Excluding *FMR1* gene, RepeatHMM results are significantly correlated with PCR results ($R^2=0.73$, p -value: 1.07×10^{-31}).

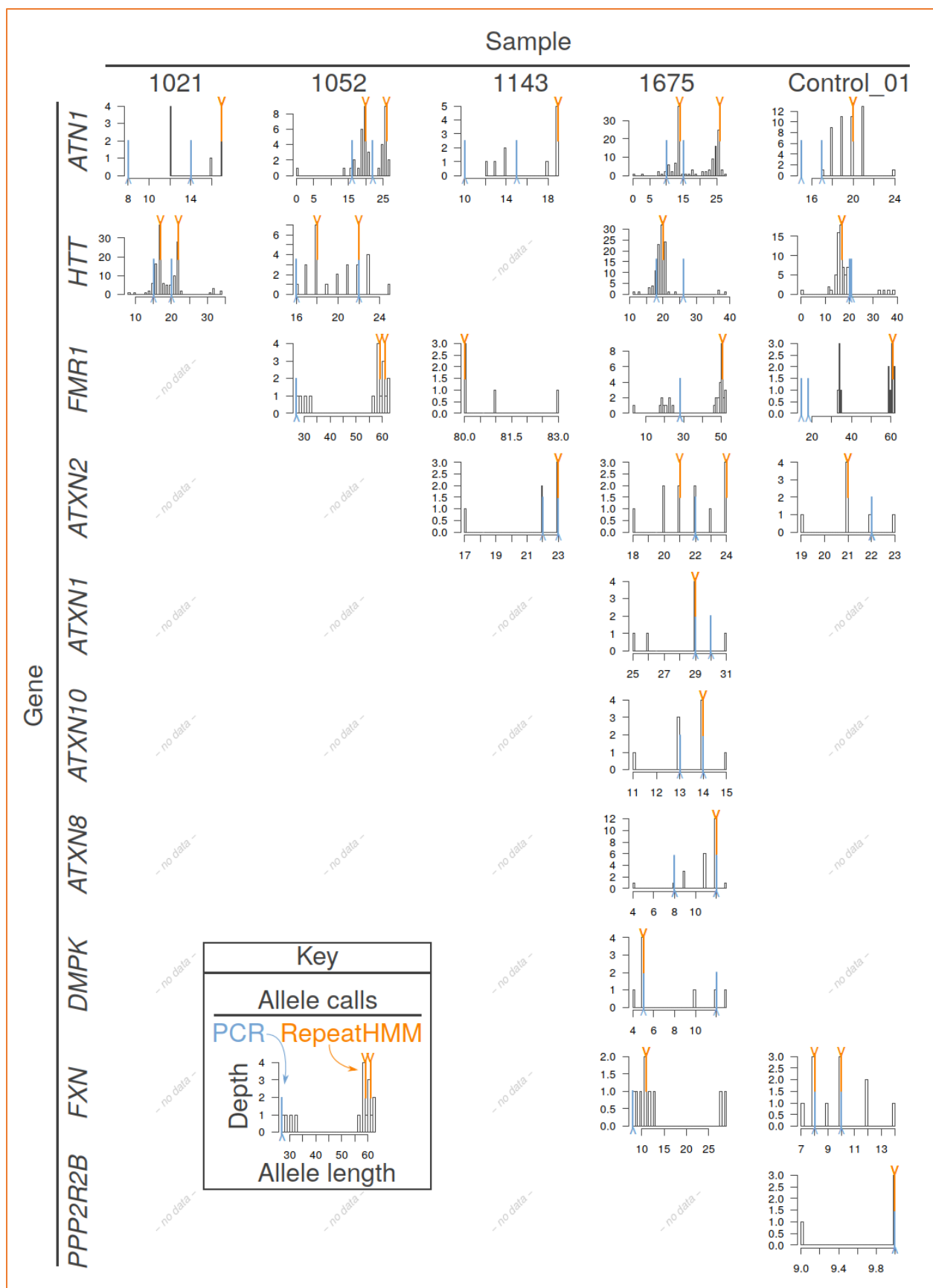


Figure 5-4 Distribution of repeat counts within individual reads supporting RepeatHMM genotypes.

Individual plots are shown for each gene sequenced in each individual, and represent the number of reads that contained alleles of various repeat counts, which were used in RepeatHMM's Gaussian mixture model to call diploid genotypes. For each plot, the final genotype called by RepeatHMM is shown in orange, and for comparison, the corresponding PCR-based genotype is shown in blue. Even with high-coverage data, there is sometimes considerable variance around the apparently "true" allele (eg for *HTT* in sample 1675).

