Identification and characterisation of genetic variation that modifies age at onset in Huntington's disease

by

William Branduff McAllister

A thesis submitted for the degree of

Doctor of Philosophy

CARDIFF UNIVERSITY

PRIFYSGOL CAERDYDD

September 2019

Summary

Huntington's disease (HD) is a progressive and ultimately fatal neurodegeneration caused by a CAG repeat expansion in the huntingtin gene (*HTT*). The length of the CAG repeat is strongly inversely correlated with disease onset, but there remains considerable onset variation that is unexplained, even between individuals with the same repeat length. This thesis details the investigation of genetic modifiers of HD onset using next-generation sequencing (NGS), with an emphasis on rare coding variant identification.

Chapter 1 gives a general introduction. **Chapter 2** details the experimental procedures used in this study. **Chapter 3** explores several HD onset phenotypes in the Registry-HD study, derived using the clinician's estimate of onset and the HD clinical characteristics questionnaire. An extreme HD onset cohort (N=500) is then selected using residual age at motor onset.

Chapter 4 uses whole-exome sequencing to sequence the 500 HD patients selected in the previous chapter. I identified rare damaging variation in several DNA repair genes associated with altered disease onset, including *FAN1*, *EXO1*, *MSH3, LIG1* and *PMS1*. Unbiased whole-exome burden and SKAT(-O) analyses identified *NOP14* as an exome-wide significant gene. Investigation revealed *NOP14* strongly tagged *HTT* allele structure, identifying it as a major modifier of disease onset.

Chapter 5 confirms the *HTT* allele structures using an independent NGS method. *HTT* alleles possessing additional interruptions were associated with late disease onset; whereas alleles lacking interruptions were uniformly found in early onset individuals. I also detail three novel *HTT* alleles associated with extremely delayed onset and explore MiSeq-based instability measurements.

In **Chapter 6**, I discuss the study generally. I give an overall model describing genetic factors that modify HD onset, and present my two-fated pathway model for somatic instability. I then highlight future studies and approaches in HD given what we have learned from this work. The results presented here have important implications for understanding the mechanisms underlying HD onset, underscoring DNA and DNA repair as critical components in HD.

Lay summary

Huntington’s disease (HD) is a destructive and untreatable brain disease that affects one in 10,000 in the UK. HD is usually inherited from an affected parent and has a 50% chance of being passed on to children. The later stages of the disease require full-time nursing care, and HD is usually fatal 15-20 years after symptoms start. We know what causes HD – an error in the huntingtin gene. We all have two copies of this gene (one from mum and one from dad), and it’s important for our health. Our genes are made up of four DNA chemicals called adenine, cytosine, guanine and thymine, and these are abbreviated as the letters A, C, G and T. In HD, three DNA letters in the huntingtin gene, C-A-G, are repeated too many times, like a genetic stutter, making the gene faulty. The number of these CAG letters is important, as people with more CAGs on average develop HD earlier (an earlier ‘onset’). But there is still enormous variability between when people start experiencing HD symptoms that we cannot explain. For instance, two people with the same number of CAGs could have HD symptoms years or even decades apart from one another. The work detailed in this thesis explores how a person’s genetic makeup (the small genetic differences that all make us unique) can affect when HD onset occurs.

To begin, I give a background to the current HD research field and introduce important concepts to consider in the thesis (**chapter 1**). Following this, I set out the experimental procedures used (**chapter 2**). In my first results chapter, I explore the symptoms experienced by HD patients (**chapter 3**). HD patients can experience a wide range of different symptoms including uncontrollable muscle movements (chorea), depression, cognitive decline and psychosis (hallucinations/delusions). Using clinical data, I select a group of 500 HD patients (a ‘cohort’) that have early or late onset of motor symptoms given their CAG length, and these individuals are then analysed in subsequent chapters.

In the next results chapter (**chapter 4**), I use a technique called ‘whole-exome sequencing’ (WES) in the 500 HD patients selected from the previous chapter. Exome sequencing captures all of someone’s DNA that used to make protein. This ‘protein-coding’ part of our DNA is considered the most important of our genetic makeup, and large differences to this part of our DNA can cause diseases like HD. Small differences in this DNA, called DNA ‘variants’, however, are normal, and studying these can give critical insight into how we genetically differ to one another. Using exome sequencing, I identify several DNA variants in genes that are associated with early or late onset HD, mostly in genes involved in maintaining and repairing DNA. I also find small differences (independent of length) in the Huntingtin gene that are associated with different HD patient onsets.

My last results chapter (**chapter 5**) uses another technique, called targeted ‘MiSeq’ sequencing, to confirm the differences seen in Huntingtin. Most Huntingtin genes have a single DNA letter difference where a CAG becomes a CAA. So, writing the letters out would look something like: ‘CAGCAG…CAG**CAA**CAG’. But I find there are more genetic spelling mistakes that can occur, and these are strongly associated with different onsets in HD patients. Whenever additional CAAs instead were found, this was associated with later disease onset than expected (for example, ‘CAGCAG…CAG**CAACAA**CAG’). However, some patients instead only had CAGs with no CAAs, and this was only found in people with an earlier than expected HD onset.

In my final chapter (**chapter 6**), I discuss the project generally. My data have important implications for understanding of HD and reinforce that both huntingtin DNA and DNA repair mechanisms are important in disease. Understanding these mechanisms is crucial as these can identify potential drug targets for therapies. Using my data, and supported by other recent work, I present my two-fated pathway model for HD. This model attempts to explain how the genes I find in the current study may act to slow or accelerate when people have HD onset. Finally, I consider future directions given what we have learned from my study. Further work, building on the data presented here, is likely to identify additional genes or processes that affect HD, and these may have pharmacological relevance.

Acknowledgements

I would like to express my sincere gratitude to my supervisors for their incredible support during this project. Professor Lesley Jones for her enthusiasm, knowledge and availability (despite a busy schedule) throughout my PhD. Dr Thomas Massey for his clinical perspective and the many hours we spent whittling away discussing analyses and other topics. Professor Peter Holmans for all his fantastic statistical advice and guidance, and for bearing with me when something required a second explanation! And Professor Nigel Williams for his encouragement and scientific insight. I would also like to thank the patients, families, carers and clinical staff who donated both the time and the samples that made Registry-HD (and thus this research) possible. Thank you to the Cardiff University School of Medicine who kindly funded this studentship.

A huge thanks to Dr Elliott Rees for his bioinformatic expertise, and for taking the time to introduce me to the wonderful world of Linux supercomputing and bioinformatics. I am very grateful to Professor Darren Monckton and his team at Glasgow University for their hospitality during my stay there, especially to Dr Marc Ciosi for his MiSeq sequencing method and training. I am also grateful to Áine Moylett for her work optimising the *HTT* repeat Sanger sequencing and to Dr Alex Richards for his help with SKAT(-O). Many thanks to Jo Morgan and Sarah Jacques for their patience in training me on NGS library prep and for their sequencing of our exomes. A big thanks to Alex Evans for her work genotyping our patient samples. And a special thanks to rest of the MRC core team for their incredible work keeping the labs functioning.

To my fellow PhD-students-at-arms Jasmine Donaldson, John Hubert and Ant Warland, thanks for all the fun and laughs we've shared over the years. Thank you to my other office compatriots Carolina Toste, Matt Bracher-Smith, Natalie Simon and Sum Mistry for always brightening my day – it means so much! To my wonderful colleagues in the wider HD group, thank you all! Dr Sergey Lobanov, Joe Stone, Freja Sadler, Dr Caroline Binda, Dr Laura-Nadine Schuhmacher and especially Lyn Elliston for her steadfast kindness and moral support. Many thanks also to Dr Georgina Menzies and Dr Emma Yhnell for their career advice and support so freely given.

And finally, a thank you to my family. My aunt and uncle for helping to house me during my foray up to Scotland. And to my parents for always being there for me. Although I don't get to visit as much as either of us would like (the geography does not help), their unwavering love, encouragement and support helped me more than I can put into words. Thank you.

Publications and presentations

Forthcoming publications based on the work presented here

McAllister, B., … Williams, N., Holmans, P., Jones, L., Massey, T. Symptom onset in Huntington's disease: analysis of the European Registry population (in preparation).

McAllister, B., … Williams, N., Holmans, P., Jones, L.$^+$, Massey, T$^+$. Exome sequencing the extremes of a stratified population reveals modifiers of Huntington's disease onset (in preparation).

Publications to which I contributed to during this project

GeM-HD Consortium† (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell 178: 887–900. PubMed: https://www.ncbi.nlm.nih.gov/pubmed/31398342.

†Published as a consortium; I am (joint) first author for Cardiff University (Group 2):
McAllister, B.*, Massey, T.*, Medway, C., ... Jones, L.$^+$, Holmans, P.$^+$.

Ellis N.*, Tee A.*, **McAllister B.**, Massey T., McLauchlan D., … Jones L., and Holmans P. Genetic risk underlying psychiatric and cognitive symptoms in Huntington's Disease. [in review]. Bioarchive: https://www.biorxiv.org/content/10.1101/639658v3.

Massey T., **McAllister B.**, and Jones L. (2018). Methods for Assessing DNA Repair and Repeat Expansion in Huntington's Disease. Methods in molecular biology 1780: 483–495. PubMed: https://www.ncbi.nlm.nih.gov/pubmed/29856032.

Oral presentations

CAG Triplet Repeat Disorders, Gordon Research Conference, Lucca, Italy, 2019.
CAG Triplet Repeat Disorders, Gordon Research Seminar, Lucca, Italy, 2019.
Genomics of Brain Disorders Conference, Wellcome Genome Campus, Hinxton, UK, 2018.
Sequencing for Science and Impact. Heath Park Campus, Cardiff University, 2017.

Poster presentations

McAllister B., Massey T., Rees E., Holmans P., Williams N., Jones L. (2019). Whole exome sequencing and regression analyses identify damaging DNA repair variants associated with altered age at onset in Huntington's disease. CAG Triplet Repeat Disorders, Gordon Research Conference, Lucca, Italy.

McAllister B., Massey T, Rees., E., Holmans P., Williams N., Jones L. (2018). Exome sequencing identifies differences in repeat structure as being associated with altered onset in Huntington's patients. European Huntington's Disease Network Plenary Meeting, Vienna, Austria.

McAllister B., Massey T., Holmans P., Williams N., Jones L. (2018). Exome sequencing of Huntington's disease patients identifies mismatch repair machinery and repeat structure as modifiers of disease onset. Genomics of Brain Disorders Conference, Wellcome Genome Campus, Hinxton, Cambridge, UK.

Contents

Summary iii

Lay summary iv

Acknowledgements vi

Publications and presentations vii

Contents viii

Figure list xv

Table list xvii

Equation list xviii

Abbreviations xix

Chapter 1: General introduction 1

1.1 A background to Huntington's disease 1

1.2 Clinical symptoms 2

1.2.1 An overview of clinical symptoms 2

1.2.2 Motor symptoms 2

1.2.3 Cognitive symptoms 3

1.2.4 Psychiatric and behavioural symptoms 4

1.2.5 Secondary symptoms 5

1.2.6 Juvenile Huntington's disease 5

1.3 Genetics of HD 6

1.3.1 *HTT* and the CAG repeat tract 6

1.3.2 Epidemiology of HD 8

1.4 Neuropathology 9

1.4.1 Normal role of the striatum 9

1.4.2 Dysfunction of the striatum in HD 11

1.4.3 Other neuropathology 12

1.5 Molecular pathogenesis 13

1.5.1 Wild-type Huntingtin 13

1.5.2 The generation of toxic species in HD 13

1.5.3 Downstream cellular pathology 17

1.6 Other repeat diseases 20

1.7 Genetic modifiers of Huntington’s disease 25

1.7.1 Human genetic studies 25

1.7.2 Functional work in model systems 26

1.8 Thesis aims 31

Chapter 2: Materials and methods 32

2.1 Study design 32

2.1.1 HD participants 32

2.1.2 Participant data 32

2.2 Age at HD onset determination 37

2.2.1 Rater’s estimate of onset 37

2.2.2 Deriving ages at onset (best-estimate) 37

2.2.3 Individual symptom determination using CCQ data 38

2.2.4 Calculating binary symptom onset using CCQ data 38

2.3 Statistical analyses of clinical data 39

2.3.1 Linear modelling 39

2.3.2 Intergenerational anticipation estimates 40

2.4 Selecting an extreme onset cohort 41

2.5 DNA Preparation 42

2.5.1 DNA used for sequencing 42

2.5.2 DNA Quantitation 42

2.6 Whole-exome sequencing (WES) 44

2.6.1 Whole-exome library preparation 44

2.6.2 Sequencing of WES libraries 45

2.7 Analyses of whole-exome sequencing data 47

2.7.1 Pipelines for variant calling, QC and annotation 47

2.7.2 Exome candidate gene analysis 50

2.7.3 Burden regression analyses 51

2.7.4 Whole-exome SKAT and SKAT-O analyses 51

2.7.5 Pathway analysis 52

2.8 MiSeq sequencing 53

2.8.1 MiSeq library generation 53

2.8.2 Sequencing on the MiSeq 53

2.9 MiSeq data analyses 55

2.9.1 *HTT* genotyping with Scale-HD 55

2.9.2 Estimating instability in MiSeq with Scale-HD 55

2.9.3 Redefining onset groups using MiSeq 55

2.9.4 Generalised linear models for MiSeq data 56

2.10 Sanger sequencing 57

2.10.1 Variant confirmation 57

2.10.2 Sanger sequencing of the *HTT* repeat 57

2.10.3 Primer design 58

2.11 Genescan sizing of *HTT* CAG 63

2.12 Detecting copy number variants in select samples 63

2.13 Computing facilities and analytical tools 64

Chapter 3: Phenotypical analysis and cohort selection in Registry-HD 66

3.1 Introduction 66

3.2 Initial onset determination in Registry-HD 68

3.2.1 Exploring onset types in Registry 68

3.2.2 Rater’s estimate of HD onset 71

3.2.3 Motor onset derivation (best-estimate) 72

3.2.4 Cognitive onset determination (best-estimate) 76

3.2.5 Psychiatric onset determination (best-estimate) 79

3.3 Deriving symptom onset using the CCQ 82

3.3.1 Adjusting CCQ for earlier symptom onset 82

3.3.2 The relationship between age at onset of different symptoms 86

3.4 Symptom prevalence in HD 89

3.4.1 Initial modelling of symptom prevalence 89

3.4.2 Extending the symptom GLMs with additional covariates 94

3.4.3 Using the HADS-SIS scales to investigate phenotype 99

3.4.4 Correlations for all covariates, HADS-SIS and binary CCQ 101

3.5 Quantifying variation in symptom onset explained by CAG length 103

3.6 Estimating anticipation in Registry 109

3.7 Selecting an extreme onset population of Huntington's disease patients 111

3.7.1 Selecting early and late onset HD patients using AMO 111

3.7.2 Comparing the extreme cohort across symptoms 115

3.8 Discussion 119

3.8.1 Overview of results 119

3.8.2 Best-estimate symptom onset estimation 119

3.8.3 Symptom onset determination using CCQ 120

3.8.4 The effect of CAG length on symptom onset 122

3.8.5 Anticipation in Registry 125

3.8.6 Sex differences in HD symptoms 125

3.8.7 Selection of an extreme onset cohort 128

Chapter 4: Exome sequencing of an extreme motor onset cohort of HD patients 130

4.1 Introduction 130

4.2 Exome sequencing of HD patient DNA 132

4.3 Exome quality control pipeline 133

4.3.1 Establishing exome quality using Picard and Hail 133

4.3.2 Detecting exome contamination 136

4.3.3 Comparing imputed sex to our patient database 137

4.3.4 Principal component analysis and ancestry determination 138

4.3.5 Removal of related individuals 140

4.3.6 Redefining early and late populations post re-genotyping 142

4.4 Annotation pipeline and variant prioritisation 145

4.5 Calling *HTT* CAG structure with WES 147

4.6 Candidate gene analysis 149

4.6.1 Selection of candidate genes 149

4.6.2 *FAN1* 149

4.6.3 *EXO1* 154

4.6.4 MSH3 158

4.6.5 *TCERG1* 161

4.6.6 *LIG1* 165

4.6.7 Other candidate genes 167

4.7 Logistic burden regression analysis 169

4.8 Linear burden regression analysis 174

4.9 SKAT and SKAT-O analyses 179

4.10 Pathway analysis 183

4.11 Discussion 186

4.11.1 Overview of results 186

4.11.2 Development of in-house pipelines for exome analysis 186

4.11.3 Short tandem repeats in *HTT*, *MSH3* and *TCERG1* 188

4.11.4 NSD variation in FAN1 189

4.11.5 NSD variation in *EXO1* 191

4.11.6 LoF variants in MSH3 192

4.11.7 WES of other GWAS candidate genes 192

4.11.8 Whole-exome analyses 194

Chapter 5: Using MiSeq to investigate structure and instability of the *HTT* CAG repeat 196

5.1 Introduction 196

5.2 MiSeq sequencing and Scale-HD 198

5.3 Determining *HTT* allele structure 199

5.3.1 Interruptions in the CAG repeat tract 199

5.3.2 The effect of interruption structures on age at onset 204

5.4 Validating MiSeq using genescan data 206

5.4.1 CAG length 206

5.4.2 Repeat instability 210

5.5 Comparing lymphoblastoid and blood DNA 212

5.6 Repeat instability in early onset and late onset individuals 215

5.7 Disease progression and repeat instability 217

5.8 Discussion 219

5.8.1 Overview of results 219

5.8.2 Robustness of the MiSeq protocol and analysis 219

5.8.3 Lymphoblastoid and blood DNA 221

5.8.4 Atypical repeat structures and HD onset 222

5.8.5 Repeat instability in early and late onset 223

Chapter 6: General discussion 225

6.1 Summary of findings 225

6.2 Study limitations 225

6.2.1 Calculating residual age at onset 225

6.2.2 Sample size and extreme phenotype sampling 226

6.2.3 Lymphoblastoid and blood DNA 227

6.2.4 General challenges of NGS and WES 227

6.3 Mechanisms underlying HD onset 230

6.3.1 An overall model for onset 230

6.3.2 Allelic interruptions and their contribution to HD onset 231

6.3.3 DNA repair mechanisms may stabilise or destabilise repeats 234

6.3.4 Contribution of downstream modifiers of HD onset 241

6.4 Future directions 242

6.4.1 Additional disease phenotypes 242

6.4.2 Additional sequencing and analysis 243

6.4.3 Considering other sequencing approaches 244

6.4.4 Downstream analysis of implicated variants 245

6.5 Final conclusions 246

Appendices 247

Appendix 1 – Adjusted CCQ data for 5 and 10 year cut-offs 247

Appendix 2 – Correlation matrices for CCQ onset data by sex 251

Appendix 3 – Full binary symptom, HADS-SIS and covariate correlation matrix 254

Appendix 4 – *FAN1* Sanger sequencing traces 256

Appendix 5 – Non-coding variation identified by WES in other candidate genes 262

Appendix 6 – Coding variation identified in other candidate genes 274

Appendix 7 – Coding variation in *CUBN*, *MUT* and *NOP14* 280

Appendix 8 – Modelling estimated STR length in MSH3 285

Appendix 9 – Average coverage of target genes from WES 286

Appendix 10 – Variant-by-variant logistic regression in exomes 287

Appendix 11 – Whole-exome logistic regression weighting on MAF 289

Appendix 12 – Whole-exome logistic regression weighted on deleteriousness 291

Appendix 13 – Whole-exome linear regression weighted on deleteriousness 292

Appendix 14 – Whole-exome continuous SKAT-O test weighted on deleteriousness 294

Appendix 15 – Whole-exome auxiliary analysis excluding non-European individuals 295

Appendix 16 – FastQC QC metrics for MiSeq 304

Appendix 17 – Inter-run consistency of MiSeq 306

Appendix 18 – Alternative GLMs for allele structure on residual HD age at motor onset 307

Appendix 19 – SNP genotyping array of lymphoblastoid cells 308

References 309

Figure list

Figure 1.1: Structure and penetrance of HTT exon 1 8

Figure 1.2: The normal function of the basal ganglia and the striatum. 10

Figure 1.3: Dysfunction of the basal ganglia and striatum in early HD. 12

Figure 1.4: A repetitive hairpin/stem-loop in RNA/DNA. 15

Figure 1.5: Mechanisms by which toxic species are generated in HD. 16

Figure 1.6: Downstream molecular and cellular dysfunction in HD. 19

Figure 2.1: PicoGreen™ standard curve. 42

Figure 2.2: WES index overview for plates (96 plate format). 44

Figure 2.3: First QC step in WES using a DNA 1000 chip. 46

Figure 2.4: Second QC step in WES with a HS DNA chip. 46

Figure 2.5: An overview of the exome quality control pipeline. 48

Figure 2.6: Overview of the annotation pipeline. 49

Figure 2.8: Bioanalyser traces of MiSeq libraries pre- and post-clean-up. 54

Figure 2.9: Corrected and uncorrected MiSeq genotyping. 56

Figure 3.1: HD onset types observed in Registry. 68

Figure 3.2: HD onset type by age at onset in Registry. 70

Figure 3.3: Comparison of sxrater and CCQ for motor HD onset. 73

Figure 3.4: Motor onset in Registry across different ages and CAG repeat lengths. 74