Sniffles2 and SquiggleKit

To explore the potential of Nanopore sequencing for genotyping the long *C9orf72* repeat expansion, we employed Sniffles2 to detect structural variants in the Nanopore data. The results indicate that only one sample was able to yield a *C9orf72* repeat expansion (RE) genotype due to sufficient depth of coverage. Notably, this sample (designated as 1675) was negative by rpPCR (Figure 5-5).

The Sniffles RE genotype for sample 1675 was homozygous for a large 15,845 basepair "insertion" consisting of approximately one hundred instances of GGCCCC, followed by a region enriched for the repeat motif AGCACC, and then another hundred or so GGCCCC repeats, interspersed with non-GGCCCC sequence (Figure 5-6). Intriguingly, the start position of the genotype was located 180 base pairs telomeric from the expected position of the RE.

Based on these findings, it was unclear whether this genotype represented a true alternative genetic variant contributing to *C9orf72*-mediated ALS, or the established *C9orf72* hexanucleotide repeat expansion. For example, the complex allele shown in Figure 5-6 could indeed be the true sequence of a large insertion/expansion mutation, but the mutagenic mechanism for such a variant is not parsimonious. Alternatively, the sequence composition of this allele could derive from basecalling inaccuracy with Guppy's recurrent neural network, followed by imprecise alignment and variant calling that is unable to correctly genotype the expansion. We therefore sought to further investigate the nature of sequence reads contributing to this allele.

Upon examining the Nanopore data of *C9orf72* in sample 1675 using IGV, it appears that misalignment over low-complexity sequence around the RE locus, combined with the way the aligner treated the data, resulted in an inaccurate genotype. Specifically, a read with a length of 39,938bp contained some misaligned repeat sequence and had a very long CIGAR string with

13,476 bases soft-clipped at the beginning of the alignment (ie. 13,476 bases not considered by the aligner but left in the aligned sequence).

To further explore the nature of this sequence read, extracted raw picoamp current signal from the fast5-format Nanopore data using SquiggleKit. We then plotted the extracted signal for this read in Figure 5-7. Notably, the expected current signal for normal genome sequences does not follow any discernible pattern, whereas GGCCCC repeating sequences should follow a very tight pattern with regular periodicity. Visualising the data this way, we observed that this sequence read did indeed include a long stretch of signal with roughly 6 basepair periodicity, and this repetitive region began within the ngmlr soft-clipped region. Measuring the time elapsed between the start and end of this repetitive region allowed us to approximately estimate the number of hexanucleotide repeats (assuming a sequencing speed of 450 bases per second) as 1,908 repeats.