Figure 3.5: Comparison of sxrater and CCQ for cognitive HD onset. 76

Figure 3.6: Cognitive onset in Registry across different ages and CAG repeat lengths. 77

Figure 3.7: Initial APO estimation (before adjustment with sxrater). 79

Figure 3.8: Initial APO estimation (after adjustment with sxrater). 80

Figure 3.9: Distribution of psychiatric symptoms in Registry (before sxrater adjustment). 82

Figure 3.10: Comparing onset estimation between sxrater and CCQ. 84

Figure 3.11: Distribution of psychiatric symptoms in Registry (after sxrater adjustment). 85

Figure 3.12: Sex differences for ages at onset for symptoms in Registry. 87

Figure 3.13: Correlation plots for ages at onset comparing between sexes. 88

Figure 3.14: Correlation plot for binary-derived CCQ data, HADS-SIS and covariates. 102

Figure 3.15: Regression for apathy against CAG length. 105

Figure 3.16: Anticipation in Registry. 110

Figure 3.17: A comparison of the Langbehn model in Registry cohort (AMO). 112

Figure 3.18: Extreme AMO cohort selection for sequencing. 114

Figure 3.19: Ages at onset for other symptoms in the extreme AMO cohort. 118

Figure 4.1: Quality control metrics for exome libraries using Picard. 134

Figure 4.2: Quality control metrics for exome libraries using Hail. 135

Figure 4.3: Detection of exome contamination using VerifyBamID. 136
Figure 4.4: Sex checks using Peddy. 137
Figure 4.5: Hail-derived principal components. 138
Figure 4.6: Ancestry of WES samples estimated by Peddy. 139
Figure 4.7: Identity by decent ratios for WES samples. 140
Figure 4.8: Genetic relationship matrices for WES samples. 141
Figure 4.9: A flowchart of the exomes kept/lost in the quality control pipeline. 143
Figure 4.10: Redefining early and late populations with MiSeq genotypes. 144
Figure 4.11: Non-synonymous damaging (NSD) variants across the exome. 146
Figure 4.12: The canonical *HTT* CAG repeat structure. 148
Figure 4.13: Structural overview of identified *FAN1* variants. 150
Figure 4.14: Structural overview of identified *EXO1* variants. 157
Figure 4.15: Structure of the TCERG1 repeat. 164
Figure 4.16: *TCERG1* genotype length plotted against corrected AMO residual. 164
Figure 4.17: Structural overview of identified LIG1 variants. 165
Figure 4.18: Effect of the *MLH1* Ile219Val variant (rs1799977) on residual AMO. 168
Figure 4.19: Q-Q plot of whole-exome burden regression. 173
Figure 4.20: Q-Q plot of whole-exome linear regression (corrected residual). 178
Figure 4.21: Q-Q plots for SKAT and SKAT-O analyses. 180
Figure 4.22: Members of the top pathways from gene set analysis. 185
Figure 5.1: Visualisation of observed *HTT* alleles from MiSeq. 201
Figure 5.2: Comparing CAG lengths from MiSeq, genescan and local labs. 207
Figure 5.3: Residual ages at motor onset from MiSeq, genescan and local labs. 208
Figure 5.4: CAG lengths before and after using MiSeq against age at motor onset. 209
Figure 5.5: CAG repeat read frequency plots for wild type and expanded alleles. 211
Figure 5.6: MiSeq and genescan somatic instability measures. 211
Figure 5.7: CAG instability in blood and lymphoblastoid cells. 213
Figure 5.8: Longitudinal lymphoblastoid cell CAG length and mosaicism. 214
Figure 5.9: MiSeq mosaicism and age at onset (LBC and blood). 216
Figure 5.10: Comparing progression and age at motor onset. 217
Figure 5.11: Relationship between MiSeq mosaicism and progression (LBC and blood). 218
Figure 6.1: Proposed mechanisms for onset modification in HD. 231
Figure 6.2: Repeat sequence and secondary DNA structure. 233
Figure 6.3: A two-fated pathway model for DNA repair in HD. 240

Table list

Table 1.1: The polyglutamine repeat diseases. 22
Table 1.2: The non-coding repeat diseases and disease-associated repetitive fragile sites. 24
Table 1.3: Candidate genetic modifiers of Huntington’s disease. 28
Table 1.4: An overview of the DNA canonical DNA repair pathways. 30
Table 2.1: Select demographic and phenotypic data terms. 33
Table 2.2: The clinical characteristics questionnaire (CCQ). 34
Table 2.3: Hospital anxiety and depression scale (HADS) questionnaire. 35
Table 2.4: Snaith’s irritability scale (SIS) questionnaire. 36
Table 2.5: *FAN1* Sanger sequencing primers. 60
Table 2.6: Other Sanger sequencing primers (not *FAN1*). 62
Table 2.7: Databases used for analyses. 64
Table 2.8: Bioinformatic tools used for analyses. 65
Table 3.1: Onset summaries by CAG (general onset). 71
Table 3.2: Age at motor onset summary statistics. 75
Table 3.3: Age at cognitive onset summary statistics. 78
Table 3.4: Age at first psychiatric symptom summary statistics. 81
Table 3.5: Prevalence of HD symptoms in men and women. 91
Table 3.6: Unadjusted CCQ data GLMs for symptom prevalence. 92
Table 3.7: Adjusted CCQ data GLMs for symptom prevalence. 93
Table 3.8: Unadjusted CCQ data extended GLMs. 96
Table 3.9: Adjusted CCQ data extended GLMs. 98
Table 3.10: GLMs for the HADS-SIS scale. 100
Table 3.11: GLMs examining expanded and wild-type CAGs on age at symptom onset. 104
Table 3.12: Variance explained by CAG length for symptoms across sexes. 106
Table 3.13: CAG length on less stringent CCQ-derived data. 108
Table 3.14: Parental HD onset summary information. 109
Table 3.15: Selected individual CAGs for exome sequencing. 113
Table 3.16: Frequencies of symptoms in the extreme AMO cohort (CCQ). 116
Table 4.1: An overview of the samples exome sequenced. 132
Table 4.2: Atypical *HTT* allele structures identified by WES. 148
Table 4.3: Coding variation from *FAN1* from WES. 152
Table 4.4: Sanger sequencing confirmation of *FAN1* variants. 153
Table 4.5: Coding variation from *EXO1* from WES. 156
Table 4.6: Coding variation in *MSH3* from WES. 160
Table 4.7: Coding variation in *GPR151* from WES. 162

Table 4.8: Coding variation in *TCERG1* from WES. 163
Table 4.9: Coding variation in *LIG1* from WES. 166
Table 4.10: Logistic regression for candidate genes. 171
Table 4.11: Logistic regression of the entire exome (MAF 0.1, 1 and 2%), no weighting. 172
Table 4.12: Linear regression (uncorrected residual) of the entire exome (MAF 0.1, 1 and 2%), no weighting. 175
Table 4.13: Linear regression for candidate genes. 176
Table 4.14: Linear regression (corrected residual) of the entire exome (MAF 0.1, 1 and 2%), no weighting. 177
Table 4.15: SKAT and SKAT-O dichotomous and continuous tests. 181
Table 4.16: Whole-exome dichotomous and continuous SKAT and SKAT-O tests. 182
Table 4.17: Pathway analysis of NSD variants. 184
Table 5.1: Details of DNA sequenced using MiSeq. 198
Table 5.2: Identified *HTT* structures in an extreme onset HD cohort. 202
Table 5.3: Comparing WES and MiSeq *HTT* structural calls. 203
Table 5.4: GLMs for *HTT* allele structure on residual HD age at motor onset. 205
Table 5.5: Differences between blood and lymphoblastoid CAG lengths (MiSeq). 212
Table 5.6: GLMs interrogating modifiers of somatic mosaicism. 216
Table 5.7: GLMs interrogating modifiers of HD disease progression. 218

Equation list

Equation 2.1: Full parametric Langbehn survival model. 41
Equation 2.2: Simplified Langbehn model for mean onset at a given CAG length. 41
Equation 2.3: MiSeq mosaicism (Scale-HD). 55
Equation 2.4: MiSeq slippage (Scale-HD). 55

Abbreviations

ACO: Age at cognitive onset

AMO: Age at motor onset

ANX: Anxiety

APO: Age at psychiatric onset

APT: Apathy

ARCCA: Advanced Research Computing Division at Cardiff University

BAM: Binary alignment map (file-type)

BER: Base excision repair

β-ME: β-mercaptoethanol (also known as 2-mercaptoethanol)

BVR: Baseline variant rate

CADD: Combined Annotation Dependent Depletion

CAG: Cytosine-adenosine-guanine (repeat)

CCQ: Clinical characteristics questionnaire

CEX: Coding exome oligonucleotides

CNV: Copy number variant

COG: Cognitive (symptoms)

CR: Call rate

DCL: Diagnostic confidence level

DEP: Depression

DM: Myotonic dystrophy

DMSO: Dimethyl sulfoxide

DP: (Read) depth

E: Early

EBV: Epstein Barr virus

fs: Frameshift

GATK: Genome analysis toolkit

GLM: Generalised linear model

GO: Gene ontology

GQ: Genotype quality

GRCh37: Genome reference consortium human genome build 37

GRM: Genetic relationship matrix

GWA(S): Genome-wide association (study)

HADS(-SIS): Hospital anxiety and depression scale (with Snaith's irritability scale)

HD: Huntington's disease

Het: Heterozygote

Hg19: Human genome assembly GRCh37

HomR: Homozygote reference

HomV: Homozygote variant

HTT: Human Huntingtin gene

IBD: Identity by descent

ICD: International Statistical Classification of Disease

ICL: Inter-strand crosslink

iPS(C): Induced pluripotent stem (cell)

IRB: Irritability

JHD: Juvenile Huntington’s disease

KIN: Karyomegalic interstitial nephritis

L: Late

LBC: Lymphoblastoid cell

LoF: Loss-of-function

MAF: Minor allele frequency

MMR: Mismatch repair

MTR: Motor (symptoms)

MutLα: Heterodimer of MLH1-PMS2

MutLβ: Heterodimer of MLH1-PMS1

MutLγ: Heterodimer of MLH1-MLH3

MutSα: Heterodimer of MSH2-MSH6

MutSβ: Heterodimer of MSH2-MSH3

NFE: Non-Finnish Europeans

NGS: Next-generation sequencing

NS: Non-synonymous

NSD: Non-synonymous damaging

nt: Nucleotide

O_{exp}: Expected onset

O_{obs}: Observed onset

ORF: Open reading frame

PBA: Problem behaviours assessment

PC: Principal component

PCA: Principal component analysis

PCNA: Proliferating cell nuclear antigen

PCR: Polymerase chain reaction

POB: Perseverative/obsessive behaviour

PolyP: Polyproline

PolyQ: Polyglutamine

PRS: Polygenic risk score

PSY: Psychosis

Refseq: Reference sequence

RTR: Shorthand for sxrater (rater’s best estimate of onset)

SAP: SAF-A/B, Acinus and PIAS domain

SD: Standard deviation

SDMT: Symbol digit modalities test

SI: Somatic instability

SIS: Snaith’s irritability scale

SKAT: Sequence kernel association test

SKAT-O: Optimised sequenced kernel association test

SMB: Streptavidin magnetic beads

SNP: Single nucleotide polymorphism

SNV: Single nucleotide variant

SPRI: Solid phase reversible immobilisation

STR: Short tandem repeat

Sxrater: Rater’s best estimate of onset

SZ: Schizophrenia

TAS: Total anxiety score

TDS: Total depression score

TFC: Total functional capacity

TIS: Total irritability score

TMS: Total motor score

TPR: Tetratricopeptide repeat domain

TWAS: Transcriptome-wide association study

UHDRS: Unified Huntington’s disease rating scale

UHDRS-b: Behavioural component of the UHDRS

UTR: Untranslated region

VAB: Violent/aggressive behaviour

VCF: Variant calling file

VDS: Variant dataset file (used by Hail)

VEP: Variant effect predictor tool (ensembl)

VQSR: Variant Quality Score Recalibration

VRR-Nuc: Virus-type replication-repair nuclease

WES: Whole-exome sequencing

WGS: Whole-genome sequencing

Chapter 1: General introduction

1.1 A background to Huntington's disease

In 1872, George Huntington gave his seminal lecture 'On Chorea', published shortly thereafter in *The Medical and Surgical Reporter* (Huntington, 1872). Huntington described what he called the 'hereditary chorea', commonly known at the time in New England as 'magrums', in striking detail, possible only through his substantial contact with affected individuals and their families. In his paper, Huntington succinctly described three of the primary features of the disease: (1) its hereditary inheritance, (2) its motor and psychiatric involvement and (3) its fatal outcome. While not the first, the vividness and clarity of Huntington's description attracted significant worldwide interest by the medical community, and the disease became known as Huntington's disease (HD).

Significant progress has been made since the original descriptions of the disease, and one of the greatest advances was the discovery of HD in several fishing communities on the Venezuelan Lake Maracaibo by physician Americo Negrette (Negrette, 1955; Okun and Thommi, 2004). These communities have the highest incidence of HD in the world due to a founder mutation, and have become exceptionally well characterised through prolonged longitudinal study by Nancy Wexler and colleagues, now over several decades (Wexler et al., 2004; Wexler, 2013). Genetic material donated by the Venezuelan HD kindreds greatly contributed towards HD research, and in 1983 the locus (4p16.3) containing the causative gene for HD was identified (Gusella et al., 1983). 10 years later in 1993, closely following the original description of repeat disease by La Spada et al. (La Spada et al., 1991), the causative mutation for HD was found: a trinucleotide repeat expansion mutation in the *HTT* (huntingtin) gene (The Huntington's Disease Collaborative Research Group, 1993). It is now known HD is one of a family of polyglutamine diseases (see 1.6 & Table 1.1), a group of diseases caused by CAG (encoding glutamine) triplet repeat expansions, and these are part of a broader group of repeat diseases (see 1.6 & Table 1.2).

HD is a destructive neurodegeneration, typically fatal 15-20 years following its onset. It often presents mid-life (~40-50 years), but can also occur much earlier or later, and juvenile onset (<20 years) occurs in ~5% of cases (Quarrell et al., 2012). The clinical symptoms of HD are varied, and encompass motor abnormalities, cognitive decline and psychiatric/behavioural changes, as explored in the following section (1.2). The size of the CAG repeat expansion is a major determinant of disease onset, with CAG repeat size strongly inversely correlated with onset (Andrew et al., 1993; Wexler et al., 2004; Lee et al., 2012c) (1.3). Mechanistically, the expanded *HTT* gene predominantly acts *via* a gain of function mechanism that leads to neuronal death (1.5), preferentially (although not wholly) in the medium spiny projection neurons of the striatum, part of the basal ganglia (1.4). Following a general overview of HD and a brief introduction of other repeat diseases (1.6), I then discuss the elucidation of pathologically relevant genetic modifiers of HD (1.7). The aims of the current thesis are set out in 1.8. Principally, these aims are to explore genetic modifiers of HD using next-generation sequencing techniques in a large cohort of HD patients (Registry-HD, 2.1).

1.2 Clinical symptoms

1.2.1 An overview of clinical symptoms

HD is often described as having a triad of symptomatic domains: motor, cognitive and psychiatric/behavioural domains. While motor symptoms are outwardly the most obvious symptoms experienced by HD patients – indeed, originally HD was called Huntington's chorea – it is increasingly recognised non-motor symptoms of HD play a prominent role for patients, who often find these the most debilitating aspects of disease (Vamos et al., 2007; Ready et al., 2008; Paulsen et al., 2010; Tabrizi et al., 2013; Bachoud-Lévi et al., 2019). The most common clinical assessment for HD is the Unified Huntington's Disease Rating Scale (UHDRS) (Huntington Study Group, 1996), which contains a number of motor, behavioural, cognitive, emotional and functional elements. These include Total Functional Capacity (TFC) score, a measure of independence that ranges from 13 (no or minimal impairment) to 0 (advanced disease with total dependence on others), and Total Motor Score (TMS), a measure of motor impairment (see 1.2.2). The diagnostic confidence level (DCL) component of the UHDRS is often used for defining HD motor onset. The DCL measures how confident the assessing clinician that the patient's motor symptoms are the result of HD, with a score of 0 indicating no motor symptoms suggestive of HD, and a score of 4 indicating a ≥99% confidence that motor symptoms are due to HD (and thus a motor HD onset). However, studies have shown subtle motor, psychiatric and cognitive symptoms precede overt motor symptoms in HD individuals, as long as 10-15 years before clinical motor diagnosis (Folstein et al., 1983a; Huntington Study Group, 1996; Paulsen et al., 2008, 2014, 2017; Stout et al., 2011; Tabrizi et al., 2013; Paulsen and Long, 2014; Reilmann et al., 2014; Ross et al., 2014; Huntington Study Group PHAROS Investigators et al., 2016; Martinez-Horta et al., 2016).

1.2.2 Motor symptoms

Chorea, derived from the Greek "dance", are non-repetitive, involuntary jerky or writhe-like movements found in most, but not all, manifest HD patients, although HD individuals may be unaware of these involuntary movements (anosognosia) until the movements become more advanced (McCusker et al., 2013; Sitek et al., 2014). Choreic movements usually begin in the distal extremities (especially the fingers and toes) and facial muscles. Chorea often worsens over disease course, and can severely impact normal functioning leading to imbalance, difficulty walking, gait change and risk of falling (Grimbergen et al., 2008; Busse et al., 2009). Chorea can also cause speech difficulties (dysarthria), swallowing problems (dysphagia) and breathing trouble, depending on the muscle groups involved. Other motor abnormalities include muscle rigidity, dystonia (uncontrollable muscle contractions that can alter posture), bradykinesia (slowness of movement), akinesia (impaired ability to start movement), myoclonus (sudden muscle contractions) and bruxism (involuntary teeth grinding/jaw clenching) (Bachoud-Lévi et al., 2019). Together with chorea, motor abnormalities eventually result in an inability to walk. The extent of these symptoms varies from patient to patient and can often change over the course of disease. For instance, more advanced HD patients may have less chorea and more rigidity compared with earlier in disease (Roos, 2010).

More subtle motor symptoms can occur before more overt chorea and/or clinical motor diagnosis (Biglan et al., 2009). These include slowed finger tapping (Rowe et al., 2010; Tabrizi et al., 2012, 2013), gait alteration (Rao et al., 2011) and oculomotor changes. Oculomotor HD symptoms include slowed horizontal smooth pursuit of the eye (Winder and Roos, 2018), impaired initiating and slowing of saccadic eye movement (voluntary rapid movement of the eye between fixation points) and reduction of optokinetic nystagmus (normal eye reflex for tracking of moving objects) (Oepen et al., 1981; Blekher et al., 2004, 2006; Biglan et al., 2009). Motor abnormalities in HD can be clinically assessed using the total motor score (TMS) component of the UHDRS, which contains 31 individual motoric items assessed by the clinician (reviewed by (Reilmann and Schubert, 2017)).

1.2.3 Cognitive symptoms

Subtle cognitive deficits occur at least 10 years before clinical HD onset. During this pre-manifest period, HD individuals experience slight cognitive decline among several domains including executive function, visual motor integration skills, emotion recognition and psychomotor ability (Lemay et al., 2005; Say et al., 2011; Stout et al., 2011, 2012; Tabrizi et al., 2011; Harrington et al., 2012; Papp et al., 2013; Papoutsi et al., 2014; Huntington Study Group PHAROS Investigators et al., 2016; Martinez-Horta et al., 2016). Cognitive dysfunction worsens over time and working memory, episodic memory and learning are often impaired by HD motor onset (Dumas et al., 2013; Papoutsi et al., 2014). Similar to chorea, patients may be unaware of these changes as cognitive symptoms are often underreported by patients (McCusker and Loy, 2014; Sitek et al., 2014). Certain functions are relatively spared in the early stages of disease, including semantic memory, language and spatial awareness (reviewed by (Dumas et al., 2013)). However, advanced HD is overwhelmingly associated with cognitive impairment and dementia, although HD dementia tends to have a lesser memory component compared to dementia in Alzheimer's disease (Aretouli and Brandt, 2010; Peavy et al., 2010). HD cognitive deficits can be assessed using cognitive tests part of the UHDRS – these tests include the Stroop inference test (executive function), symbol digit modality test (SDMT) (psychomotor ability), circle tracing test (visuospatial ability) and emotion recognition test.