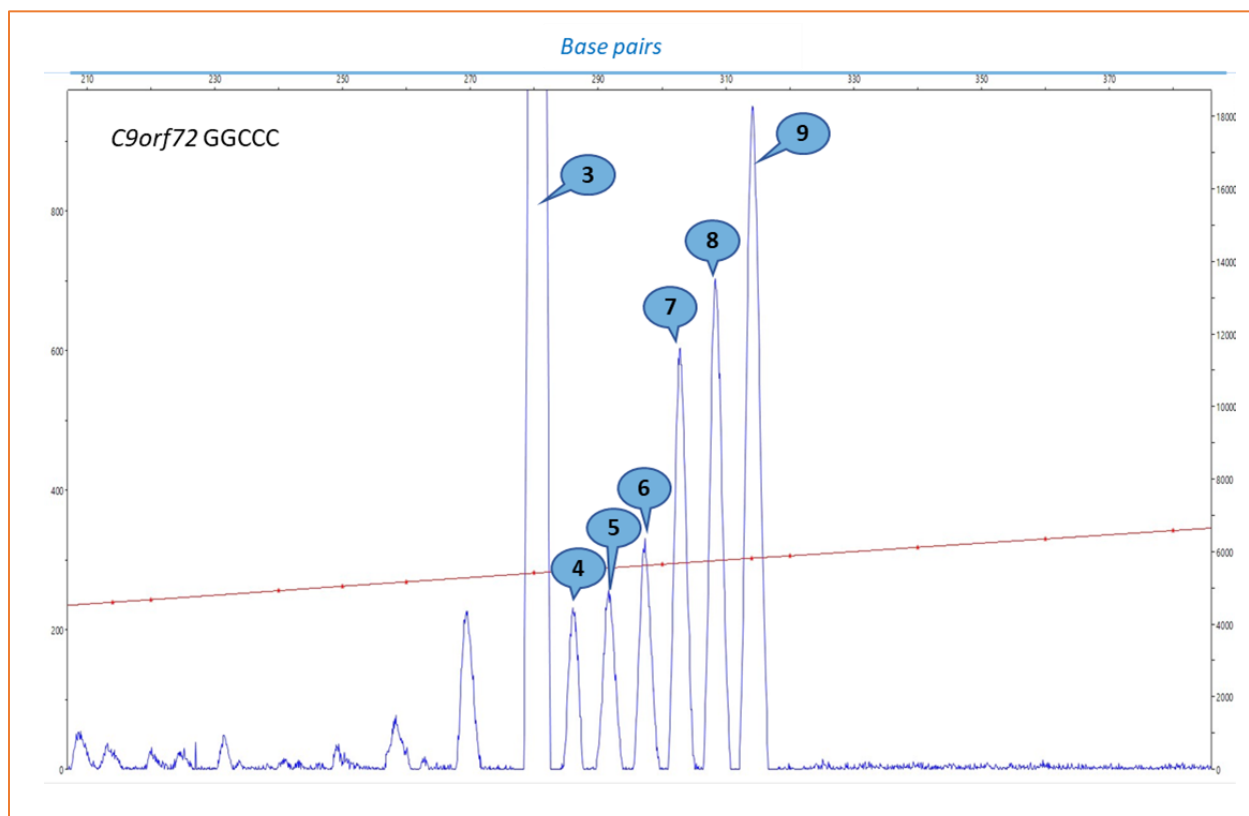


Figure 5-5 Capillary electropherogram of the *C9orf72* rpPCR for sample 1675.

The panel shows the smooth peaks pattern obtained by the rpPCR of the *C9orf72* gene starting around 275-280 bp. The first peak has 3 GGCCCC repeats, increasing by 1 repeat every following peak (marked by the blue call outs). Sample 16 has 9 hexanucleotide repeats. We used the GeneScan™ 1200LIZ™ dye size standard (not shown in panel, but based on it, on top of the panel are the base pairs).