1.2.4 Psychiatric and behavioural symptoms

HD can encompass a variety of behavioural and psychiatric disturbances, although the extent to which individual patients experience these symptoms is quite variable. Depression is very common in HD patients and can frequently occur before clinical HD onset (van Duijn et al., 2007, 2014; Epping et al., 2013; Martinez-Horta et al., 2016), although depression may decline in the latter stages of disease (Paulsen et al., 2005b). Suicide ideation is also common in individuals with HD, particularly before receiving a diagnosis of HD or when independence is diminished later in disease (Paulsen et al., 2005a; Larsson et al., 2006; Robins Wahlin, 2007; Eddy et al., 2016; van Duijn et al., 2018). Approximately ~5-7% of HD individuals commit suicide (Sørensen and Fenger, 1992; Cardoso, 2017). Irritability, anxiety and perseveration are also experienced by many HD patients (van Duijn et al., 2014), and, similar to depression, irritability occurs at elevated rates in pre-manifest HD individuals (Bouwens et al., 2015; Martinez-Horta et al., 2016). Psychosis can also be a symptom of HD, although this is rarer (only ~5-10% of patients) (van Duijn et al., 2014; Rocha et al., 2018). Apathy is the most common neuropsychiatric disturbance experienced, being reported in at least ~50% of HD patients (Paulsen et al., 2001; van Duijn et al., 2014). Apathy is present in pre-manifest HD (Martinez-Horta et al., 2016), and consistently worsens during disease course (Tabrizi et al., 2013).

Due to the wide range of behavioural and psychiatric disturbances, there is no single accepted assessment battery for these symptoms, although the problem behaviours assessment (PBA) (Craufurd et al., 2001) and Hospital Anxiety and Depression Scale (HADS) (Zigmond and Snaith, 1983) are commonly used. The long-form PBA contains 40 items addressing various behavioural abnormalities common in HD (*e.g.* depression, irritability, aggression, *etc*.), whereas the HADS is a shorter 14 item questionnaire addressing depression and anxiety specifically (see later in Table 2.3). Snaith's irritability scale (SIS) can be used for assessing the severity of irritability in HD, and is a short 8 item questionnaire similar to the HADS (Snaith et al., 1978) (Table 2.4). For an excellent review of behavioural rating scales in the context of HD, see Mestre et al. (Mestre et al., 2016).

1.2.5 Secondary symptoms

In addition to the primary HD domains, HD patients can also experience a variety of peripheral symptoms. Sleep disturbance is very common, affecting ~70% of patients (Arnulf et al., 2008; Videnovic et al., 2009; Bellosta Diago et al., 2017; Bachoud-Lévi et al., 2019). Chronic pain is also common amongst HD patients (Arran et al., 2014), with approximately 40% of patients reporting pain in a recent study (Underwood et al., 2017), although the aetiology of this pain is not well understood. Cardiac failure is common in more advanced disease, and is the second most common cause of death in patients (second to aspiration pneumonia) (Sørensen and Fenger, 1992). There is also some evidence of HD-related cardiomyopathy more generally in disease (reviewed by (Critchley et al., 2018)). In male HD patients, limited testicular atrophy and decreased testosterone levels can also be apparent (Markianos et al., 2005; Van Raamsdonk et al., 2007), although notably this does not appear to affect fertility (Pridmore and Adams, 1991).

Perhaps the most reported secondary HD symptom is weight loss, which can occur in pre-manifest disease (Djousse et al., 2002; Mochel et al., 2007) and also increases in severity with *HTT* CAG length (Aziz et al., 2008). The cause of weight loss in HD is poorly understood and is probably multifactorial. Weight loss may partially arise from dysphagia (problems swallowing) and increased energy expenditure as motor symptoms advance (Trejo et al., 2004; Gaba et al., 2005; Brotherton et al., 2012). Gastrointestinal problems, another symptom experienced by many HD patients, could also contribute (Andrich et al., 2009; Aziz et al., 2010; McCourt et al., 2015; Sciacca et al., 2017). Still, evidence in patients and model HD systems have shown weight loss is probably at least partially a consequence of gross metabolic imbalance in HD: mitochondria and hepatic (liver) dysfunction (Panov et al., 2002; Chiang et al., 2007; Josefsen et al., 2010; Hoffmann et al., 2014), insulin insensitivity (Podolsky et al., 1972; Podolsky and Leopold, 1977; Lalić et al., 2008) and adipose tissue abnormalities (Lakra et al., 2019) have all been reported as possible mechanisms underlying this weight loss phenotype.

1.2.6 Juvenile Huntington's disease

Juvenile HD (JHD) occurs in ~5% of HD cases (Potter et al., 2004; Quarrell et al., 2012), and is defined as HD onset <20 years of age (Roos, 2010). JHD can arise in individuals where *HTT* CAG is >60 CAG repeats (Douglas et al., 2013), although can also occur at smaller repeat lengths (normally ~50-60 CAGs) (Squitieri et al., 2006; Ribaï et al., 2007). Exceptionally long CAGs (>200) may even be associated with an infantile form of JHD (Nicolas et al., 2011). Considering JHD is important as its clinical manifestation can differ compared to that of adult onset HD. For instance, in JHD cognitive symptoms are often the first symptoms noticed, and JHD is associated with marked cognitive decline including speech and language delay, learning deficiencies and difficulties at school (Gonzalez-Alegre and Afifi, 2006; Squitieri et al., 2006; Yoon et al., 2006; Ribaï et al., 2007). JHD patients, especially those that are younger, can also present with seizures and epilepsy (Gonzalez-Alegre and Afifi, 2006; Cloud et al., 2012), a symptom not typically seen in adult onset HD. Chorea in many cases only presents later (10-20 years of age), and JHD is often associated more with bradykinesia and rigidity (Roos, 2010). Even within JHD there is marked heterogeneity between patients, with larger *HTT* CAG sizes being associated with a more severe disease (Fusilli et al., 2018).

Approximately two-thirds of JHD cases are paternally transmitted (Ridley et al., 1988; Myers et al., 1993; Ranen et al., 1995; Gonzalez-Alegre and Afifi, 2006). This is a consequence of intergenerational repeat instability, wherein the *HTT* CAG repeat can become larger when transmitted to offspring (this concept is detailed in the next section, 1.3), and most large intergenerational CAG expansions occur in the paternal line during spermatogenesis (Yoon et al., 2003; Wheeler et al., 2007; Simard et al., 2014; Neto et al., 2017; Jamali et al., 2018). Large CAG expansions can also originate in the maternal line (Nahhas et al., 2005), although these events are rarer.

1.3 Genetics of HD

1.3.1 *HTT* and the CAG repeat tract

A triplet CAG repeat expansion in the *HTT* (huntingtin) gene is the aetiological cause of HD. The expansion arises from a native repetitive CAG sequence highly polymorphic in size, usually between 6-35 CAGs in the normal population (Gardiner et al., 2019). The size of the CAG repeat expansion has been shown to be the major determinant of HD onset (Andrew et al., 1993; Duyao et al., 1993; Snell et al., 1993), with larger CAG repeat expansions being strongly inversely correlated with earlier disease onset (see Fig. 3.4C later). CAG length accounts for ~50-60% of the age at motor onset variation in HD (Andrew et al., 1993; Duyao et al., 1993; Snell et al., 1993; Illarioshkin et al., 1994; Kieburtz et al., 1994; Rosenblatt et al., 2001; Aylward et al., 2004; Wexler et al., 2004; Lee et al., 2012c; Rinaldi et al., 2012) (discussed in more detail in 3.8.4). The association between CAG length and onset has been reported to be as high as ~70% in Venezuelan kindreds (Wexler et al., 2004; Andresen et al., 2007b, 2007a), although this may partially be due to inter-relatedness. Using CAG length, it is also possible to predict the age at onset of individuals with HD, although not with enough accuracy to be clinically reliable (Brinkman et al., 1997; Langbehn et al., 2004). The remaining onset variability (~40%) is partially heritable, indicating other genes have a role in determining HD onset (Duyao et al., 1993; Wexler et al., 2004; Andresen et al., 2007b, 2007a) (see 1.7 for consideration of non-*HTT* genetic factors that affect HD).

Disease penetrance is also closely tied to *HTT* CAG repeat length (see Fig. 1.1). *HTT* alleles with 36-39 CAGs are associated with partial disease penetrance, *i.e.* these alleles can cause disease, but only in a subset of individuals. When associated with disease, reduced penetrance alleles are frequently associated with late disease onset and an incomplete and often milder HD phenotype (Rubinsztein et al., 1996; Migliore et al., 2019). Recently, studies have shown reduced penetrance alleles can occur infrequently (~1/400) in the general population (Kay et al., 2016; Gardiner et al., 2019). 40 CAG alleles have a high disease penetrance, with an estimated ~93% of these alleles being associated with disease (Langbehn et al., 2004), and ≥41 CAGs is associated with complete, or very near complete, disease penetrance.

Intermediate alleles, commonly defined as *HTT* alleles with between 27-35 CAG repeats, are not associated with disease, but may become pathogenic in subsequent generations (Semaka et al., 2013; Migliore et al., 2019), resulting in *de novo* HD (Myers et al., 1993; Kay et al., 2018). This phenomenon of sporadic *de novo* HD is the result of repeat expansion, wherein large CAG tracts can become increasingly long intergenerationally (*i.e.* repeat expansion occurs in gametic cells). How prone intermediate alleles are to vertical instability and repeat expansion is somewhat unclear, however, and it appears most intergenerational expansions occur towards the larger allelic range (34-35 CAGs) (Brocklebank et al., 2009; Hendricks et al., 2009; Semaka et al., 2010, 2013). Semaka and Hayden suggested further subdividing intermediate alleles into CAG sizes of low (27-29 CAGs), moderate (30-33 CAGs) and high (34-35 CAGs) risk (Semaka and Hayden, 2014) to more accurately represent the liability of intermediate alleles to expand.

In addition to penetrance, intergenerational CAG repeat instability can also cause clinical anticipation, wherein HD tends to have earlier onset in subsequent generations in HD families. Anticipation is caused by already pathogenic CAG alleles expanding further when vertically transmitted, thus leading to earlier HD onset in the children of HD individuals (Ranen et al., 1995; Teisberg, 1995; McInnis, 1996; Demetriou et al., 2018). As alluded to previously in 1.2.6, the sex of the transmitting parent is important. Paternally transmitted

HTT alleles are more prone to instability and larger CAG repeat expansions, thus leading to a larger average anticipation when passed through the male germline (Ridley et al., 1988; Duyao et al., 1993; Zühlke et al., 1993; Wheeler et al., 2007; Aziz et al., 2011; Ramos et al., 2012a). This seems to be due to spermatogenesis being more prone to repeat expansion (Yoon et al., 2003; Wheeler et al., 2007; Simard et al., 2014; Neto et al., 2017; Jamali et al., 2018), although the mechanisms of increased paternal expansion rate are poorly understood. It is worth noting that as well as intergenerational repeat expansion (in gametic cells), somatic instability (in non-gametic (somatic) cells) occurs in HD and many other repeat diseases. This concept of somatic *HTT* repeat instability may be an important pathogenic driver of disease and is considered in 1.5 & 1.7.

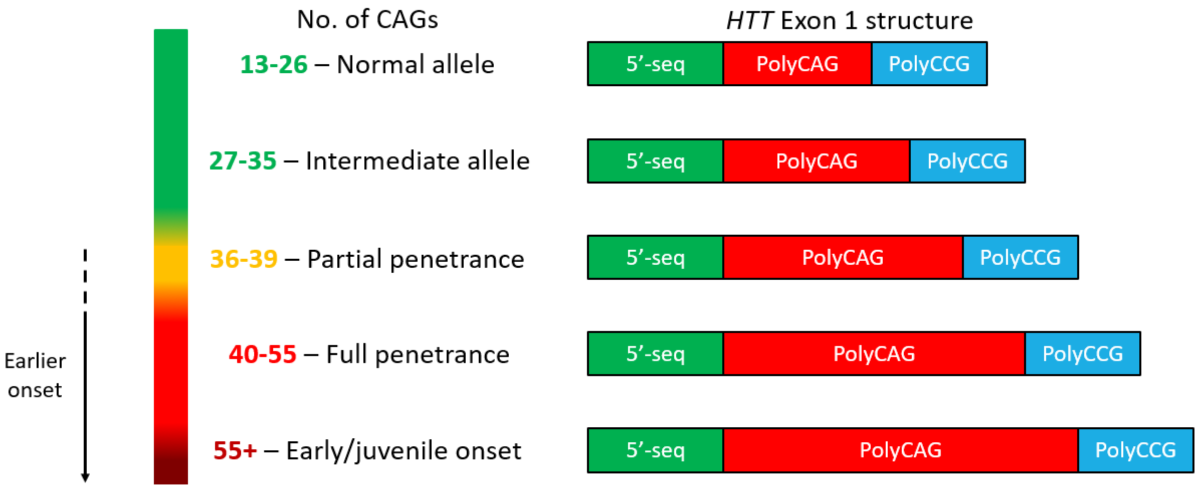


Figure 1.1: Structure and penetrance of HTT exon 1. Indicated are the ranges for the CAG repeat in *HTT* on the left. Longer CAGs are associated with earlier onset once into the fully penetrant pathological range (≥40 CAGs). On the right, a cartoon structure of *HTT* exon 1 is shown. Note this figure does not include the sequence 3' to the polyCCG repeat. 5'-seq: The first 51 nucleotides of *HTT* (which encode the N-terminal first 17 amino acids of HTT); PolyCAG: Repetitive CAG tract; PolyCCG: Repetitive CCG tract.

1.3.2 Epidemiology of HD

HD is most common in those of European descent where estimates vary between ~5-13 cases per 100,000 (Sackley et al., 2011; Evans et al., 2013; Fisher and Hayden, 2014; Rawlins et al., 2016; Squitieri et al., 2016; Kay et al., 2018) (also reviewed generally by (Kay et al., 2017)). Comparatively, HD is rarer in those of Finnish decent (2.12 per 100,000) (Sipilä et al., 2015), owing to Finland's genetic heritage compared to the rest of Europe. Both African (Hayden et al., 1980; Scrimgeour and Pfumojena, 1992; Baine et al., 2016) and East Asian (Chen and Lai, 2010; Kim et al., 2015) ancestries have a low prevalence of HD, with

<1 case per 100,000. Expanded *HTT* alleles from African/Asian populations appear to have unique *HTT* haplotypic origins compared to the commonly expanded European haplotypes (Warby et al., 2011; Baine et al., 2013). As demonstrated by the Venezuelan kindred, founder effects can result in much higher rates of disease in a local area (Wexler et al., 2004). Indeed, Huntington originally described HD as mostly confined to the Long Island region of New York state (Huntington, 1872). Recently, Kay et al. found the frequency of intermediate *HTT* alleles (27-35 CAGs) closely mirrored the prevalence of HD (Kay et al., 2018), suggesting that (1) rare expansions in intermediate alleles underlies the prevalence of HD in different populations and (2) longer intermediate alleles (which are more common in Europeans) are more prone to intergenerational expansion (Kay et al., 2018).

1.4 Neuropathology

1.4.1 Normal role of the striatum

Degeneration of the striatum is the neuropathological hallmark of HD (de la Monte et al., 1988; Aylward et al., 1998; Vonsattel and DiFiglia, 1998). Understanding the normal role of the striatum is vital to understanding the neuropathology and dysfunction that occurs in HD (1.4.2). The striatum is part of the subcortical basal ganglia, and the basal ganglia has roles in regulating motor, mood and cognitive circuitry. The striatum is composed of the dorsal (upper) and ventral (lower) striatum. The dorsal striatum is made up of the caudate and putamen, separated by the internal capsule, and the ventral striatum is made of the nucleus accumbens and olfactory tubercle. The striatum inputs into both the globus pallidus internal and globus pallidus externus. The striatum receives dopaminergic input from the substantia nigra pars compacta and glutaminergic input from the cortex. The predominant striatal cell type (~90-95%) are the medium spiny projection neurons (MSNs) (Kita and Kitai, 1988; Waldvogel et al., 2015), which release the inhibitory neurotransmitter γ-aminobutyric acid (GABA). Other neuronal cell types, (*e.g.* cholinergic interneurons) are also present in the striatum but are fewer in number.

The striatum has an integral role in the corticostriatal loop, or ‘motor loop’, as indicated in Fig 1.2 (Alexander and Crutcher, 1990; Parent and Hazrati, 1993, 1995; Graybiel, 1995). In healthy individuals, the striatum acts *via* two pathways: the direct pathway (which facilitates movement) and the indirect pathway (which inhibits movement). Notably, these pathways are not entirely separate (Haber et al., 2000; Cui et al., 2013; Parker et al., 2018), but here we will just consider them as discrete pathways for simplicity. In the direct pathway, inhibition of the globus pallidus internal and substantia nigra pars reticularis leads to dis-inhibition of the thalamus. In contrast, in the indirect pathway the striatum inhibits the globus pallidus externus; this in turn dis-inhibits the subthalamic nucleus which activates the globus pallidus internal, leading to inhibition of the thalamus. The thalamus is a central coordinator of motor signalling, and thus the inhibition of the thalamus represses movement (as per the indirect pathway), whereas excitation/dis-inhibition of the thalamus promotes movement (as per the direct pathway).

The striatum is further subject to modulatory signalling from the substantia nigra pars compacta (Albin et al., 1989). The direct pathway is activated by dopaminergic signalling from the pars compacta, whereas the indirect pathway is inhibited by this signalling. Fine control of the striatum is important for normal basal ganglia-mediated brain functions, and breakdown of the corticostriatal loop has important consequences which can lead to the symptoms seen in Huntington's disease or Parkinson's disease (Rinne et al., 1989).

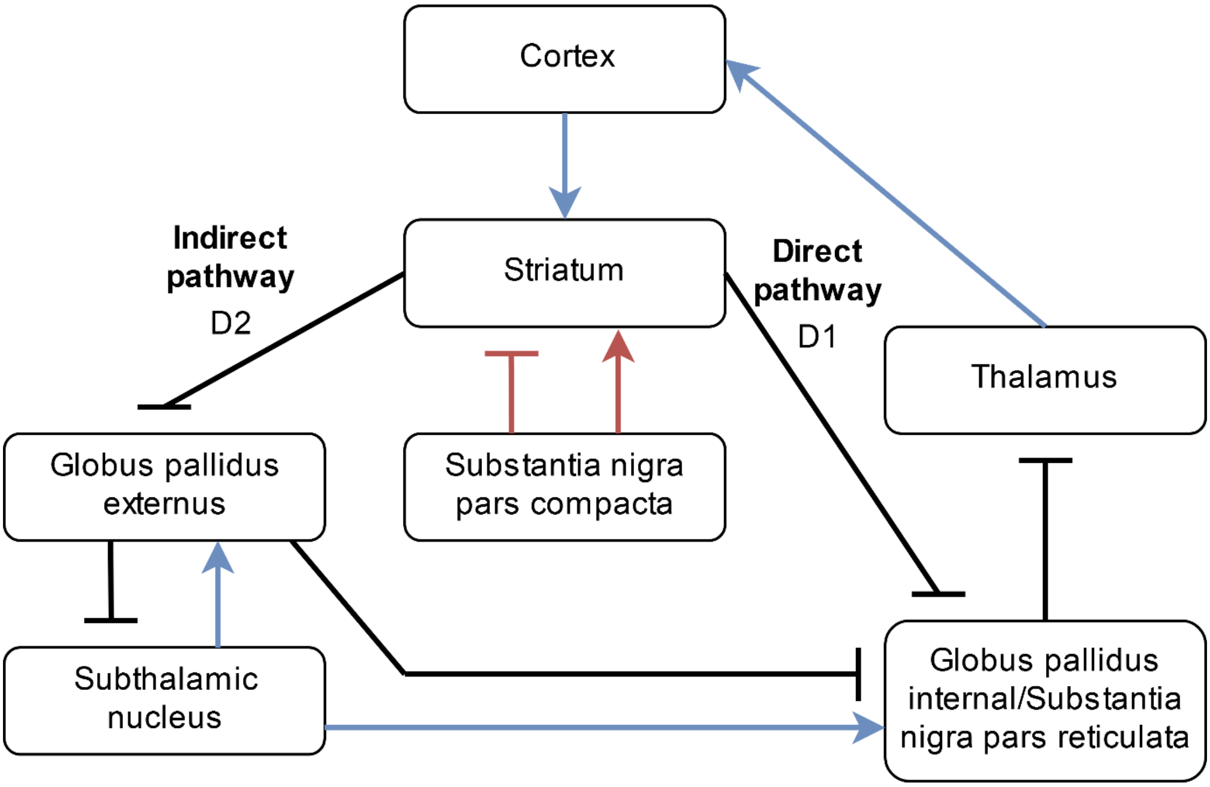


Figure 1.2: The normal function of the basal ganglia and the striatum. Indicated is a schematic overview of the corticostriatal loop in a healthy individual. Red arrows are dopaminergic; blue arrows are glutamatergic; black arrows are GABAminergic. Excitatory signals are shown by lines with arrow heads, whereas blunted, flat arrows show inhibitory signals. D1: dopamine receptor type 1; D2: dopamine receptor type 2. Loosely adapted from (Sgroi and Tonini, 2018).