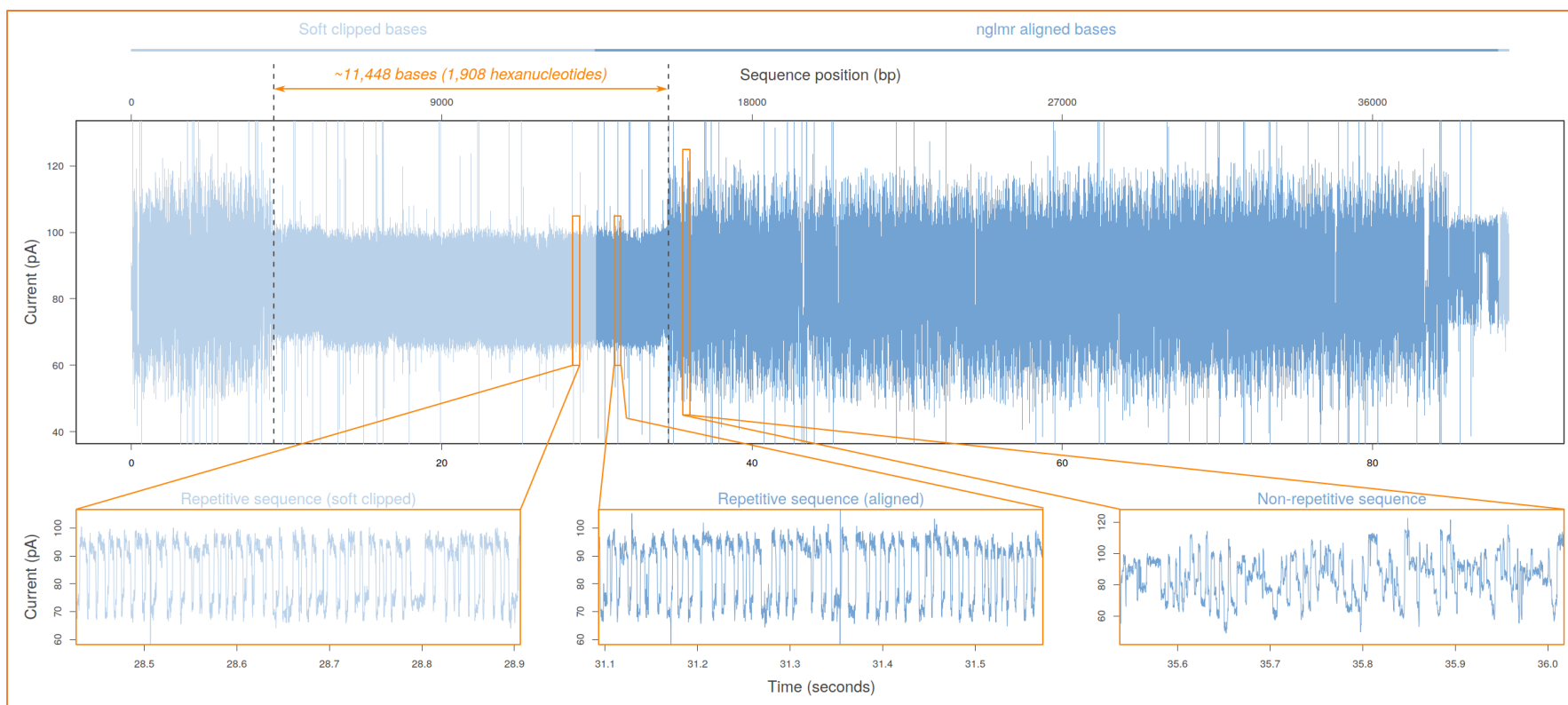


Figure 5-7 Raw current signal from sample 1675 on the *C9orf72* repeat expansion region.

Top plot shows the raw current signal in pA from the sample 1675, read number 9da7c231-00e3-4b4e-8b21-b092c07 at the *C9orf72* repeat expansion region. Darker blue signal corresponds to the sequence aligned by ngmlr software; lighter blue signal corresponds to the soft-clipped bases, not aligned, or used by ngmlr. Bottom panel contains 3 expanded figures from the raw current on the top, which shows how the current varies in a given time and its behaviour during sequencing. This shows that the soft-clipped bases have the same repetitive pattern as the aligned repetitive sequenced. When calculating the number of repeats including the soft-clipped, sample 1675 has 1,908 GGCCC repeats in the *C9orf72* gene.

5.4 Discussion

Our study provides an insight on the potential of long read sequencing and non-PCR Cas9-enrichment protocol by ONT for genotyping repeat expansions. Our findings indicate that this protocol is effective for sequencing long reads, despite facing some challenges in designing crRNAs for certain genes due to their locations and low yield. We also encountered issues with resuspending the pooled crRNAs, which formed a jelly-like solution that required dilution. To improve future experiments, we suggest splitting the crRNA pool and exploring new probes for genes with low or no coverage. Our study highlights the importance of optimizing protocols for repeat expansion genotyping and suggests avenues for future research in this area.

RepeatHMM

To assess the accuracy of RepeatHMM in calling the number of repeats from Guppy-basecalled data, we analysed the results obtained by RepeatHMM and compared them with PCR genotyping for the genes with sufficient sequence data to support RepeatHMM genotypes. Our findings showed reasonable agreement between the two methods, except for the *FMRI* gene ($R^2=0.73$, p -value: 1.07×10^{-31}). We observed that the variance in allele calls by RepeatHMM was well explained by PCR alleles, but there were still some discrepancies between the two methods. We found that RepeatHMM accurately called the genotypes for allele 1 and 2 in the *PPP2R2B* gene, but there were mismatches in some repeats for the *HTT*, *ATXN1*, *ATXN2*, and *ATXN10* genes. Notably, there was a clear difference in the number of repeats called by RepeatHMM and PCR genotyping for the *FMRI* gene, suggesting that PCR genotyping may not be very reliable for this gene. For the other genes, it is likely that discrepancies between Nanopore and PCR genotypes are at least partially a function of low sequence coverage in some cases (as detailed in figure 5-4); however, for some genes, despite high sequence coverage, substantial variance for repeat length inferred within individual reads still remained. Further improvements in long-read sequencing analysis pipelines and statistical models supporting RE genotypes may lead to better genotypes at RE loci; this should be an area of future work. In summary, our results suggest that RepeatHMM is generally effective in calling the number of repeats, with reasonable concordance with PCR genotyping results. However, some discrepancies were observed, indicating that caution should be exercised when interpreting

results obtained using RepeatHMM, and future developments in RE genotyping approaches with long-read sequence data may yield improvements in accuracy for these loci.

Sniffles2 and SquiggleKit

Our study demonstrates the potential of Nanopore sequencing for genotyping *C9orf72* repeat expansions, but also highlights the challenges associated with accurately genotyping such expansions. Our findings show that relying solely on alignment-based methods based on hard base calls for genotyping *C9orf72* repeat expansions may lead to inaccuracies and errors, particularly in low-complexity regions. The use of alternative methods, such as the analysis of raw signal data, may provide a more accurate and reliable means of genotyping *C9orf72* repeat expansions.