1.4.2 Dysfunction of the striatum in HD

As discussed in 1.4.1, the primary neurodegenerative consequence of HD is striatal degradation. More specifically, the striatal medium spiny projection neurons (MSNs) are the principal cell population affected by HD neurodegeneration (de la Monte et al., 1988; Aylward et al., 1998; Vonsattel and DiFiglia, 1998), although more widespread cellular loss probably also contributes towards disease (see 1.4.3). Importantly, two sub-populations of MSNs can be distinguished: (1) striatonigral MSNs expressing dopamine type 1 receptors (D1), substance P and dynorphin and (2) striatopallidal MSNs expressing dopamine type 2 receptors (D2) and enkephalin (Gerfen, 1992; Gertler et al., 2008; Bunner and Rebec, 2016). Although the direct and indirect pathways are not completely separated (Haber et al., 2000; Cui et al., 2013), the direct corticostriatal pathway is mostly associated with D1 MSNs, whereas the indirect pathway is mostly associated with D2 MSNs (Surmeier et al., 2007).

In HD, striatal neurodegeneration preferentially affects the indirect pathway in early disease (Deng et al., 2004; Starr et al., 2008), as D2 striatopallidal MSNs undergo more marked neurodegeneration than D1 striatonigral MSNs (Reiner et al., 1988; Albin et al., 1992; Richfield et al., 1995; Augood et al., 1996; Wilson et al., 2017; Niccolini et al., 2018). Consequently, loss of D2 striatopallidal MSNs leads to repression of the indirect pathway and (relative) overactivation of the direct pathway (see Fig. 1.3). In this paradigm, overactivation of the direct pathway inhibits the globus pallidus internal and pars reticula of the substantia nigra, whereas the subthalamic nucleus is instead inhibited. Dis-inhibition of the thalamus promotes motoric activity, and contributes towards the development of hyperkinetic motor symptoms (*e.g.* chorea) seen in HD (Bateup et al., 2010).

Later in disease, substantial loss of both D1 and D2 MSNs occurs (Glass et al., 2000; Deng et al., 2004), subsequently resulting in dysfunction of both direct and indirect pathways. The dysfunction of both striatal pathways probably leads to the apparent hypokinetic phenotypes more often seen in advanced HD (Bateup et al., 2010). As well as affecting the direct and indirect corticostriatal pathways, striatal degradation probably also affects other brain circuitry. The nucleus accumbens, part of the ventral striatum, is part of the mesolimbic pathway and has roles in motivation and reward, for instance (reviewed by (Klawonn and Malenka, 2018)). Dysfunction of the limbic systems may, then, play a part in HD neuropathology (Mehrabi et al., 2016). The limbic systems help regulate memory, mood and emotion (reviewed by (Rolls, 2015)), and their improper functioning could precipitate some of the cognitive or behavioural symptoms experienced by HD patients.

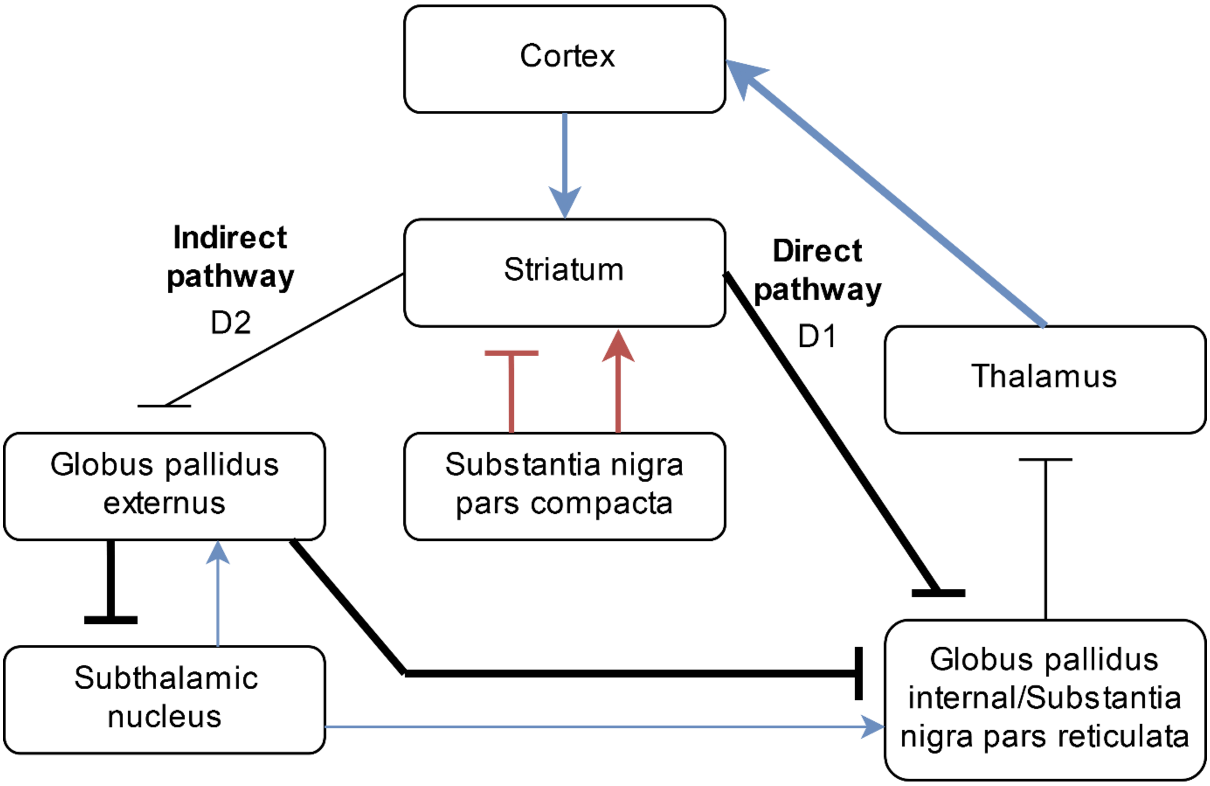


Figure 1.3: Dysfunction of the basal ganglia and striatum in early HD. Indicated is a schematic overview of the corticostriatal loop in an individual with HD. Over activation of the direct pathway leads to suppression of inhibition in the thalamus. Red arrows are dopaminergic; blue arrows are glutamatergic; black arrows are GABAminergic. Excitatory signals are shown by lines with arrow heads, whereas blunted, flat arrows show inhibitory signals. D1: dopamine receptor type 1; D2: dopamine receptor type 2. Loosely adapted from (Sgroi and Tonini, 2018).

1.4.3 Other neuropathology

Although the striatum experiences the highest degree of neurodegeneration in HD, other cell types are also vulnerable. For instance, the Vonsattel grading system is commonly used to score post-mortem HD brains between 0-4 depending on the level of macroscopic neurodegeneration (Vonsattel et al., 1985). Higher grades (3-4) in more advanced HD are associated with both striatal and cortical neurodegeneration (Vonsattel et al., 1985; de la Monte et al., 1988; Rosas et al., 2002; Ruocco et al., 2008; Thu et al., 2010). Gross degeneration also occurs in other brain regions including the hypothalamus, cerebellum and other parts of the basal ganglia (reviewed in (Waldvogel et al., 2015)). Substantial evidence from neuroimaging studies indicate striatal and white matter loss is apparent in pre-manifest HD (Rosas et al., 2003; Reading et al., 2004; Ciarmiello et al., 2006; Tabrizi et al., 2009; Aylward et al., 2011; Klöppel et al., 2015; Wu et al., 2017), and some patient symptoms are associated with these measurable neuropathological changes (Tabrizi et al., 2009, 2013;

Langbehn et al., 2019). Thus, pathologically relevant neurodegeneration occurs ahead of clinical onset, and is probably the basis for many of the subtle pre-manifest symptoms HD patients experience.

1.5 Molecular pathogenesis

1.5.1 Wild-type Huntingtin

HTT is a large protein (348 kDa) found in both the nucleus and cytoplasm, and is structurally composed of two predominantly HEAT repeat-containing domains connected by an interlinking bridge (Guo et al., 2018). The HEAT motif (named after the proteins where these repeats are found: Huntingtin, elongation factor 3, protein phosphatase 2A and the Tor1 kinase in yeast) contains two alpha helices connected by a linking segment, and HEAT repeats are found in several regulatory and transport-associated proteins (Andrade and Bork, 1995). HTT is also subject to a variety of post-translational modifications such as phosphorylation, ubiquitylation and proteolytic cleavage by various enzymes (Ehrnhoefer et al., 2011). Although the wild-type functions of HTT are still not well characterised (reviewed by (Saudou and Humbert, 2016)), HTT appears to be multifaceted scaffold protein, with roles in vesicular trafficking (DiFiglia et al., 1995; Caviston et al., 2007, 2011; Colin et al., 2008), autophagy (Steffan, 2010; Martin et al., 2014; Wong and Holzbaur, 2014), transcription (Steffan et al., 2000; Dunah et al., 2002; Takano and Gusella, 2002) and DNA repair (Maiuri et al., 2017). The role of the polyglutamine repeat in native HTT is unclear. Importantly, HTT is required for normal development, as knockout of *Htt* (the mouse isoform of *HTT*) is embryonically lethal (Duyao et al., 1995; Nasir et al., 1995; Zeitlin et al., 1995).

1.5.2 The generation of toxic species in HD

HD is conventionally considered a mechanistic gain of function disease, wherein the expanded *HTT* CAG repeat confers toxicity that leads to HD. This is supported by evidence from HD mouse models where human transgenic N-terminal *HTT* fragments containing a large CAG repeat confer an HD phenotype (Mangiarini et al., 1996; Schilling et al., 1999; Bradford et al., 2009). Moreover, the strong relationship between CAG length and disease onset also suggests a primarily toxic gain of function (Andrew et al., 1993; Duyao et al., 1993; Snell et al., 1993). Haploinsufficiency of HTT (*i.e.* partial loss of HTT function) could nevertheless exacerbate or modify HD pathology (see (Saudou and Humbert, 2016)). Furthermore, there is some evidence of a dominant negative function (where mutated HTT interferes with normal HTT function), and this may serve as another contributory mechanism in HD (Elias et al., 2014; Molina-Calavita et al., 2014; Lopes et al., 2016; Ruzo et al., 2018).

As shown in Fig. 1.5, there are several ways by which toxic moieties are generated in HD. The classical view is once the encoded HTT polyglutamine segment reaches a critical pathogenically relevant size (~36-40 glutamines in HD), mutant HTT (mutHTT) can go on to affect many downstream biochemical and cellular systems (see 1.5.3 & Fig. 1.6). As well as the production of full-length HTT protein, the generation and subsequent aggregation of small N-terminal HTT fragments is characteristic of HD pathology (Davies et al., 1997; Mende-Mueller et al., 2001), and N-terminal HTT fragments appear to be especially cytotoxic (Mangiarini et al., 1996; Davies et al., 1997; Scherzinger et al., 1997; Barbaro et al., 2015). N-terminal fragments can arise through post-translational proteolytic cleavage of HTT *via* a diverse array of enzymes such as caspases (Goldberg et al., 1996; Wellington et al., 1998; Kim et al., 2001), calpains (Gafni and Ellerby, 2002; Gafni et al., 2004) and metalloproteinases (Miller et al., 2010). Splicing of *HTT* RNA has been implicated as an alternative mechanism by which small *HTT* exon 1-containing fragments may arise (Sathasivam et al., 2013; Neueder et al., 2017; Franich et al., 2019).

Apart from polyglutamine proteins, other toxic protein species may also arise in HD from repeat associated non-ATG (RAN) translation (Bañez-Coronel et al., 2015). In RAN translation, out of frame translation by ribosomal machinery of expanded *HTT* RNA leads to polyalanine and polyserine proteins derived from sense RNA, as well as polyleucine, polycysteine and polyalanine proteins from antisense RNA (Bañez-Coronel et al., 2015). How (or if) these repetitive proteins contribute to pathology in HD is unknown (reviewed by (Cleary et al., 2018)), although RAN proteins were found in proteinaceous aggregates in post mortem HD brains (Bañez-Coronel et al., 2015).

In addition to protein-mediated gain of function, toxic RNA species are also implicated as a mechanism in HD. RNA containing CAG repeats can form secondary hairpin/stem-loop structures (Fig. 1.4) (Sobczak et al., 2003; Kiliszek et al., 2010; de Mezer et al., 2011; Yildirim et al., 2013; Tawani and Kumar, 2015). *HTT* mRNA with pathogenic CAG hairpins can exert toxicity by the sequestration of RNA-binding proteins including MBNL1 (muscleblind-like protein 1) and other spliceosome pathway components (de Mezer et al., 2011; Mykowska et al., 2011; Urbanek et al., 2016; Schilling et al., 2019). Dysregulation of cellular splicing machinery may contribute towards pathogenesis (Mykowska et al., 2011). CAG repeat-containing RNA can also interfere with other RNA metabolism such as Dicer and the RNA-induced silencing complex (RISC) pathway (Bañez-Coronel et al., 2012).

Notably, although *HTT* DNA isn’t directly toxic itself, all toxic species (protein or RNA) originate from the DNA. Like RNA, repetitive CAG sequence in DNA can form secondary hairpins (Gacy et al., 1995; Mitas et al., 1995; Grabczyk and Usdin, 2000; Pearson et al., 2002; Napierala et al., 2005; Sobczak and Krzyzosiak, 2005; Liu et al., 2010a). DNA hairpins can act as an interface to which DNA repair components may bind (Owen et al., 2005; Burdova et al., 2015; Guo et al., 2016), and mishandling by DNA repair machinery may lead to somatic CAG repeat expansion as indicated in Fig. 1.5 (this concept is elaborated more fully in 1.7 & 6.3.3). Somatic instability may act as a master toxicity regulator of sorts in HD, with successively longer CAG repeats producing more toxic downstream RNA and protein components (see also 1.7), a concept discussed by Massey and Jones (Massey and Jones, 2018).

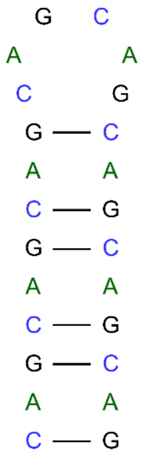


Figure 1.4: A repetitive hairpin/stem-loop in RNA/DNA. Shown is an example hairpin/stem-loop. This may arise in long, repetitive CAG tracts in either DNA or RNA (and can also occur in other repeats such as CUG repeats in myotonic dystrophy type 1 RNA). In the figure, CAG sequences bind to their complementary GAC sequence on the opposite repetitive sequence, with hydrogen bonds shown by the black lines.

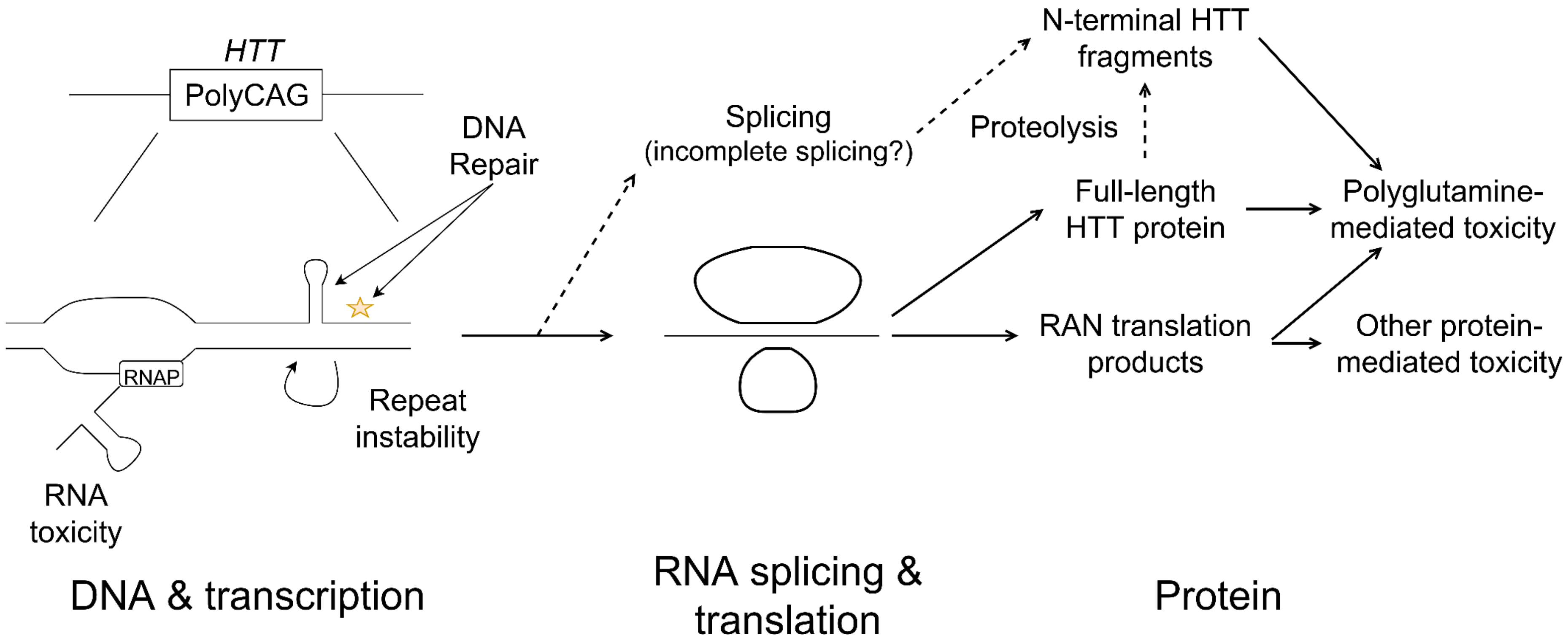


Figure 1.5: Mechanisms by which toxic species are generated in HD. Indicated are the ways in which an expanded *HTT* CAG (shown as PolyCAG) can generate toxic species that lead to disease (for downstream mechanisms, see Fig. 1.6). On the left (DNA & transcription), a section of CAG in *HTT* is being transcribed by RNA polymerase (RNAP). Here, RNA could form secondary structures that are toxic (these structures can also arise once fully transcribed). Additionally, mishandling by DNA repair could exacerbate pathology by increasing repeat size *via* somatic instability (see 1.7 & 6.3.3). *HTT* transcripts are then translated, and (incomplete) splicing may alter the size of the mature mRNA. Finally, RNA is translated to protein. Full-length or toxic N-terminal HTT fragments lead to toxicity downstream, although possibly in different ways. Proteolytic cleavage of full-length HTT may also occur. Finally, repeat-associated non-AUG (RAN) translation may result in additional protein-mediated toxicity.

1.5.3 Downstream cellular pathology

The molecular mechanisms governing HD pathogenesis are complex, poorly understood and probably multimodal, with both gain of function and loss of function mechanisms implicated. A vast array of cellular machinery is disrupted in HD cells, and trying to pinpoint the most relevant driving mechanism(s) has been difficult (*i.e.* determining pathogenic cause and effect) – see 1.7 for an overview of the genetic work trying to identify and prioritise the most relevant disease-associated pathways. Here, I will just give a brief overview of some of the more prominent proposed HD mechanisms for background, and a number of these are illustrated in Fig. 1.6. Recent reviews give more detail about the downstream pathology which occurs in HD (Sepers and Raymond, 2014; Bates et al., 2015; Saudou and Humbert, 2016; Adegbuyiro et al., 2017; Jimenez-Sanchez et al., 2017; Jodeiri Farshbaf and Ghaedi, 2017; Raymond, 2017; Dickey and La Spada, 2018; Lieberman et al., 2019).

One of the more striking features of HD pathology is the formation of proteinaceous inclusions in the brain (Davies et al., 1997; Scherzinger et al., 1997), also found in other polyglutamine diseases. Protein inclusions can be nuclear or perinuclear (cytoplasmic) in origin and arise from the misfolding of monomeric polyglutamine fragments to form oligomeric species. These oligomers can form aggregates and eventually large protein inclusions incorporating other proteins (reviewed in detail by (Adegbuyiro et al., 2017)), and N-terminal HTT fragments readily form aggregates (Scherzinger et al., 1997). The effect of HTT aggregation on disease remains uncertain, however, with both protective (Gutekunst et al., 1999; Kuemmerle et al., 1999; Arrasate et al., 2004) and neurotoxic (Davies et al., 1997; Liu et al., 2015; Woerner et al., 2016; Bäuerlein et al., 2017) effects being reported. It is possible protein aggregates could exert both protective and toxic properties, and this may be context or disease-stage relevant (Tallaksen-Greene et al., 2003). For instance, early in pathology aggregation may help to clear toxic polyglutamine proteins; however, later in disease aberrant sequestration of other proteins, overloading of proteasomal pathways or autophagy dysfunction could contribute to cell death (see reviews by (Ortega and Lucas, 2014; Martin et al., 2015; Croce and Yamamoto, 2019; Lieberman et al., 2019)).