The discovery of an expanded genotype for sample 1675, containing a large insertion of GGCCCC repeats in *C9orf72*, located 180 base pairs from the expected position of the RE, raises important questions about the alignment of this repeat expansion and optimal strategies for managing long-read sequence data at hyper-expanded loci such as *C9orf72*. The high number of repeats (around 1,908 hexanucleotide repeats) observed in this allele is intriguing and warrants further investigation; however, it is not clear whether this allele represents the canonical *C9orf72* hexanucleotide repeat expansion that causes around 10% of ALS/FTD cases, or a novel allele that is proximal to this locus but which has a different overall sequence and location. The negative rpPCR result (figure 5-5) for this sample suggests that this could be a real possibility; however, the sequence composition of the apparent expanded allele (Figure 5-6) is not consistent with straightforward established mechanisms of repeat expansion such as polymerase slippage.

Considering this, along with the treatment of sequence reads supporting this genotype (soft clipping spanning the start location of the expansion), it is more likely that the expansion is the conventional G₄C₂ allele whose role in ALS is already known. Our ability to directly estimate the repeat expansion length showcases the great potential of long-read sequencing to enable highly detailed analysis of the role that the variable length of the *C9orf72* RE plays in the sub-phenotypes of ALS and FTD. This leaves an open question around the negative rpPCR result,

however. It is possible that this discrepancy could derive from laboratory error, rpPCR insensitivity or sequence variation in primer binding sites; the latter possibility is particularly appealing given other observations of *C9orf72* discordance within pedigrees (Ryan M, Hardiman O, pers. comm.) and warrants further investigation.

A study by Giesselmann in 2019 investigated the accuracy of genotyping *C9orf72* repeat expansions using long read sequencing from ONT. Their findings revealed a systematic sequence strand bias that resulted in more accurate counts for the GGGGCC sequence compared to the complementary strand (GGCCCC) when using RepeatHMM. They concluded that while neural network base-callers can improve single-read base quality at the genomic level, they become unreliable for repeat expansions greater than 32 G₄C₂ repeats (Giesselmann et al. 2019). To address this issue, they developed the tool STRique, which employs a signal alignment algorithm to map the upstream and downstream boundaries of each repeat more precisely and quantifies the number of repeats of any given STR sequence with a hidden Markov model (HMM) (Giesselmann et al. 2019; Schreiber and Karplus 2015). These results highlight the challenges of accurately genotyping repeat expansions using Nanopore sequencing and emphasize the need for improved tools and methodologies to overcome these limitations. STRique presents a promising approach to improving the accuracy of repeat expansion genotyping and should be considered in future studies (Giesselmann et al. 2019).

Our results highlight the potential of Nanopore sequencing for detecting structural variants and identifying long-range rearrangements and expansions in the human genome. The ability to obtain long read lengths with Nanopore sequencing offers a significant advantage over other sequencing technologies and may prove invaluable in identifying complex genomic rearrangements, particularly in regions of the genome with high sequence repeats.

5.5 Conclusions and future directions

In conclusion, our study highlights the challenges of using Nanopore sequencing to genotype repeat expansions, with accuracy potentially compromised by high variability and errors associated with long-read sequencing. However, our findings also indicate that Nanopore sequencing can be a valuable tool for the discovery of repeat expansions due to its long read lengths. To maximize the accuracy of repeat expansion genotyping, future efforts should focus on generating high-coverage data where possible. While the use of RepeatHMM can improve accuracy, its limitations should be carefully considered, especially for genes with high variability or low sequence coverage. Our study suggests that Nanopore experiments may be better suited for discovery of repeat expansions rather than accurate genotyping. Analysing the Nanopore data using the raw signal, using software like SquiggleKit, is a valuable tool available to someone with interest in repeat expansions. Overall, our results highlight the importance of evaluating the accuracy of repeat expansion genotyping methods and suggest that combining multiple approaches may be necessary for accurate and comprehensive characterization of repeat expansions. Furthermore, future studies should continue to explore new approaches and strategies to overcome the limitations of current methods and further our understanding of the genetic basis of diseases caused by repeat expansions.

Chapter 6

Discussion, limitations & future direction

The overarching aim of this thesis was to enhance our comprehension of the genetic mechanisms underlying ALS. Specifically, we aimed to investigate the existence of pleiotropy in genetic contributors to ALS aetiology, as evidenced by multiple repeat expansions previously linked to other neurodegenerative diseases. By advancing our knowledge of ALS risk genes, we can pave the way for further research into disease mechanisms and risk factors, ultimately leading to improved patient diagnosis and treatment options. Through these efforts, we hope to make significant strides in enhancing patient outcomes and mitigating the impact of these devastating conditions.