Axonal vesicle transport is also impaired in HD cells, possibly through a combination of HTT loss of function and dominant negativity. HTT helps facilitate the vesicle trafficking of numerous cellular components (detailed by (Saudou and Humbert, 2016)) including organelles (Caviston et al., 2011; Liot et al., 2013; Wong and Holzbaur, 2014), GABA receptors (Twelvetrees et al., 2010) and brain-derived neurotrophic factor (BDNF) (Gauthier et al., 2004), a neurotrophin involved in the neuronal survival and maintenance. HTT also helps regulate BDNF (and other gene) transcription through the binding of various transcription factors (Zuccato et al., 2001, 2003). Altered delivery and transcription of BDNF and other cell survival proteins may contribute to the neuronal vulnerability in HD. Mitochondrial dysfunction also occurs in HD, and may be exacerbated by reduced BDNF input (see review by (Jodeiri Farshbaf and Ghaedi, 2017)).

Finally, excitotoxicity has been suggested as a mechanism by which striatal cells could die in HD. In excitotoxicity, hyperactivity of glutamate-binding *N*-methyl-D-aspartate (NMDA) receptors leads to gross calcium ion dyshomeostasis and subsequent cellular demise (see (Sepers and Raymond, 2014; Raymond, 2017)). Impaired glutamate uptake by astrocytes may enhance an excitotoxic phenotype (Faideau et al., 2010; Wood et al., 2018), and aberrant activity by reactive glial astrocytes (and possibly other glia) could further contribute to neurotoxicity (Liddelow et al., 2017). It is not clear which mechanisms are leading to cell death in HD patients, however, and multiple mechanisms of cell death may be operating at different points in disease.

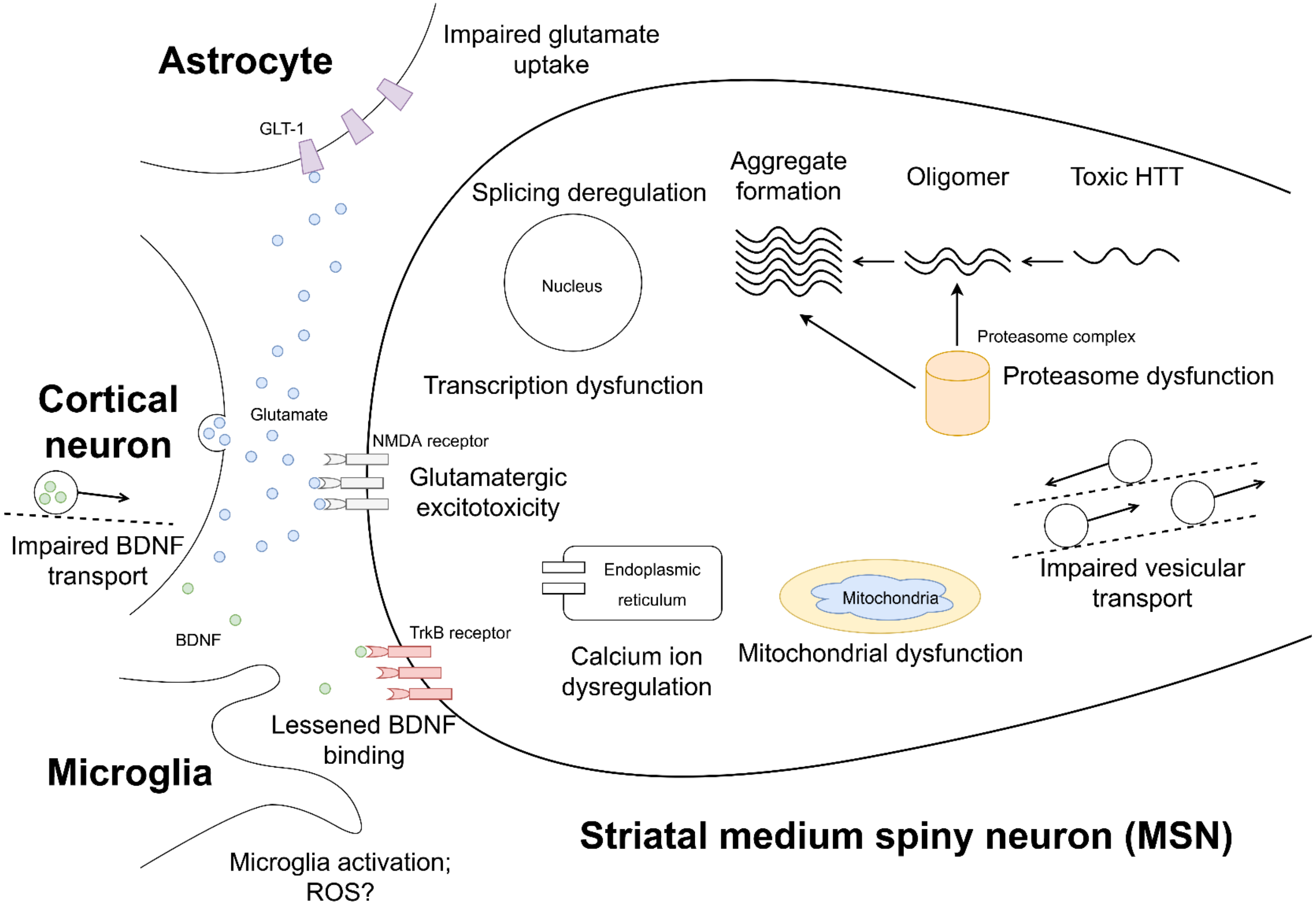


Figure 1.6: Downstream molecular and cellular dysfunction in HD. Shown are some of the pathogenic mechanisms by which HD may lead to cell dysfunction and death. Abbreviations: BDNF: brain-derived neurotrophic factor; GLT-1: Glutamate transporter 1; NMDA receptor: *N*-methyl-D-aspartate receptor (binds glutamate); ROS: Reactive oxygen species; TrkB receptor: Tropomyosin receptor kinase B (binds BDNF).

1.6 Other repeat diseases

HD is one of nine polyglutamine diseases, all of which are caused by exonic CAG repeat expansions (Table 1.1). Although the repeat expansions occur in different genes, all the polyglutamine diseases are neurodegenerations that preferentially atrophy different regions of the brain or motor neurons. Eight of the nine polyglutamine disorders are autosomal dominant, with the exception being the X-linked recessive spinal and bulbar muscular atrophy (SBMA) which only presents in males (heterozygous female carriers are usually asymptomatic (Mariotti et al., 2000)). CAG-driven gain of function pathology is thought to be an important mechanism in all the polyglutamine diseases, although both dominant negativity and/or loss of function mechanisms may contribute to the distinct neuropathological and clinical phenotypes seen (the polyglutamine diseases are reviewed generally by (Lieberman et al., 2019)). All the polyglutamine diseases show clinical anticipation, *i.e.* the children of affected parents are likely to have earlier ages of disease onset caused by intergenerational CAG repeat expansion (Lieberman et al., 2019). As with HD, the other polyglutamine diseases are rare, and the epidemiology varies depending on ancestry. For instance, Ruano et al. estimates a worldwide prevalence of the dominant cerebellar ataxias (which includes some non-polyglutamine ataxias) to be 2.7 per 100,000 (Ruano et al., 2014).

The polyglutamine diseases are part of a larger group of repeat disorders, encompassing >30 diseases (see Table 1.2). The coding polyalanine repeat diseases have been omitted (excepting oculopharyngeal muscular dystrophy (OPMD)) from Table 1.2 as their aetiology is likely to be different; these are considered elsewhere (Brown and Brown, 2004). The majority of the non-glutamine repeat diseases occur in non-coding regions of the genome, and various repeat expansion species have been identified as potentially pathogenic. Most are trinucleotide repeats (CAG, CTG and CGG and related repeats), but more complex quadri-, penta-, hexa- and even dodecanucleotide repeats have been described. Among these diseases are Fragile-X syndrome (FRAXA), the most common cause of inherited mental disability, Fuch's endothelial dystrophy type 3 (FECD3), a partially penetrant repeat disease that causes corneal dystrophy and visual impairment, and a subtype of amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD) caused by an expansion mutation in *C9orf72*. Hexanucleotide expansions in *C9orf72* account for ~40% of familial ALS, ~25% of familial FTD, and between ~3-15% of sporadic ALS/FTD cases (Renton et al., 2011; Pliner et al., 2014). Almost all repeat diseases are associated with a neurological phenotype of some kind.

The mechanisms by which non-coding repeat diseases cause pathology are diverse. Silencing (loss of function) of genes occurs in autosomal recessive repeat diseases such as Friedrich's ataxia and glutaminase deficiency (GLD) (van Kuilenburg et al., 2019). Gene silencing also occurs in Fragile-X syndrome where large repeats (>200 CGGs) in the X-linked *FMR1* gene lead to hypermethylation and loss of the important FMRP protein in males (Verkerk et al., 1991; Coffee et al., 1999). Similar to HD, RNA toxic gain of function mechanisms are prominent in several repeat diseases such as in the myotonic dystrophies (DM1 and DM2) (reviewed in (Sznajder and Swanson, 2019)). Repetitive RNA can form hairpins and hence sequester RNA binding and processing proteins including muscleblind-like protein family members (Timchenko et al., 2001; Fardaei et al., 2002), leading to widespread splicing dysregulation.

One of the more intriguing mechanisms in non-coding repeat disease is repeat associated non-ATG (RAN) translation (although this may occur in coding repeat disorders too, as it does in HD). Here, even normally intronic RNA containing large repeats may be translated to form a variety of in- and out-of-register repetitive protein species. RAN translation was originally identified in spinal cerebellar ataxia (SCA) type 8 (Zu et al., 2011), but has since been implicated across a range of repeat diseases including HD (Bañez-Coronel et al., 2015), *C9orf72* ALS/FTD (Mori et al., 2013), DM1 (Zu et al., 2011), DM2 (Zu et al., 2017) and others (see review by (Cleary et al., 2018)). As with HD, the degree to which RAN translation contributes towards repeat disease pathology is yet unclear; however, through RAN translation, normally non-coding DNA could exert pathogenic effects through toxic protein species, blurring the line between traditionally coding and non-coding repeat diseases.

Disease	Repeat	Normal	Pathogenic	Locus	Gene	Gene symbol	Location in gene	Reference
HD	CAG	6-35	40->200	4p16.3	*Huntingtin*	*HTT*	Exon 1	(The Huntington's Disease Collaborative Research Group, 1993)
DRPLA	CAG	6-35	48-93	12p13	*Atrophin 1*	*ATN1*	Exon 5	(Koide et al., 1994)
SBMA	CAG	11-34	38-68	Xq11-q12	*Androgen receptor*	*AR*	Exon 1	(La Spada et al., 1991)
SCA1	CAG	6-35	39-51	6p23	*Ataxin 1*	*ATXN1*	Exon 8	(Orr et al., 1993)
SCA2	CAG	13-31	35->200	12q24	*Ataxin 2*	*ATXN2*	Exon 1	(Sanpei et al., 1996)
SCA3	CAG	12-44	60-87	14q24-q31	*Ataxin 3*	*ATXN3*	Exon 10	(Stevanin et al., 1994)
SCA6	CAG	4-18	20-33	19p13	*CaV2.1*	*CACNA1A*	Exon 47	(Jodice et al., 1997)
SCA7	CAG	10-27	36->400	3p21-p12	*Ataxin 7*	*ATXN7*	Exon 3	(David et al., 1998)
SCA17	CAG	25-40	49-66	6p27	*TATA binding protein*	*TBP*	Exon 3	(Nakamura et al., 2001)

Table 1.1: The polyglutamine repeat diseases. Indicated are the known polyglutamine (polyQ) diseases. Note these all occur in protein coding exons. The pathogenic repeat range shows alleles with full (or very high) disease penetrance; smaller expanded repeats (*e.g.* 36-39 for HD) between the normal and pathogenic ranges (reduced penetrance alleles) may still cause disease in some individuals. HD: Huntington's disease; DRPLA: Dentatorubral-pallidoluysian atrophy (also known as Haw-River syndrome); SBMA: Spinobulbar muscular atrophy (also known as Kennedy's disease); SCA: Spinal cerebellar ataxia. SCA3 is also known as Machado-Joseph disease, and SCA17 as Huntington disease-like syndrome 4 (HDL4).

Disease	Repeat	Locus	Gene	Gene symbol	Location in gene	Inheritance modality	Reference(s)
ALS/FTD	GGGGCC	9p21.2	*Chromosome 9 open reading frame 72*	*C9orf72*	Intron	AD	(DeJesus-Hernandez et al., 2011; Renton et al., 2011)
BAFME	TTTCA/ TTTTA	8q24.11-q24.12	*Sterile alpha motif domain containing 12*	*SAMD12**	Intron	AD	(Ishiura et al., 2018; Mizuguchi et al., 2019)
BSS	GGC	16p12.3	*Xylosyltransferase 1*	*XYLT1*	5'-UTR	AR	(LaCroix et al., 2019)
CANVAS	AAGGG	4p14	*Replication factor C subunit 1*	*RFC1*	Intron	AR	(Cortese et al., 2019; Rafehi et al., 2019)
DM1	CTG	19q13.32	*Myotonic dystrophin kinase*	*DMPK*	3'-UTR	AD	(Brook et al., 1992; Mahadevan et al., 1992)
DM2	CCTG	3q21.3	*CCHC-type zinc finger nucleic acid binding protein*	*CNBP*	Intron	AD	(Liquori et al., 2001)
EPM1A	CCCCGC CCCGCG	21q22.3	*Cystatin B*	*CSTB*	5'-UTR	AR	(Lalioti et al., 1997)
FECD3	TGC	18q21.2	*Transcription factor 4*	*TCF4*	Intron	AD	(Wieben et al., 2012; Mootha et al., 2014)
FRDA	GAA	9q21.11	*Frataxin*	*FXN*	Intron	AR	(Campuzano et al., 1996)
FRAXA/ FXTAS/ FXPOI	CGG	Xq27.3	*Fragile X mental retardation 1*	*FMR1*	5'-UTR	X-LD	(Fu et al., 1991; Verkerk et al., 1991)
FRAXE	CCG	X28	*Fragile X mental retardation 2*	*FMR2*	5'-UTR	X-LD	(Knight et al., 1993)
FRAXF	GCC	X28	*Transmembrane protein 185A*	*TMEM185A*	5'-UTR	X-LD	(Parrish et al., 1994)
FRA2A	CGG	2q11	*AF4/FMR2 family member 3*	*AFF3*	Intron	AD	(Metsu et al., 2014b)
FRA7A	CGG	7p11.2	*Zinc finger protein 713*	*ZNF713*	Intron	AD	(Metsu et al., 2014a)
FRA11A	CGG	11q13	*Chromosome 11 open reading frame 80*	*C11orf80*	5'-UTR	AD	(Debacker et al., 2007)
FRA12A	CGG	12q13.12	*Disco interacting protein 2 homolog B*	*DIP2B*	5'-UTR	AD	(Winnepenninckx et al., 2007)
GLD	GCA	2q32.2	*Glutaminase*	*GLS*	5'-UTR	AR	(van Kuilenburg et al., 2019)

Disease	Repeat	Locus	Gene	Gene symbol	Location in gene	Inheritance modality	Reference(s)
HDL2	CTG	16q24.2	*Junctophilin 3*	*JPH3*	Exon 2A	AD	(Holmes et al., 2001)
NIID(RD)	GGC	1q21.2	*Notch 2 N-terminal like C*	*NOTCH2NLC*	5'-UTR	AD	(Okubo et al., 2019; Sone et al., 2019; Tian et al., 2019)
OPMD	GCG	14q11.2	*Poly(A) binding protein nuclear 1*	*PABPN1*	Exon 1	AD	(Brais et al., 1998)
RCPS	CGCA	17q25.3	*Eukaryotic translation initiation factor 4A3*	*EIF4A3*	5'-UTR	AR	(Favaro et al., 2014)
SCA8	CTG	13q21.33	*Ataxin 8 opposite strand*	*ATXN8OS*	3'-UTR	AD	(Koob et al., 1999)
SCA10	ATTCT	22q13.31	*Ataxin 10*	*ATXN10*	Intron	AD	(Matsuura et al., 2000)
SCA12	CAG	5q32	*Protein phosphatase 2 regulatory subunit beta*	*PPP2R2B*	5'-UTR	AD	(Holmes et al., 1999)
SCA31	TGGAA	16q21	*Brain-expressed associated with NEDD4*	*BEAN*	Intron	AD	(Sato et al., 2009)
SCA36	GGCCTG	20p13	*Nucleolar protein 56*	*NOP56*	Intron	AD	(Kobayashi et al., 2011)
SCA37	ATTTC	1p32.2	*DAB adaptor protein*	*DAB1*	Intron	AD	(Seixas et al., 2017)

Table 1.2: The non-coding repeat diseases and disease-associated repetitive fragile sites. Note this table is not exhaustive; see (Brown and Brown, 2004) and (Sznajder and Swanson, 2019) for more non-canonical repeat diseases (primarily polyalanine insertions). The OPMD repeat is coding. In BAFME (*), repeat expansions were also reported in *TNRC6A* and *RAPGEF2* (Ishiura et al., 2018). Disease abbreviations: ALS/FTD: Amyotrophic lateral sclerosis/Frontotemporal dementia; BAFME: Benign adult familial myoclonic epilepsy; BSS: Baratela-Scott syndrome; CANVAS: Cerebellar ataxia, neuropathy, vestibular areflexia syndrome; DM1: Myotonic dystrophy (DM) type 1; DM2: DM type 2; EPM1A: Epilepsy, progressive myoclonic type 1A (also known as Unverricht-Lundborg disease); FECD3: Fuch's endothelial dystrophy type 3; FRDA: Fredrich's ataxia; FRAXA: Fragile-X syndrome; FXTAS: Fragile-X associated tremor/ataxia syndrome; FXPOI: Fragile-X associated primary ovarian insufficiency; FRAXE: Fragile-X E syndrome; FRAXF: Fragile-X F syndrome; FRA2A: Folate-sensitive fragile site (FSFS) 2A; FRA7A: FSFS 7A; FRA11A: FSFS 11A; FRA12A: FSFS 12A; GLD: Glutaminase deficiency; HDL2: Huntington disease-like 2; NIID(RD): Neuronal inclusion disease (and related disorders); OPMD: Oculopharyngeal muscular dystrophy; RCPS: Richieri-Costa-Pereira syndrome; SCA: Spinal cerebellar ataxia (SCA). Inheritance abbreviations: AD: Autosomal dominant; AR: Autosomal recessive; X-LD: X-linked dominant.

1.7 Genetic modifiers of Huntington's disease

1.7.1 Human genetic studies

A significant portion of HD age at onset is governed by the length of the *HTT* CAG repeat (~50-60%) in an inverse fashion, with earlier onset occurring at larger repeat lengths (Andrew et al., 1993; Duyao et al., 1993; Wexler et al., 2004) (see Fig. 3.4C later), leaving ~40-50% of onset unexplained. Hence two HD individuals with identical *HTT* CAG lengths could have disease onset as much as 40-50 years apart from each other. Evidence from the Venezuelan kindred study (Wexler et al., 2004) showed that ~40% of the remaining variation in HD onset is heritable, implicating other genetic factors are responsible for a substantial portion of HD onset and therefore disease pathogenesis (Duyao et al., 1993; Wexler et al., 2004; Andresen et al., 2007b, 2007a). Given the vast array of cellular machinery that is dysregulated in HD cells (see 1.5), genetic modifiers of HD can help narrow down which pathway(s) are the most critical in disease pathogenesis (Gusella and MacDonald, 2009; Gusella et al., 2014). Elucidating important disease components and mechanisms may then indicate therapeutically relevant targets.

Early human genetic studies adopted a candidate gene approach, wherein genes thought to be involved in HD pathogenic mechanisms were investigated in patients (Gusella and MacDonald, 2009; Zuccato et al., 2010). These pioneering human genetic studies implicated genes such as *APOE* (lipid metabolism) (Kehoe et al., 1999), *TCERG1* (transcriptional/splicing regulation) (Holbert et al., 2001), *TP53* (transcriptional and other regulation) (Chattopadhyay et al., 2005), *GRIK2* (glutamatergic synaptic transmission) (Rubinsztein et al., 1997; MacDonald et al., 1999; Cannella et al., 2004), *PPARGC1A* (energy metabolism) (Taherzadeh-Fard et al., 2009; Weydt et al., 2009) and *HAP1* (HTT interaction and vesicular transport) (Metzger et al., 2008), among others (see (Gusella and MacDonald, 2009)). But many of these initial findings failed to replicate in further study (Saft et al., 2004; Arning et al., 2005; Andresen et al., 2007a; Lee et al., 2012a; Ramos et al., 2012b). The reasons for these somewhat contradictory findings in early human genetic work are discussed by Gusella et al. (Gusella et al., 2014), but are predominantly due to small cohort sizes, population stratification (*i.e.* patient ethnicity) not being properly accounted for and multiple testing correction issues. Many of the same problems have been experienced more broadly across human genetic studies of the era (Gusella et al., 2014).