This thesis describes the development of a highly multiplexed genotyping panel for NDREs that is both cost-effective and efficient. Through these optimization efforts, we were able to generate a final, highly optimized genotyping panel that allowed us to genotype 35 NDREs in a cohort of 1500 ALS patients and 1700 neurologically normal controls from the Irish ALS DNA bank, as described in Chapter 3.

The use of PCR, the gold standard protocol, ensured the accuracy and precision of the resulting data, which will serve as a robust baseline for future studies. Overall, the development of this genotyping panel represents a significant contribution to the field of neurodegenerative disease research, as it will allow for more accurate and efficient identification of NDREs in patient populations. This, in turn, will help to improve our understanding of the genetic causes of ALS and related diseases, and pave the way for the development of more targeted treatment approaches.

The hypothesis of this thesis is based on the general pleiotropy of genetic risk factors for ALS and other neurological conditions, as there are overlapping phenotypes with multiple neurodegenerative diseases. This idea is supported by the fact that ALS shares many clinical and pathological features with other neurodegenerative diseases, such as Alzheimer's and Parkinson's disease, and that established ALS risk genes such as *C9orf72*, *ATXN2* and *NIPA1* also have roles in other neurological diseases. The aim of this study was to identify genetic risk factors for ALS and to explore the potential that disease mechanisms are convergent between ALS and other neurodegenerative diseases – namely, the pathophysiological effects of expression of potentially multiple repeat expansions.

The results of this study are important for future research in multiple fields. This thesis showed 3 novel genes for ALS risk and an increase in ALS risk when harbouring repeat expansions in 5 or more genes at the same time. Additionally, an increase in ALS risk was observed with cognitive/behavioural site of onset when harbouring intermediate repeats in the *JPH3* gene, which is a phenotype for HDL2. These findings are not only compelling as they indicate an intuitive endophenotype of ALS conferred by a risk allele normally associated with a neurocognitive disease, but they also open many doors and raise many other questions in the field of ALS research.

However, it is important to note that this project should be replicated in a larger cohort of patients and controls. Although the cohort used in this study was relatively large by the standards of similar work in other studies (Figley et al. 2014), the statistical power is still likely insufficient to detect all potential genetic risk factors for ALS. Increasing the number of cases and controls in a different replication cohort will confirm the results obtained in this thesis and only increase the statistical power, which will lead to more accurate and reliable results.

Moreover, this study serves as a pivotal gateway to a multitude of unexplored research domains, warranting comprehensive investigations not only within the confines of this intricate system model but also delving into the intricacies of iPSC-derived cells. This innovative avenue offers the prospect of validating the pleiotropic effects induced by diverse RNA and protein encoding intermediate repeat expansions. By deciphering these intricate interplays, we can unravel the

enigmatic mechanisms that potentially contribute to the augmentation of ALS susceptibility associated with specific gene combinations.

Nevertheless, the scope of this study is not confined solely to ALS; rather, its implications resonate across the spectrum of neurodegenerative disorders, igniting the pursuit of an enhanced comprehension of the multifarious pathways through which repeat expansions underpin a myriad of brain-related diseases. Huntington's disease, Spinocerebellar ataxias, and Parkinson's disease stand as emblematic examples. The envisaged expansion of this gene panel to encompass a broader repertoire, embracing various diseases, has the potential to germinate an extensive database. From this, a refined understanding of repeat expansion genes, responsible for augmenting susceptibility to diverse disorders, can be distilled and replicated within cell models to discern the underlying mechanisms unique to each ailment.

Ultimately, this versatile and comprehensive panel, coupled with the ensuing database, lays the groundwork for early diagnosis paradigms. Through the analysis of an individual's DNA in tandem with the obtained outcomes, the inherent risk for a plethora of diseases can be inferred. By merely quantifying the repeat expansion length per gene, a proactive understanding of disease susceptibility emerges, fostering personalized preventive measures and interventions. In essence, this study's overarching potential resides in its capacity to transcend ALS research, morphing into an instrumental tool for revolutionizing disease diagnosis and paving the way for targeted therapeutic strategies.

One of the most important reasons for researching the cause of the disease is to find a treatment. In recent years, antisense oligonucleotides (ASOs) have revealed themselves as an increasingly promising modality. These are short, synthetic, single-stranded oligonucleotides that can alter RNA expression and therefore reduce, restore, or modify protein expression (Aoki and Wood 2021; Talbot and Wood 2019). The type of ASO used in this research prevents gene expression by binding to messenger RNA (mRNA) strands. Once this binding takes place, this hybrid sequence is targeted and naturally degraded by enzymes in the cell. The appeal of developing an ASO to treat ALS and other neurodegenerative diseases is its simplicity (Khamaysa and Pradat 2022). ASOs are essentially anti-messenger RNA agents. Using genetically targeted ASOs, it is possible to design an antisense oligonucleotide sequence that binds to that mRNA

so that its expression is lowered, and consequently the mutant protein never gets made. After establishing how to deliver an ASO to a certain type of cells, theoretically, it should be possible to repeat for other neurodegenerative diseases (Aoki and Wood 2021; Khamaysa and Pradat 2022; Talbot and Wood 2019).