It has not been until more recently that modern and unbiased SNP-based approaches have been utilised to find modifiers of HD in larger cohorts of human patients. The first GeM-HD consortium genome-wide association study (GWAS) identified the locus containing *FAN1*, a

DNA repair gene involved in interstrand crosslink repair, as significant for HD onset modification (GeM-HD Consortium, 2015). *MLH1*, a mismatch repair protein, was also found as nominally significant (GeM-HD Consortium, 2015), and confirmed as genome-wide significant in a follow-up study shortly after (Lee et al., 2017). The locus containing *RRM2B*, a nucleotide scavenging gene, was also found as nominally significant (GeM-HD Consortium, 2015). The findings from the first GeM-HD GWAS had a profound effect on the HD field, highlighting that DNA repair was a critical system by which pathogenesis was mediated in HD. Following the first GeM-HD study, a relatively small GWAS in a deeply phenotyped HD cohort, TRACK-HD, identified a further DNA repair gene as significant modifying the clinical trajectory of HD patients, *MSH3* (Hensman Moss et al., 2017), and this finding was replicated in the recent Flower et al. study (Flower et al., 2019). There is also some evidence similar DNA repair mechanisms may be shared between multiple repeat diseases (Bettencourt et al., 2016), although there is currently lacking an equivalent GWAS in these other repeat diseases to confirm more conclusively.

A second, larger GWAS from the GeM-HD consortium was recently published (GeM-HD Consortium, 2019), which highlighted several more loci containing DNA repair genes as HD onset modifying, including *MSH3*, *PMS1*, *PMS2* and *LIG1*. The study also confirmed the *RRM2B* locus as genome-wide significant, as well as the *TCERG1* locus (GeM-HD Consortium, 2019). So far, *TGERG1* is the only gene from earlier candidate gene study which has been confirmed using an unbiased SNP imputation approach (Holbert et al., 2001). The *HTT* allele structure was also tagged as onset modifying by the GeM-HD study and others very recently at the time of writing this introduction (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). These data are considered in the context of the current thesis in substantially more detail in chapters 5 and 6. Table 1.3 gives an overview of the current candidate modifiers of HD from human genetics and model systems (see also the next section, 1.7.2).

1.7.2 Functional work in model systems

DNA repair is thought to mechanistically alter somatic instability, although other interaction mechanisms have been proposed (Maiuri et al., 2019). As touched on in 1.5.2, somatic repeat expansion may act to accelerate the production of toxic species in HD (discussed in (Massey and Jones, 2018)). Indeed, shortly after the *HTT* was discovered, somatic instability was reported in the brain of HD patients (Telenius et al., 1994), and instability occurs prominently in the striatum and other disease relevant tissues (Kennedy et al., 2003; Shelbourne et al., 2007). Increased somatic instability has been associated with earlier disease onset in post mortem HD brains (Swami et al., 2009). Critically, this repeat instability occurs in terminally differentiated postmitotic neurons, thereby implicating DNA maintenance and repair, not replicative machinery, as underlying somatic expansion (Shelbourne et al., 2007; Gonitel et al., 2008).

In hindsight, many of the DNA repair genes and pathways implicated by GWAS have been examined by repeat disease animal work over a decade prior (Table 1.3). Knockout of the mismatch repair gene *Msh2* ablated CAG somatic instability in HD mice (Manley et al., 1999; Wheeler et al., 2003), and reduction of *Msh3* had a similar effect (Dragileva et al., 2009; Tomé et al., 2013). *Mlh1* and *Mlh3* were also found to be necessary for repeat expansion in HD mice (Pinto et al., 2013), and *Pms2* in myotonic dystrophy type 1 mice (Gomes-Pereira et al., 2004). Importantly, the slowing of somatic instability has been shown to delay HD pathology in mice (Wheeler et al., 2003; Kovalenko et al., 2012; Budworth et al., 2015). There is also some evidence from HD cells and mice that base excision repair glycosylases (OGG1 and NEIL1) may be involved in somatic instability (Kovtun et al., 2009; Møllersen et al., 2012; Budworth et al., 2015). More recently *FAN1*, the top candidate gene from GWAS (GeM-HD Consortium, 2015, 2019), was found to protect against repeat instability in an HD cell model (Goold et al., 2019) and a fragile X mouse model (Zhao and Usdin, 2018).

As shown in Table 1.4, the primary pathway implicated in HD disease modification is the mismatch repair pathway. Mismatch repair canonically repairs small mismatches in DNA; see (Jiricny, 2006; Hsieh and Zhang, 2017) for reviews. Briefly, DNA mismatches are recognised by one of two MutS homologs, MutSα (MSH2-MSH6) or MutSβ (MSH2-MSH3). Recruitment of one of the MutL homologs (usually MutLα (MLH1-PMS2) in post replicative repair) then follows, and MutL introduces a nick near the site of damage. EXO1 then removes the damaged base region *via* its exonuclease activity. Polδ repairs the gap left by EXO1, and the DNA is re-ligated by LIG1. Improper repair by this system may lead to somatic repeat instability (see Fig. 6.3 and 6.3.3 later). Additionally, elements of inter-strand crosslink repair could be involved in somatic instability, given the signal seen in *FAN1* through GWAS (GeM-HD Consortium, 2015, 2019; Lee et al., 2017). It is also possible elements of base excision repair could have a role in HD somatic mosaicism (reviewed by (Polyzos and McMurray, 2017)), however, currently base excision repair components have not been found as modifiers of disease in HD patients (excepting *LIG1* which is shared between multiple pathways). As there is considerable crosstalk between DNA repair pathways, these systems may not be entirely separate (see 6.3.3 for a model exploring this).

Candidate gene	DNA repair?	Function of encoded protein	Evidence from model systems	Evidence from human genetics
FAN1	Yes	Endo/Exonuclease in inter-strand crosslink repair	(Zhao and Usdin, 2018)† (Goold et al., 2019)	(GeM-HD Consortium, 2015, 2019; Bettencourt et al., 2016; Lee et al., 2017)
LIG1	Yes	Ligase, re-ligates DNA during various repair pathways	(Tomé et al., 2011)*	(GeM-HD Consortium, 2019)
MLH1	Yes	Endonuclease in mismatch repair (MutLα/ MutLβ/MutLγ complexes)	(Pinto et al., 2013)	(Lee et al., 2017)
MLH3	Yes	Endonuclease in meiosis? (MutLγ complex)	(Pinto et al., 2013) (Zhao et al., 2018) †	
MSH2	Yes	Mismatch recognition (MutSα/MutSβ)	(Manley et al., 1999; Wheeler et al., 2003; Seriola et al., 2011; Kovalenko et al., 2012)	
MSH3	Yes	Mismatch recognition (MutSβ)	(Dragileva et al., 2009; Tomé et al., 2013)	(Hensman Moss et al., 2017; Flower et al., 2019; GeM-HD Consortium, 2019)
MSH6	Yes	Mismatch recognition (MutSα)	(Dragileva et al., 2009; Kantartzis et al., 2012; Nakatani et al., 2015)	
NEIL1	Yes	DNA glycosylase in base excision repair	(Møllersen et al., 2012)	
OGG1	Yes	DNA glycosylase in base excision repair	(Kovtun et al., 2009; Budworth et al., 2015)	
PMS1	Yes	Unknown, but complexes with MLH1 (MutLβ)		(GeM-HD Consortium, 2019)
PMS2	Yes	Endonuclease (MutLγ complex)	(Gomes-Pereira et al., 2004)*	(Bettencourt et al., 2016; GeM-HD Consortium, 2019)
RRM2B (*UBR5*?)	No	Nucleotide scavenging		(GeM-HD Consortium, 2015, 2019)
TCERG1	No	Transcription and splicing regulation	(Arango et al., 2006)	(Holbert et al., 2001; GeM-HD Consortium, 2019)

Table 1.3: Candidate genetic modifiers of Huntington's disease. Shown are candidate HD modifiers from model systems (cells or animals) or humans. Studies marked by a (*) are from a myotonic dystrophy type 1 model and (†) from a fragile X model. Loci encompassing *SYT9*, *GSG1L* and *CCDC82* from (GeM-HD Consortium, 2019) and *SOSTDC1* from (Chao et al., 2018) are not included as the genes driving these signals are not confirmed. *RRM2B* and *UBR5* are both given as it is somewhat unclear which is driving the signal seen in GWAS (GeM-HD Consortium, 2019). See (Polyzos and McMurray, 2017) for a consideration of other base excision repair components that may also be involved in repeat disease (*e.g. FEN1*). The role of *HTT* allele structure is not indicated; see chapters 5 and 6 later for an in-depth discussion.

Pathway	Associated lesion	Damage sensors	Signalling/mediator proteins	Effector proteins	Effector polymerases	Effector ligases
a-EJ/MMEJ	DSB, broken DNA ends	PARP1	ATM?	CtIP, FEN1, MRN complex, XPF-ERCC1, XRCC1	Polθ	LIG1, LIG3
HR (SDSA and dHJ)	DSB, typically during DNA replication	MRN complex	ATM, ATR, MK2, CtIP, RPA, BRCA1-BARD1, BRCA2, PALB2	RAD51, RAD54, BLM, EXO1, FANCJ, GEN1, SMX complex, BTR complex	Polδ, Polε?	LIG1, LIG3
NHEJ	DSB, broken DNA ends (also in CSR and V(D)JR	Ku70, Ku80	DNA-PK	Artemis, APLF, WRN, PAXX, XLF	Polμ, Polλ, TdT	LIG4-XRCC4
SSA	DSB, mechanisms not well defined	MRN complex	CtIP	RAD52, ERCC1	?	?
ICL repair (canonical FA)	Interstrand crosslinks, thought to interface with SSB/DSB DNA repair machinery depending on damage context	FANCM	FA core complex, ATR	BRCA2, BRIP1, PALB2, RAD51C, SLX4, FANCD2, FANCDI, FAN1, others from HR	Polθ, Pol*ν*, REV1, other low fidelity translesion polymerases	LIG1
BER	Removal of smaller damaged DNA bases	APE1, PNKP, DNA glycosylases (*e.g.* OGG1 and NEIL1)		FEN1	Polβ	LIG1, LIG3
MMR	Removal of mismatched bases or small loop repair	MutSα (MSH2-MSH6), MutSβ (MSH2-MSH3)	MutLα (MLH1-PMS2)	EXO1	Polδ, Polε?	LIG1
NER	Removal of larger/bulky damaged DNA bases in DNA replication	XPC, RAD23B	TFIIH complex (XPB-XPD), RPA	ERCC1, XPG, XPA	Polδ, Polε, Polκ?	LIG1, LIG3?
SSB repair	Single-strand breaks	PARP1		SRCC1, PNKP, FEN1, TDP1, APTX	Polβ	LIG1, LIG3

Table 1.4: An overview of the DNA canonical DNA repair pathways. This table shows some of the major proteins involved in various repair processes. Note that there is significant blending of DNA repair pathways depending on damage and cell context; thus, the components of repair may vary. In the context of HD research, genes in blue are implicated by functional models as modifying somatic instability and/or disease phenotype; genes in red are implicated by both functional models and human genetics as HD modifiers. Note BIR (break-induced replication, a subset of homologous recombination for single-stranded double-strand break repair) is not shown on the basis it is poorly described in mammalian cells; see (Kramara et al., 2018) for a review. Pathway abbreviations: a-EJ/MMEJ: Alternative-end joining/Microhomology-mediated end-joining; HR (SDSA and dHJ): Homologous recombination, synthesis-dependent strand-annealing or double Holliday junction (these are two outcomes/sub-pathways of canonical homologous recombination; SDSA is preferred in non-meiotic contexts, dHJ resolution occurs during meiotic crossing over to generate genetic variability); NHEJ: Non-homologous end-joining; SSA: Single-strand annealing; ICL repair: Interstrand crosslink repair; BER: Base excision repair; MMR: Mismatch repair; NER: Nucleotide excision repair (note the global-NER pathway components are shown; a sister pathway, the transcription-coupled repair (TCR) pathway uses the CSA and CSB proteins as initial damage sensors during RNA transcription repair); SSB repair: Single strand break repair. Lesion abbreviations: CSR: Class switch recombination; DSB: Double-strand break; V(D)JR: V(D)J recombination (generation of immunoglobulins). Complex abbreviations: MRN complex: MRE11, RAD50, NBS1; BTR complex: BLM-TOPOIIIα-RMI1-RMI2; SMX complex: MUS81, EME1, SLX1, SLX4, XPF-ERCC1; FA core complex: FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL, FAAP100. Adapted from (Brown et al., 2017).

1.8 Thesis aims

Both human genetics and animal models have underscored DNA repair as important in the modification of HD pathology. However, many questions remain as to the specific mechanisms by which DNA repair contributes to HD, especially regarding *FAN1* which appears to protect against repeat expansion. Additionally, how non-DNA repair factors may contribute (or protect against) HD pathology is mostly unknown, although several candidate loci have been identified through GWAS. The primary goal of this project is to identify modifiers of HD age at onset, principally using next-generation sequencing (NGS) modalities. NGS offers several advantages over more traditional SNP-array and imputation approaches, including (1) the deconvolution of genes at implicated loci from GWAS, (2) the possible identification of rare modifiers not well captured by common variation alone and (3) the investigation of rare (and often coding) variants that may provide insight into specific regions or domains of a gene/protein implicated with disease phenotype. NGS techniques, therefore, have the potential to both better characterise and validate existing candidate disease modifiers through rare variants, whilst potentially identifying novel modifiers of HD poorly tagged by common variation. This thesis will investigate modifiers of HD onset through the following aims:

(1) Explore phenotypes in HD using a combination of the clinician's estimate of onset (sxrater) in addition to measures derived from the clinical characteristics questionnaire (CCQ).

(2) Stratify the large Registry-HD cohort by an age at motor onset residual to find the earliest and latest onset HD patients given their *HTT* CAG length.

(3) Use whole-exome sequencing to investigate rare variation occurring in previously highlighted genes of interest from GWAS and other functional study of HD.

(4) Implementation of an unbiased, rare variant whole-exome approach to find novel modifiers of disease onset.

(5) Study how *HTT* CAG length and sequence vary between early and late onset HD patients, and characterise the effect *HTT* sequence may have on repeat instability.

Chapter 2: Materials and methods

2.1 Study design

2.1.1 HD participants

Participants were taken from the observational European Huntington's disease Network's (EHDN) Registry-HD study (https://clinicaltrials.gov/ct2/show/NCT01590589). Ethical approval for Registry was obtained in each participating country, and participants gave written informed consent. All experiments described herein were conducted in accordance with the declaration of Helsinki.

2.1.2 Participant data

R2 and R3 Registry data cuts were used, with R3 data preferentially used when available as R3 includes the clinical characteristics questionnaire (CCQ). Where data was only given as a year, which was common for estimates of symptom onset given by the rater (sxrater), the 15th of July (15/07/xxxx) was used for estimation purposes. Table 2.1 contains the details of select demographic terms used for calculations. Table 2.2 contains the questions asked in the CCQ for all eight symptoms. Finally, Tables 2.3 and 2.4 contain the questions included in the hospital anxiety and depression scale with Snaith irritability scale (HADS-SIS) questionnaire, originally devised by (Zigmond and Snaith, 1983) and (Snaith et al., 1978). 8265 participants had CAG length data available, determined by local diagnostic labs. These were used in the clinical phenotype analyses in chapter 3. See also the standard Registry-HD protocols available online: https://www.enroll-hd.org/enrollhd_documents/2016-10-R1/registry-protocol-3.0.pdf.

Data term	Notes
Brthdtc	Date of birth (semi-anonymised)
Allele 1l	Smaller CAG size (determined by local labs)
Allele 2l	Larger CAG size (determined by local labs)
Sxrater	Rater's estimate of symptom onset (date)
Sxraterm	Rater's judgement of initial major symptom: 1 (motor), 2 (cognitive), 3 (psychiatric), 4 (oculomotor), 5 (other) or 6 (mixed)
Sxfam	[What date were] symptoms first noted by family
Sxfamm	Initial major symptom noted by family: 1 (motor), 2 (cognitive), 3 (psychiatric), 4 (oculomotor), 5 (other) or 6 (mixed)
Sxsubj	[What date were] symptoms first noted by participant
Sxsubjm	Initial major symptom noted by participant: 1 (motor), 2 (cognitive), 3 (psychiatric), 4 (oculomotor), 5 (other) or 6 (mixed)
Hddiagn	Date of clinical HD diagnosis
Momhd/dadhd	Mother/Father affected [by HD]
Momagesx/dadagesx	Age at onset of symptoms in mother/father [of HD]
Eduyrs	Years of education
Alcab	Does the participant currently drink alcohol?
Alcunits	[Number of alcohol] units per week
Tobab	Does the participant currently smoke?
Tobcpd	Cigarettes per day
Tfcscore (TFC)	[Total] functional score
Motscore (TMS)	[Total] motor score
Anxscore (TAS)	[HADS-SIS] anxiety subscore
Depscore (TDS)	[HADS-SIS] depression subscore
Irrscore (TIS)	[HADS-SIS] irritability subscore

Table 2.1: Select demographic and phenotypic data terms. Data terms were taken from the Registry-HD R3 data dictionary.

Data term	Notes
Ccmtr	Have motor symptoms ever been a part of the participant’s medical history?
Ccmtrage (MTR)	At what age did the participant’s motor symptoms begin?
Cccog	Has significant cognitive impairment (severe enough to impact on work or activities of daily living) or dementia ever been a part of the participant’s medical history
Cccogage (COG)	At what age did cognitive impairment first start to have an impact on daily life?
Ccapt	Has apathy ever been a part of the participant's medical history?
Ccaptage (APT)	At what age did apathy begin?
Ccdep	Has depression (includes treatment with antidepressants with or without a formally-stated diagnosis of depression) ever been a part of the participant’s medical history?
Ccdepage (DEP)	At what age did depression begin?
Ccpob	Has perseverative/obsessive behaviours ever been a part of the participant's medical history?
Ccpobage (POB)	At what age did perseverative/obsessive behaviour begin?
Ccirb	Has irritability ever been a part of the participant’s medical history?
Ccirbage (IRB)	At what age did the irritability begin?
Ccvab	Has violent or aggressive behaviour ever been a part of the participant’s medical history?
Ccvabage (VAB)	At what age did the violent or aggressive behaviour begin?
Ccpsy	Has psychosis (hallucinations or delusions) ever been a part of the participant’s medical history?
Ccpsyage (PSY)	At what age did psychosis (hallucinations or delusions) begin?

Table 2.2: The clinical characteristics questionnaire (CCQ). Taken from the Registry-HD R3 data dictionary. The CCQ consists of an initial binary yes/no question for if the participant has ever experienced a symptom, and, if yes, what the approximate age at symptom onset was.

Question (TDS)	Reply	Question (TAS)	Reply
I still enjoy the things I used to enjoy	- Definitely as much (0) - Not quite so much (1) - Only a little (2) - Hardly at all (3)	I feel tense or 'wound up'	- Most of the time (3) - A lot of the time (2) - From time to time, occasionally (1) - Not at all (0)
I can laugh and see the funny side of things	- As much as I always could (0) - Not quite so much now (1) - Definitely not so much now (2) - Not at all (3)	I get a sort of frightened feeling as if something awful is about to happen	- Very definitely and quite badly (3) - Yes, but not too badly (2) - A little, but it doesn't worry me (1) - Not at all (0)
I feel cheerful	- Never (3) - Not often (2) - Sometimes (1) - Most of the time (0)	I can sit at ease and feel relaxed	- Definitely (0) - Usually (1) - Not often (2) - Not at all (3)
I feel as if I am slowed down	- Nearly all the time (3) - Very often (2) - Sometimes (1) - Not at all (0)	Worrying thoughts go through my mind	- A great deal of the time (3) - A lot of the time (2) - Not too often (1) - Very little (0)
I have lost interest in my appearance	- Definitely (3) - I don't take as much care as I should (2) - I may not take quite as much care (1) - I take just as much care as ever (0)	I feel restless as if I have to be on the move	- Very much indeed (3) - Quite a lot (2) - Not very much (1) - Not at all (0)
I look forward with enjoyment to things	- As much as I ever did (0) - Rather less than I used to (1) - Definitely less than I used to (2) - Hardly at all (3)	I get a sort of frightened feeling like 'butterflies' in the stomach	- Not at all (0) - Occasionally (1) - Quite often (2) - Very often (3)
I can enjoy a good book or radio or television programme	- Often (0) - Sometimes (1) - Not often (2) - Very seldom (3)	I get sudden feelings of panic	- Very often indeed (3) - Quite often (2) - Not very often (1) - Not at all (0)

Table 2.3: Hospital anxiety and depression scale (HADS) questionnaire. Taken from the Registry-HD R3 data dictionary, originally devised by (Zigmond and Snaith, 1983). Questions on the left correspond to the total depression score (TDS) and questions on the right with total anxiety score (TAS).

Question	Reply
I lose my temper and shout or snap at others	- Yes, definitely (3) - Yes, sometimes (2) - No, not much (1) - No, not at all (0)
I am patient with other people	- All of the time (0) - Most of the time (1) - Some of the time (2) - Hardly ever (3)
I get angry with myself and call myself names	- Yes, definitely (3) - Sometimes (2) - Not often (1) - No, not at all (0)
I feel like harming myself	- Yes, definitely (3) - Yes, sometimes (2) - No, not much (1) - No, not at all (0)
The thought of hurting myself occurs to me	- Sometimes (3) - Not very often (2) - Hardly ever (1) - Not at all (0)
I feel I might lose control and hit or hurt someone	- Sometimes (3) - Occasionally (2) - Rarely (1) - Never (0)
People upset me so that I feel like slamming doors or banging about	- Yes, often (3) - Yes, sometimes (2) - Only occasionally (1) - Not at all (0)
Lately I have been getting annoyed with myself	- Very much so (3) - Rather a lot (2) - Not much (1) - Not at all (0)

Table 2.4: Snaith's irritability scale (SIS) questionnaire. Taken from the Registry-HD R3 data dictionary, devised originally by (Snaith et al., 1978).

2.2 Age at HD onset determination

2.2.1 Rater’s estimate of onset

General ages at onset were calculated using the assessing clinician’s (the rater’s) best estimate of disease onset (sxrater), coded as a date. The sxrater was used regardless of onset type (sxraterm), including if the onset type was unknown. The participant’s birthday (brthdtc) was then used to derive an estimated age ((sxrater-brthdtc)/365.25).

2.2.2 Deriving ages at onset (best-estimate)

Ages at onset were calculated using both the clinician’s estimate for HD onset (sxrater) and CCQ data. This age at onset is called a ‘best-estimate’ as it uses the two primary patient onset data (sxrater and CCQ) together (note, however, this is not a composite score, as detailed here). A best-estimate age at motor onset (AMO) was determined using both the clinician’s estimate for HD onset (sxrater) and an individual’s CCQ data. The sxrater was used where onset was classed as motor, oculomotor or mixed. Where onset type (as assessed by the clinician) was non-motor (psychiatric, cognitive or other), or where sufficient data were unavailable, the CCQ for motor onset (ccmtrage) was instead used to obtain an AMO. Sxrater was preferentially used where onset was motor, and non-motor onset individuals lacking a motor CCQ were excluded. Best-estimate ages at cognitive onset (ACO) were calculated similarly. Sxrater was used for ACO estimation for cognitive onset types; for all other onsets, cognitive CCQ (cccogage) was used for estimation when available. If the cognitive CCQ was not available for non-cognitive onset types, these data were excluded.

Quality control (QC) was used to improve onset estimation. For AMO best-estimates, onsets classed as motor, oculomotor or mixed were expected to have <2 years difference between motor CCQ and sxrater. For non-motor onsets, a motor CCQ was expected to be no more than 2 years earlier than sxrater. Likewise, for ACO best-estimates, individuals with cognitive onsets were expected to have <2 years difference between sxrater and cognitive CCQ. For other onset types, cognitive CCQ was expected to be no more than 2 years earlier than sxrater. AMO or ACOs with discrepancies greater than these allowances were manually curated. To this end, the sxrater, sxsubj (participant’s estimate of HD onset), sxfam (family’s estimate of HD onset), hddiagn (date of clinical HD diagnosis) and the entire CCQ data were considered along with clinical notes to assess data consistency and symptom history. Two assessors, including one clinician (Dr Thomas Massey), had to agree for data inclusion. Onset estimates that could not be accurately determined were excluded.

Age at first psychiatric symptom onset (APO) was calculated using either: (1) sxrater for psychiatric onsets or (2) the earliest recorded psychiatric CCQ (depression, irritability, VAB, apathy, POB or psychosis), whichever occurred earliest. An adjusted version of APO removed CCQ data occurring earlier than sxrater at 2, 5 and 10 year cut-offs.

2.2.3 Individual symptom determination using CCQ data

Unadjusted individual symptom onsets simply used the recorded CCQ data (see Table 2.2). For adjusted CCQ data, the sxrater (regardless of onset type) was used to remove CCQ data occurring earlier than sxrater at various cut-offs (2, 5 or 10 years earlier than the sxrater). A 2 year cut-off difference was most frequently used for downstream calculations. For binary symptom data, individuals having a symptom, regardless of age, were recorded as 1 (symptom experienced at some point) or were otherwise recorded as 0 (no symptom).

2.2.4 Calculating binary symptom onset using CCQ data

Binary CCQ data used the yes/no part of the CCQ data where 0 = never experienced the symptom and 1 = experienced the symptom at some point in life. A positive response for the yes/no segment of the CCQ was treated as a 1 regardless of if the age at onset for the symptom was known or not. The adjusted version of this metric removed individual CCQ data for which an age at onset for both sxrater and the CCQ data in question was available (for cases where either was missing, these data were excluded from the adjusted binary CCQ measures). For those with an sxrater/CCQ age at onset, CCQ data occurring >2 years earlier than the sxrater were removed. Thus, roughly, 0 = never experienced the symptom, 1=experienced the symptom in disease course of HD.

2.3 Statistical analyses of clinical data

2.3.1 Linear modelling

2.3.1.1 Multivariate symptom analysis using binary CCQ

Multivariate logit generalised linear models were constructed regressing on binary CCQ data, both adjusted and unadjusted binary CCQ (2.2.4). Two different series of models were constructed. The first, simpler model used sex, CAG length, disease duration and onset (as defined by the sxrater, regardless of onset type) as covariates for individuals with CAGs 36-99 and a known sxrater. Here, disease duration was derived from the individual's latest visit date on record (*i.e.* latest visit – sxrater disease onset).

The extended second model included many of the same covariates as the (Dale et al., 2016) study excluding medication: sex, CAG, disease duration, age at onset (sxrater), alcohol consumption in units/week, tobacco usage as cigarettes/day, education years, total functional capacity (TFC) score and total motor score (TMS). Only individuals with CAGs 39-55 and a TFC score >0 (excluding very advanced HD patients) were considered. All juvenile HD cases (<20 years at onset) were excluded. The most recent visit data was used where multiple visit data were available. Disease duration for the extended model was calculated using the visit date for the TFC score. For both the simpler and extended models, both unadjusted and adjusted CCQ data were used for comparative purposes. For psychosis symptoms only, individuals having a comorbid diagnosis of schizophrenia, schizotypal disorder or schizoaffective disorder were excluded from analysis. These were defined as F20, F21 or F25, respectively, in the International Statistical Classification of Disease-10 (ICD) diagnostic criteria.

2.3.1.2 HADS-SIS multivariate analysis

Total scores from the Hospital anxiety and Depression Scale with the Snaith Irritability Scale (HADS-SIS) (Table 2.4) were used to regress on using a generalised linear model: total anxiety score (TAS), total depression score (TDS) and total irritability score (TIS) (Table 2.3). The same covariates were used as in the extended model in 2.3.1.1. A second modelling approach transformed TAS/TDS/TIS scores into binary measures. TAS and TDS were 0-7 considered 'normal' (0) and 8-21 a 'case' (1). For TIS, 0-7 was normal (0) and 8-24 a case (1). These were then regressed on the same covariates as before (2.3.1.1, extended model).

2.3.1.3 Estimation of CAG length on symptom onset

Age at onset for each recorded symptom were natural log transformed. Linear models were constructed using Ln(Age at onset) ~ [CAG length] for CAGs 36-90 for whom sex was known. One individual with a highly unusual onset was found to lie in or near Cook's distance and was removed from all CAG length analyses (Cook's distance is useful in regression modelling to find unusual outliers which affect the overall model, as was found here). Ages at onset <3 years of age were removed as these were found to bias R^2 estimation.

2.3.2 Intergenerational anticipation estimates

Momagesx and Dadagesx data were used as age at onset estimates for the participant's mother and father, respectively. The rater's estimate onset for the offspring (sxrater), regardless of onset type, was used to estimate anticipation for these calculations, *i.e.* Parental age at onset (Mom/Dadagesx) – Child/Proband age at onset (sxrater).

2.4 Selecting an extreme onset cohort

An expected age at onset (O_{exp}) was calculated with the Langbehn model (Langbehn et al., 2004, 2010), shown in equations 2.1 and 2.2, using the local clinical lab CAG lengths for each individual (see 2.1). AMO residual was calculated taking the expected onset from the patient's observed age at motor onset (as in equation 2.2). 500 (250 early, 250 late) of the individuals with the largest residual AMOs were chosen for sequencing, between CAG 40-55. Expected ages at onset were calculated using the Langbehn model for CAGs 41-55. For 40 CAGs, the median age of motor onset for the Registry-HD group for individuals with 40 CAGs was used as the expected age of motor onset (O_{exp}) (median CAG_{40}=59 years). Individuals with missing onset type from the clinician were excluded from being selected for sequencing. All individuals selected for sequencing were re-genotyped by MiSeq (2.8) and their residuals recalculated using the MiSeq CAG lengths and the Langbehn model as described (2.9).

$$p(Age) = \left(1 + e^{\left(\frac{\pi}{\sqrt{3}} \times \frac{[-21.54 - e^{(9.56 - (0.146\ \times [CAG]) + Age}}{\sqrt{35.55} + e^{(17.72 - (0.327\ \times [CAG])}}\right)}\right)^{-1}$$

Equation 2.1: Full parametric Langbehn survival model. *p*(Age) = probability of not having neurological symptoms until the given age, [CAG] = CAG length.

$$O_{exp} = 21.54 + e^{9.556 - (0.146\ \times\ [CAG])}$$

$$Residual = O_{exp} - O_{obs}$$

Equation 2.2: Simplified Langbehn model for mean onset at a given CAG length. O_{exp} = expected onset, O_{obs} = observed onset, [CAG] = CAG length.

2.5 DNA Preparation

2.5.1 DNA used for sequencing

Early passage patient-derived lymphoblastoid cell (LBC) DNA was acquired for the 500 individuals selected for sequencing (2.4) from BioRep, Italy, in line with Registry protocols (see Registry protocol URL in 2.1.2). Patient whole blood DNA was obtained for a subset of these individuals and several 'normal HD onset' positive controls (total N=49), also obtained from Registry. The normal onset individuals had between -1 to +1 AMO residual using locally derived CAGs and the Langbehn model (see 2.4). Induced pluripotent stem cell (iPSC) DNA was obtained from N1 and N5 lines originally derived in (HD iPSC Consortium, 2012). The iPSCs were grown and DNA extracted by Jasmine Donaldson. Finally, longitudinal lymphoblastoid cells and their DNA were grown and extracted by Dr Thomas Massey.

2.5.2 DNA Quantitation

2.5.2.1 DNA preparation with PicoGreen™ (for plates)

Quant-iT™ PicoGreen™ (ThermoFisher, P7589) was used for preparation of plates for next-generation sequencing using standard guidelines. Briefly, a standard curve was constructed using standard DNA provided (example in Fig. 2.1). For the assay, 2 µL of DNA was added to a black Greiner plate (Greiner Bio-One, 655077) and 98 µL of working PicoGreen™ solution (1:200 dilution of original PicoGreen™ with TE buffer) was added to each well. Fluorescence was read on a microplate reader (Tecan Infinite 200Pro).

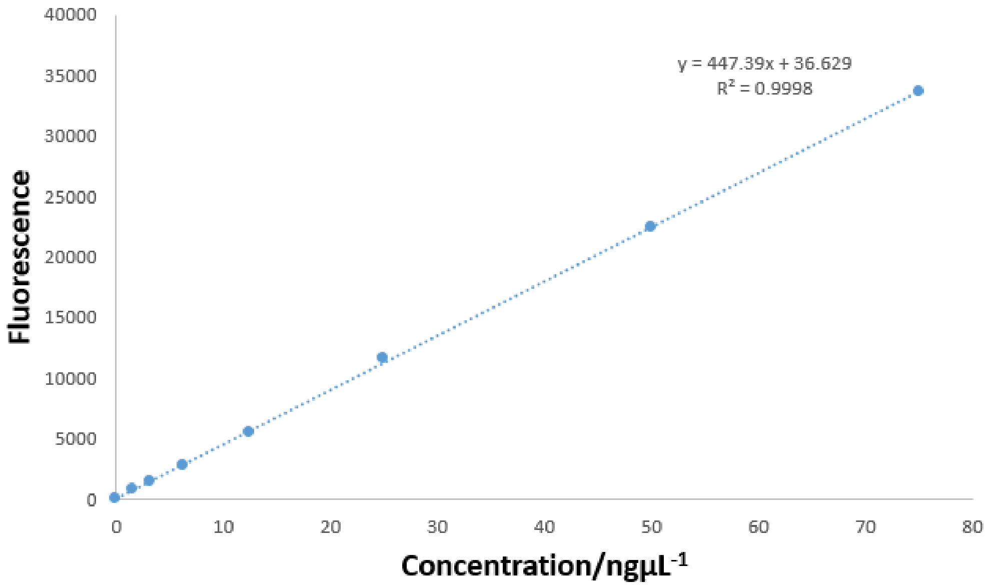


Figure 2.1: PicoGreen™ standard curve. Shown is a typical PicoGreen™ standard curve with the following DNA concentrations from smallest to highest: 0.00 (blank), 1.56, 3.13, 6.25, 12.50, 25.00, 50.00 and 75.00 ngµL^{-1}.

2.5.2.2 DNA preparation with Qubit™ (for individual samples/libraries)

Qubit™ dsDNA high-sensitivity (HS) assay kits (Q32584) were used for next-generation sequencing library preparation with a Qubit fluorometer. For Sanger sequencing and other applications involving DNA, the Qubit™ dsDNA broad range kit (Q32850) was used. Briefly, in both instances working solutions of Qubit™ were made (1:200 dilution) with the provided Qubit™ solution and buffer. 10 µL of each of the two standard DNA were added to 190 µL buffer to generate a standard curve using the Qubit™ fluorometer. DNA samples were diluted 1:100 (2 µL in 200 µL).

2.6 Whole-exome sequencing (WES)

2.6.1 Whole-exome library preparation

DNA samples from lymphoblastoid cells, iPSCs and blood DNA were taken as described (see 2.5.1). 12-plex exome libraries were created using TruSeq® rapid exome library kits (Illumina, now known as Nextera™ Exome Kit for 96 samples, 20020617). A full protocol is available online as TruSeq® Rapid Exome Reference Guide, see http://emea.support.illumina.com/downloads/truseq-rapid-exome-library-prep-reference-guide-1000000000751.html. The following is therefore an outline.

Plates containing 96 DNA samples for library preparation were prepared by PicoGreen™ to 7.5 ngµL^{-1}. In order to reduce technical variability between sequencing runs, 96-well plates were balanced equally between early and late onset patient samples where possible. In most cases, each 12-plex library consisted of 6 early and 6 late samples. Prepared DNA was fragmented enzymatically *via* transposase (TDE2). Fragmented libraries were barcoded adding i5 and i7 adapters (see Fig. 2.2) using a Biomek liquid handling workstation (Beckman Coulter FX^{P}). Following initial amplification, tagmented libraries were then cleaned-up using solid phase reversible immobilisation (SPRI) beads (part of the exome kit) added by a Biomek liquid handling workstation. Individual libraries were then quantified using PicoGreen™ and a sample taken from each row and column was diluted 1:10 and checked using a Bioanalyser (Agilent) with a DNA 1000 chip (Agilent, 5067-1505) for QC purposes. An example Bioanalyser trace for one of the plates prepared is in Fig. 2.3.

		Illumina i7 adapters											
		N701	N702	N703	N704	N705	N706	N707	N708	N709	N710	N711	N712
		1	2	3	4	5	6	7	8	9	10	11	12
Illumina i5 adapters E157	A												
E502	B												
E503	C												
E504	D												
E505	E												
E506	F												
E507	G												
E508	H												

Figure 2.2: WES index overview for plates (96 plate format). Shown are the i5 (Exxx) and i7 (N7xx) adapters used to barcode a 96-well plate in whole-exome sequencing (WES). Rows A-H and columns 1-12 refer to the position on the plate the barcodes would be positioned during exome library preparation. For instance, position C3 would be used to create an exome library with two barcodes, E503 and N703.

300 ng from each individual 12-plex library was then pooled for a total of 8 pools (Σ=3.6 µg DNA per pool) using the Biomek workstation. Following pooling, two hybridisation and capture steps were performed. Hybridisation used coding exome oligos (CEX) to enrich for the exome, cleaned up again using SPRI beads. The capture step used streptavidin magnetic beads (SMB) and an enhanced enrichment wash to select for captured exome libraries. Following hybridisation and capture, the library was amplified, and a final clean-up performed using SPRI beads. Enriched DNA libraries were quantified by Qubit™. A high-sensitivity Bioanalyser chip (Agilent, 5067-4626) was then used to (1) check library integrity and (2) estimate the average bp size of each pooled library (example in Fig. 2.4), allowing for a nM concentration for each library to be determined. Note that these traces include a pronounced PCR-bubble (the second smaller broad peak at ~500-1000bp, Fig. 2.4) observed during several plate preps, however we had no trouble sequencing these libraries.

2.6.2 Sequencing of WES libraries

Completed exome libraries were sequenced by the Medical Research Council (MRC) core team at Cardiff University on an in-house HiSeq 4000 platform (Illumina). Libraries were pooled in equimolar amounts in groups of 96 and run over 8 lanes on the HiSeq4000 patterned flow cell, excepting plates 1 (12 libraries run on a single lane) and 3 (60 libraries run across five lanes). Clustering used Illumina ExAmp regents from a HiSeq® 3000/4000 PE cluster kit (Illumina, PE-410-1001) on the cBot system. Sequencing used a 2x75bp end run with a HiSeq® 3000/4000 SBS kit for 150 cycles (Illumina, FC-410-1002).

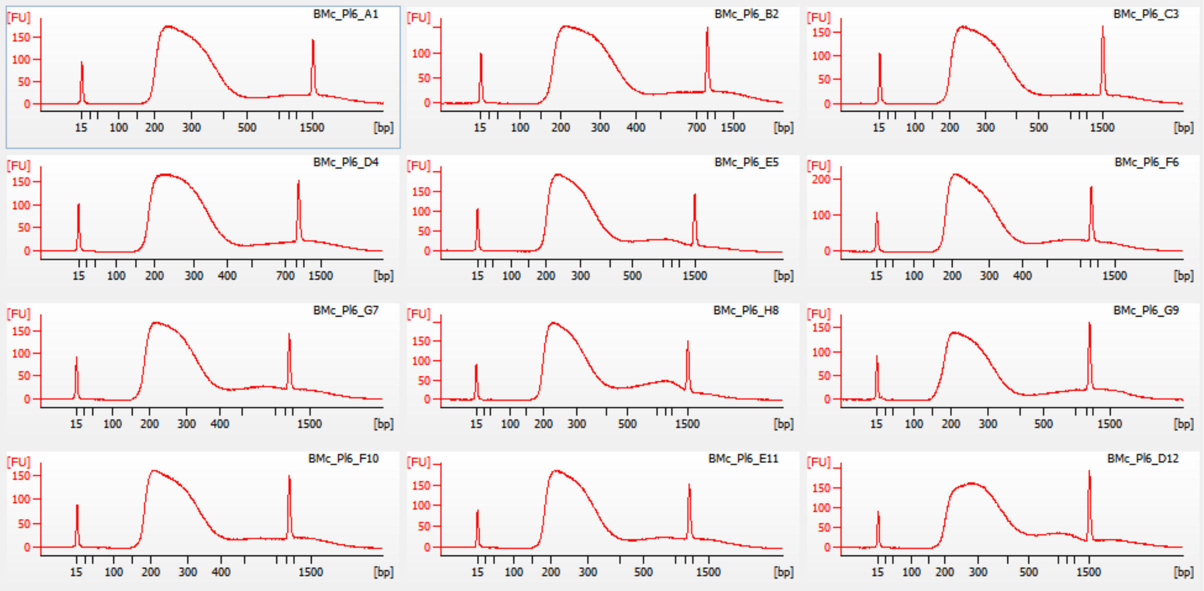


Figure 2.3: First QC step in WES using a DNA 1000 chip. Shown are Bioanalyser traces following the initial creation of the library before exome enrichment using a DNA 1000 chip (Agilent, 5067-1505). Each trace is a single library from the plate. The samples taken form a 'V' shape pattern on the plate, *i.e.* A1, B2, C3, D4 etc., in order to sample all rows and columns at least once.