For ALS, there have been some drug trials for the *C9orf72* RE. In one study, an ASO injected into the spinal canal led to a significant reduction of ALS-related neurotoxins known as dipeptide repeat proteins (DPRs) in the trial subject's spinal fluid (Tran et al. 2021). During the course of the trial, the subject's ALS functional score rating and other measures of impact were largely stable or slightly improved. The patient, who had been experiencing weakness in the legs and feet before treatment, had no neurological or medically adverse effects from the treatment (Tran et al. 2021).

However, results from ASO trials are not universally positive. BIIB078 is an ASO designed to target the mRNA of the *C9orf72* gene that began its Phase 1 in adults in 2018. This ASO functions by breaking down mRNAs containing the hexanucleotide repeats. The *C9orf72* RNA and the dipeptide proteins generated by the RE are considered harmful to motor neurons, hence the idea that reducing their levels would be beneficial. However, the outcomes of the 2022 clinical trial did not demonstrate any clinical benefit. The study revealed that the disease's mechanisms are much more intricate than solely focusing on the toxicity of *C9orf72* associated with the RE (Anon n.d.-c, Anon n.d.-b). It is therefore of paramount importance to continue basic biomedical research into the mutations underlying ALS risk, so that targeted treatments may continue to be improved towards an effective cure for this devastating disease.

An alternative strategy to address the detrimental gain of function associated with RNA/protein interactions in SOD1 ALS pathology involves the utilization of RNA interference (RNAi) tactics. Unlike antisense ASOs, which are single-stranded and readily engage their targets, RNAi employs double-stranded RNAs that require enzymatic processing prior to exerting their effects. Within this RNAi mechanism, RNA molecules are processed through the RNA-induced silencing complex (RISC) in the cytoplasm, leading to the degradation of targeted mRNA molecules, ultimately suppressing the expression of specific genes (Castanotto and Rossi 2009). Common RNAi techniques encompass short interfering RNAs (siRNAs), short hairpin RNAs

(shRNAs), and artificial mitochondrial RNAs (miRNAs). Delivery of these RNA molecules into CNS neurons can be facilitated through adeno-associated viral vectors (AAV). In experimental models featuring SOD1G93A mice, the introduction of siRNAs via AAV-mediated delivery yielded a noteworthy 39% extension in survival, albeit with varying effectiveness linked to the disease stage due to the already ongoing SOD1 dysfunction (Foust et al. 2013; Stoica et al. 2016; Thomsen et al. 2014).

Encouraging outcomes in animal studies prompted the initiation of clinical trials involving two familial ALS patients. Intrathecal infusion of a singular AAV-miR-SOD1 treatment was administered. The first patient exhibited transient lower limb strength improvement alongside side effects of meningoradiculitis. Conversely, the second patient, pre-treated with immunosuppression, experienced no side effects. Although autopsy of the first patient indicated reduced SOD1 levels, cerebrospinal fluid analysis did not reflect SOD1 reduction in either patient after two weeks. The second patient demonstrated consistent scores on an ALS functional measure and maintained vital capacity over a 12-month period. This underscores the translation of promising animal-based RNAi strategies to human trials, albeit with complex outcomes requiring further exploration (Mueller et al. 2020).

An additional gene therapy strategy for addressing ALS involves the utilization of neurotrophic factor which have garnered significant attention in recent years. Neurotrophins are signalling molecules governing neuronal function and influencing apoptosis rates. The majority of investigations into neurotrophins have centred around insulin growth factor (IGF) and vascular endothelial growth factor (VEGF) (Huang and Reichardt 2001). The application of self-complementary adeno-associated virus serotype 9 (scAAV9) encoding IGF-1 in SOD1 mice demonstrated compelling results, notably curbing motor neuron degeneration in the spinal cord's anterior horns and significantly decelerating disease onset and progression (Lin et al. 2018). Similarly, studies injecting AAV9 carrying IGF-2 into SOD1-G93A mice exhibited a noteworthy 10% increase in lifespan, pointing to its potential as a protective agent for neuronal survival (Allodi et al. 2016). The administration of scAAV9-VEGF-165 to SOD1-G93A mice underscored enhanced survival and motor strength extension (Wang et al. 2016). Intriguingly, simultaneous administration of IGF-1 and VEGF failed to exhibit additive benefits, suggesting potential overlapping mechanisms (Dodge et al. 2010)

Last but not least, is Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-Associated Protein (CRISPR/cas) strategy. CRISPR/cas, initially studied in bacteria, is gradually gaining traction in neurodegenerative research, focusing on rectifying mutant DNA to interrupt abnormal downstream pathways. This strategy holds the prospect of serving as a one-time intervention. Notably, limited studies have ventured into ALS using this method. A study in 2017 employed CRISPR to target the SOD1 gene via neonatal SOD1G93A mice, utilizing modified AAV9 to deliver *Staphylococcus aureus*-derived Cas9 (SaCas9) and a single-guide RNA (sgRNA). This intervention resulted in decreased SOD1 expression in spinal cords, an upsurge in motor neurons, delayed disease onset, and extended survival (Gaj et al. 2017). Subsequent 2020 studies echoed these findings, showcasing reduced SOD1 expression in the spinal cord and enhanced survival rates (Duan et al. 2020; Lim et al. 2020). An inherent limitation of these studies lies in their administration to young mice prior to ALS symptom manifestation. The efficacy of such treatment in older mice already demonstrating ALS-related symptoms remains uncertain due to the revised El Escorial criteria, requiring symptoms in at least one anatomical region for ALS diagnosis (Brooks et al. 2000).

The future of ALS genetics research holds great promise, as it continues to delve deeper into uncovering new REs, SVs, SNPs, and other genetic factors associated with the disease. Here are some key directions for advancing our understanding of the genetics of ALS:

1. Identification of New REs and SVs: Researchers are likely to employ advanced sequencing technologies, including long-read sequencing, to identify novel repeat expansions and structural variants that might have been previously missed. As these technologies become more accessible and affordable, they will play a crucial role in uncovering rare and elusive genetic factors contributing to ALS.
2. Functional Characterization: It's essential to investigate how structural variants, including those interrupting Topologically Associating Domains (TADs), affect gene expression and regulation in ALS. Functional studies can help elucidate the impact of these variants on cellular processes and disease pathogenesis.
3. Non-Coding Variants: While REs have gained significant attention, the role of non-coding variants in ALS should not be underestimated. Comprehensive genome-wide association studies (GWAS) can help identify SNPs and other non-

coding variants associated with ALS susceptibility, offering insights into the regulatory elements and pathways involved.

4. Splice Variants and Dysregulated Splicing: Dysregulated splicing has emerged as a crucial aspect of ALS pathogenesis. Researchers can employ RNA sequencing techniques to identify and characterize splice variants implicated in ALS. Understanding how these variants affect gene expression and protein function is vital for dissecting the disease mechanisms.
5. Epigenetic Changes: Investigating epigenetic alterations in target tissues, such as the motor neurons, is a burgeoning area of ALS research. Epigenetic studies can uncover DNA methylation, histone modifications, and non-coding RNA regulation that influence gene expression and may contribute to disease progression.
6. Multi-Omics Approaches: Integrating data from genomics, transcriptomics, proteomics, and epigenomics will provide a more comprehensive view of ALS pathogenesis. Multi-omics approaches can identify interconnected networks of genes and pathways that play pivotal roles in the disease.
7. Patient Stratification: As genetic heterogeneity within ALS becomes increasingly apparent, personalized medicine approaches may emerge. Stratifying patients based on their genetic profiles could lead to more tailored therapeutic interventions.
8. International Collaborations and Data Sharing: The genetics of ALS is a complex and multifaceted field, and international collaborations and data sharing initiatives are critical for pooling resources, standardizing methodologies, and accelerating research progress.

Overall, the future of ALS genetics research is likely to be characterized by a multidisciplinary approach, incorporating advanced sequencing technologies, functional genomics, epigenetics, and collaborative efforts to unravel the complex genetic landscape of the disease. This collective effort will not only enhance our understanding of ALS but also pave the way for the development of more effective diagnostic tools and therapeutic strategies.

In summary, this thesis has identified several genetic risk factors for ALS, including novel genes and repeat expansions. These findings shed light on the complex and heterogeneous nature of ALS, as well as the potential for personalized medicine approaches. Furthermore, this thesis has highlighted the promising role of REs in the pleiotropy of ALS and possibly other neurodegenerative diseases. Overall, this study has made significant contributions to our understanding of ALS and has opened up new avenues for research and potential treatment for neurodegenerative diseases.