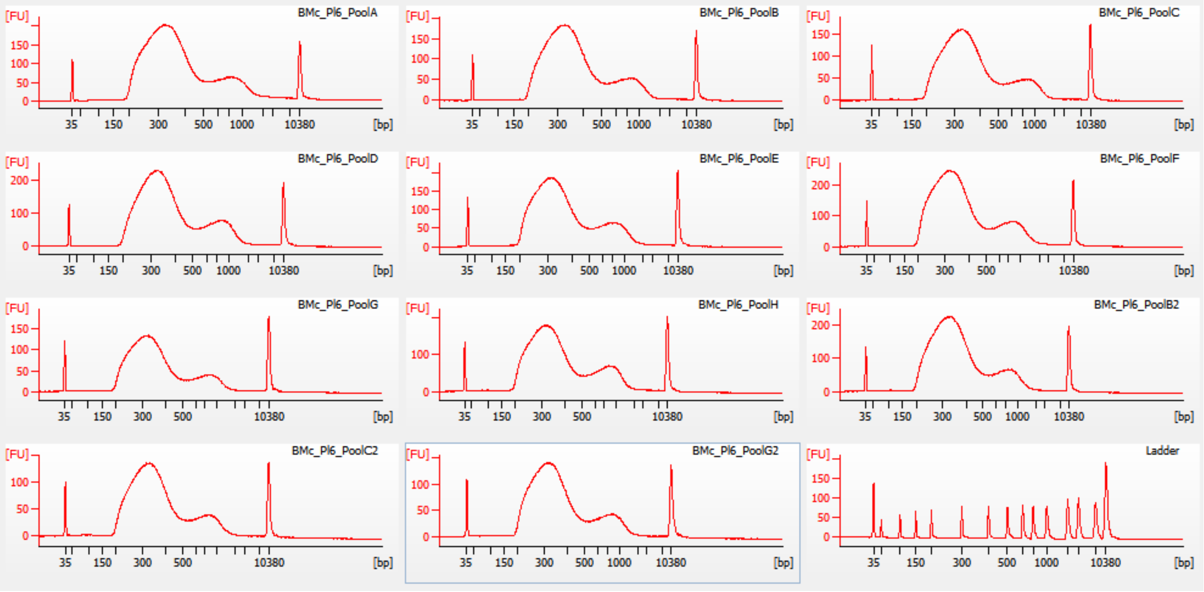


Figure 2.4: Second QC step in WES with a HS DNA chip. Shown above are the final exome-enriched sequencing libraries using a high-sensitivity Bioanalyser chip (Agilent, 5067-4626). These traces can be used to calculate the average bp size of each library. Note that there are eight pools per plate; the final four traces are the ladder (bottom right) and three repeats.

2.7 Analyses of whole-exome sequencing data

2.7.1 Pipelines for variant calling, QC and annotation

2.7.1.1 GATK pipeline

We used an in-house genome analysis toolkit (GATK) pipeline for alignment and variant discovery (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013), with pipeline scripts originally written by Dr Elliott Rees. De-multiplexed FASTQ reads were aligned to the GRCh37 reference genome utilising BWA version 0.7.5a (Li and Durbin, 2009). Duplicate reads were marked with Picard (picard-tools v1.97) (https://github.com/broadinstitute/picard). Indel realignment and base recalibration used GATK, and this generated variant-ready binary alignment map (BAM) files (Li et al., 2009). The GATK haplotype caller (v3.4) was used for calling single nucleotide variants (SNVs) and insertions/deletions (indels) to generate a variant calling file (VCF). Variants were annotated using GATK's variant quality score recalibration (VQSR). The VCF underwent initial annotation using SnpEff and SnpSift (Cingolani et al., 2012).

2.7.1.2 Exome QC pipeline

A schematic overview of the exome QC is shown in Fig. 2.5. First, Picard was used to generate QC metrics on BAM files using the Nextera™ targeted region manifest v1.2. We used 70% of the exome covered at 10X (PCT_TARGET_BASES_10X) as a cut-off, and exomes <70% were repeated. A VCF for all exomes was then imported into Hail (v0.1, devel-b08433b; https://github.com/hail-is/hail) running version 2.1.1 of Apache Spark™ (Zaharia et al., 2016). Anaconda (v2) and Java (v1.8.0_45) were also required. Per-sample QC metrics (sample_qc) were generated, and samples where either call rate, mean genotype quality or mean depth were >3 standard deviations smaller than the mean were excluded and repeated where possible. Contamination was assessed with VerifyBamID (v1.1.3; (Jun et al., 2012)) using the sequence-only estimate for contamination, 'Freemix'. Samples with Freemix > 0.075 were excluded, as per the ExAC study (Lek et al., 2016). The heterozygote/homozygote ratio (rHetHomVar) in Hail was roughly equivalent. Where possible, samples which failed contamination checks were repeated to maximise N. At this point, any sample duplicates were removed. Where both duplicates passed QC, the sample with the highest QC metrics was used.

Imputed sex was determined using Peddy (v0.3.5; (Pedersen and Quinlan, 2017), https://github.com/brentp/peddy) and these compared to sex in the Registry-HD database. Any samples which differed between imputed and recorded sex were removed from the analysis. One individual with an unknown sex in the Registry database was kept. Ancestry was estimated using Peddy by principal component analysis (PCA) against 2504 whole-genomes from the 1000 genomes project (1000 Genomes Project Consortium, 2015). The first 10 principal components (PCs) were then calculated using hail's built in pca() function, which implements PCAs as per (Patterson et al., 2006). Relatedness was calculated using PLINK v1.9 ((Purcell et al., 2007; Chang et al., 2015); www.cog-genomics.org/plink/1.9/) identity by descent (IBD) cut-off of PI_HAT > 0.5, and for related pairs, the individual with the largest residual were retained for analysis. A genetic relatedness matrix (GRM) was also calculated in Hail (grm()), equivalent to PLINK's IBD. Finally, all samples were re-genotyped using MiSeq and two early/late onset populations redefined for downstream analysis (2.8-2.9).

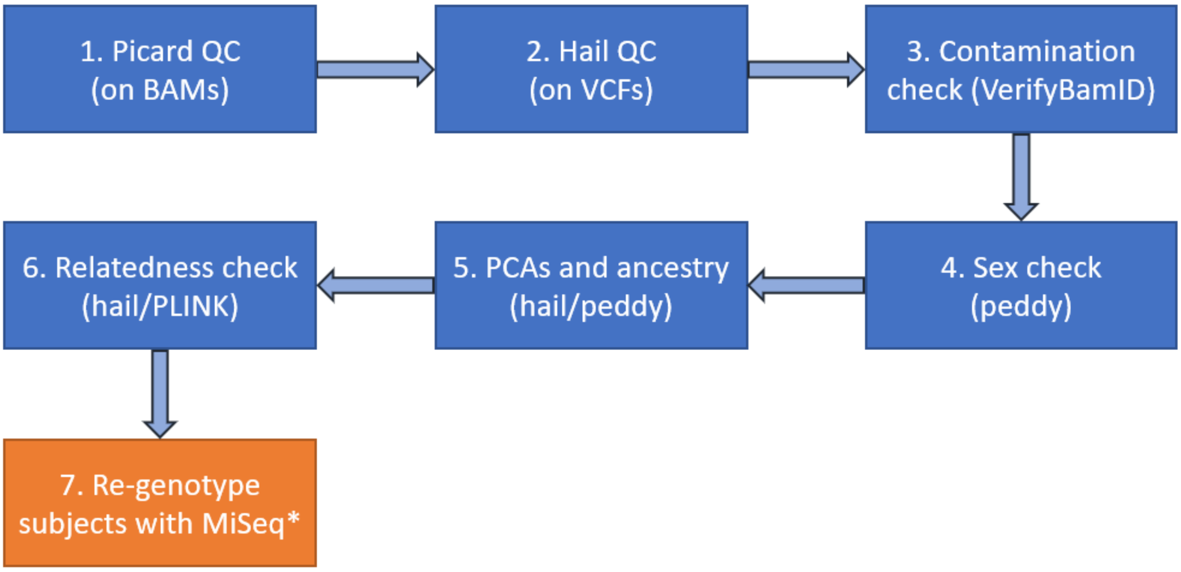

1. Picard QC (on BAMs)
2. Hail QC (on VCFs)
3. Contamination check (VerifyBamID)
4. Sex check (peddy)
5. PCAs and ancestry (hail/peddy)
6. Relatedness check (hail/PLINK)
7. Re-genotype subjects with MiSeq*

Figure 2.5: An overview of the exome quality control pipeline. Shown is an outline of the quality control (QC) pipeline employed. Note that step 7 (MiSeq genotyping of the CAG repeat) is detailed in 2.8/2.9 & chapter 5. See also Fig. 4.9.

2.7.1.3 Variant annotation pipeline

An overview of this pipeline is shown in Fig. 2.6. A VCF containing all samples was imported into hail (version devel-b08433b), and exomes failing the QC in 2.7.1.2 or were removed. Duplicates were likewise removed. Multiallelic sites were split using Hail's split_multi() function. Heterozygote and homozygote calls were then defined. Where ≤10% of read calls were an alternative allele, the variant was a homozygote reference. Equally, where ≥90% reads were the alternative allele, the variant was classed as a homozygote alternative. Heterozygotes were defined where between ≤25% and ≥75% of reads were the reference. Variant calls/genotypes which did not conform to the homozygote/heterozygote definition were filtered from the analysis. Per-variant QC filtered sample genotype calls where read depth was <10 and genotype quality was <30.

The variant effect predictor tool (VEP, version 89.6) (McLaren et al., 2016) was used to annotate all variants run alongside v86 of ensembl tools. Three variant classes were manually assigned using VEP's most_severe_consequence flag: loss-of function (LoF), non-synonymous (NS) and synonymous variants. LoF variants were defined as splice acceptor, splice donor, stop gained or frameshift variants. Non-synonymous (NS) variants included all LoF variants, inframe indels, stop or start losses and missense variants. Synonymous (SY) variants were those with no coding or functional change. The gene ID and symbol where the most severe consequence occurs (LoF > NS > SY) was associated with the variant.

Variants were annotated with minor allele frequencies (MAFs) from gnomAD's non-Finnish European (NFE) group (v2.0.2) (Lek et al., 2016; Karczewski et al., 2019). Variants in our dataset missing from gnomAD were included at every MAF cut-off. The following MAF cut-offs were used: MAF ≤0.1% (very rare); MAF ≤1% (rare) and MAF ≤2% (uncommon). Damaging scores from dbNSFP v3.0 (Liu et al., 2011, 2016) were also annotated, including CADD (Kircher et al., 2014; Rentzsch et al., 2019), SIFT (Ng and Henikoff, 2003; Sim et al., 2012) and Polyphen2 (Adzhubei et al., 2010). Non-synonymous damaging (NSD) variants were defined as rare (≤1% gnomAD MAF) and damaging (either LoF or NS variants with CADD PHRED ≥20). Individuals were marked as early, late or neither using the redefined early/late/normal populations from MiSeq (2.8/2.9) Similarly, AMO residuals from MiSeq (corrected and uncorrected) and TWAS data (GeM-HD Consortium, 2019) were also annotated. Baseline variant rates (BVRs) were calculated at different MAFs and filters by summing all the variants that pass QC for each sample for each mutation class. These were included in downstream whole-exome analyses (2.7.3-2.7.4).

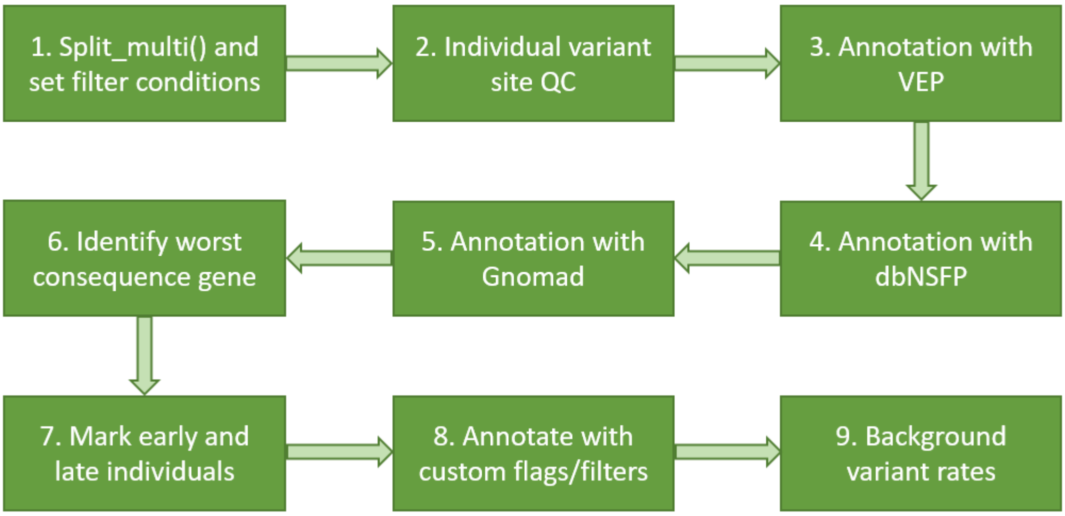

1. Split_multi() and set filter conditions
2. Individual variant site QC
3. Annotation with VEP
4. Annotation with dbNSFP
5. Annotation with Gnomad
6. Identify worst consequence gene
7. Mark early and late individuals
8. Annotate with custom flags/filters
9. Background variant rates

Figure 2.6: Overview of the annotation pipeline. Shown is an outline of the annotation pipeline employed in Hail for the exomes that passed QC.

2.7.2 Exome candidate gene analysis

2.7.2.1 Identifying NSD variation in candidate genes

All NS and LoF variants in genes of interest were extracted from genes of interest using the dichotomous (early/late) population from 2.9 (N=440). Hail aggregated numbers of homozygote references (HomR), heterozygote (Het), homozygote variant (HomV) and N/C (not called) variant calls for early and late groups. Non-synonymous damaging (NSD) variants were defined as before in 2.7.1; damaging variants (LoF or NS with CADD PHRED ≥20) either missing from gnomAD or NFE MAF≤1%.

2.7.2.2 Counting CAGs in *HTT* from WES

Raw reads from variant ready BAM files (2.7.1) were extracted using SAMtools (v1.9; (Li et al., 2009; Li, 2011)) targeting chr4 2,900,000-3,100,000 (GRCh37). Reads were manually appraised examining (1) the 5'- of the CAG region (following the TTC in the 5'-seq (encoding N17 of HTT), (2) the 3'- of the CAG region (centred around the CAA interruption) and (3) the number of CCGs observed in the polyP region. Where read-through of the wild-type repeat was possible, inference of the 3'- structure of the expanded repeat was possible.

2.7.2.3 Estimating effects of STR length in *MSH3* and *TCERG1*

For *MSH3*, exomes which had at least 2 of the 4 short tandem repeat (STR) calls identified by GATK were used in a linear regression model. In *TCERG1*, exomes were only included in a linear regression if all 5 STR variant sites were called. Here, a positive variant call simply means the site was not missing due to low coverage. Genotype length was equivalent to the number of repeating units gained (positive genotype length) or lost (negative genotype length) compared to the reference sequence (a homozygote reference has a genotype length of 0). For *MSH3*, one repeating unit was one amino acid; for *TCERG1* one repeating unit was two amino acids. In both instances, homozygote variants would be counted as having twice the effect on genotype length. For instance, a heterozygote for the Pro67Pro69del STR variant in *MSH3* would be counted as -3 (3 amino acid change, and 3 repeating unit change), whereas a homozygote for the variant would count as -6.

2.7.2.4 Determining coverage in target genes

Exon coordinates for all genes in the GRCh37 build of the genome were downloaded from GENCODE (https://www.gencodegenes.org/human/release_19.html; (Frankish et al., 2019)), accessed June 2019). Depth at each nucleotide in the canonical transcripts of target genes was extracted using custom scripts for BEDTools v2.24.0 (Quinlan and Hall, 2010). Mean depth was then calculated on a per-exome and per-plate level for each target transcript.

2.7.3 Burden regression analyses

2.7.3.1 Logistic regression (variant-by-variant)

A logistic regression for all exome variants passing QC was conducted using the Wald test, implemented in Hail using logreg(). We regressed the binary early (1) or late (0) phenotype in our dichotomous populations (N=440) (as defined in 2.9) on variant count for each variant observed. The covariates used were PC1-5 (Hail) and mean variant sample depth for each sample. NS damaging variants could have either CADD PRED ≥20 or missing CADD in this test (this differs to the following whole-exome tests where variants had to have a CADD PHRED annotation of ≥20 to be considered). Note in the variant-by-variant test, LoF variants were not included with the NS damaging variants.

2.7.3.2 Whole-exome logistic regression

Whole-exome burden regression used the Wald test implemented in Hail (logreg_burden()). Individuals were coded from the dichotomous population (2.9; N=440) as early (1) or late (0) onset. Variants had to pass the following filters to be included: non-synonymous damaging (NSD) (LoF or NS ≥20 CADD PHRED), GATK's variant quality score recalibration (VQSR≥98.5), call rate ≥75% and either (1) an missing MAF from gnomAD (v2.0.2) or (2) a MAF≤ the defined filter, if the variant was observed in gnomAD. The MAF filters used were 0.1% (very rare), 1% (rare) and 2% (uncommon). The following covariates were used: PC1-5 (from Hail), baseline variant rate (BVR) and the mean variant depth (Hail). For tests which used deleterious variant weighting, CADD PHRED was multiplied by the number of non-reference genotypes. Where imputation of CADD scores was used, missing CADD PHRED scores were imputed as 15. For weighting on MAF, we multiplied non-reference genotypes by (1/MAF). For MAF imputation where the variant was not observed in gnomAD (v2.0.2), we imputed values of 1,000,000 where (equivalent to a singleton in gnomAD).

2.7.3.3 Whole-exome linear regression

Whole-exome burden linear used linreg_burden() in Hail with the same filters (NSD, VQSR≥98.5, call rate≥75%, MAF≤1%) and covariates (PC1-5, BVR and mean variant depth) as above in 2.7.3.2. We regressed both uncorrected and corrected AMO residuals (y-variables; calculated in 2.9) on NSD variant burden (x-variable).

2.7.4 Whole-exome SKAT and SKAT-O analyses

VDS files from Hail were exported to PLINK format (.bed, .bim and .fam). A custom SetID file was created containing variants classified as NSD as in 2.7.1 (MAF ≤1% NFE and either LoF or CADD PHRED ≥20) and were VQSR≥98.5. The SKAT package (v1.3.2.1) in R (Wu et al., 2011; Lee et al., 2012d) was used to generate an SSD_SetID file to run SKAT and SKAT-O tests. PC1-5, BVR and average variant depth were used as covariates, and a missingness cut-off of 0.25 was employed. For dichotomous analyses, the N=440 population was used, coding early (1) and late (0). For continuous analyses, the 485 population was used with corrected and uncorrected residuals (2.9). For SKAT(-O) tests with weighted deleteriousness, we imputed missing scores as CADD PHRED 20 for NSD variants, and 30 for LoF variants.

2.7.5 Pathway analysis

Gene sets were taken from the Gene Ontology (GO) database (Ashburner et al., 2000; The Gene Ontology Consortium, 2019), accessed Apr 2017. *p* values for genes from whole-exome burden regression tests and SKAT(-O) tests were taken. Genes *p* values were then combined across GO pathways using Fisher's method. Fisher's method uses the -2LN(*p*) for each *p* value, returning a chi-squared distribution. These are then summed across pathways/gene sets, and a *p* value is calculated from the chi-squared distribution with the (genes with a *p* value)*2 as the degrees of freedom. Genes with missing *p* values were excluded.

2.8 MiSeq sequencing

2.8.1 MiSeq library generation

2.8.1.1 PCR amplification

This protocol is now available in Protocol Exchange (Ciosi et al., 2018), and thus the following is an outline. Libraries were prepared at Glasgow University, and all initial PCR amplification was performed in a laminar flow hood. 2 µL of DNA from patient blood or lymphoblastoid cells were equalised to 10 ngµL^{-1} using PicoGreen™ (methods 2.5.2) in 96-well plates with at least one positive and negative control on each plate. The negative control was 2 µL nuclease-free water, and the positive control was 2 µL of a known HD patient from the Venezuelan kindred. The libraries were prepared for 384-plate sequencing; to this end, 2 µL (5 µM) of each of the forward (S) and reverse (N) primers was added according to the plate map in Fig. 2.7. A full list of these primers is available in (Ciosi et al., 2018), supplementary table 2. A custom 10X master mix (Sigma-Aldrich, D4545; 45 mM Tris-HCl (pH 8.8), 11 mM $(NH_4)_2SO_4$, 4.5 mM $MgCl_2$, 0.113 mgmL^{-1} BSA, 4.4 µM EDTA and 1mM each of dATP, dCTP, dGTP and dTTP) was used after adding 0.048% v/v β-mercaptoethanol (β-ME; Sigma-Aldrich, M3148). The final master mix contained 1.5 µL of the described 10X custom master mix with 0.048% β-ME, 2.8 µL nuclease-free water, 1.5 µL DMSO (Sigma-Aldrich, D8418) and 0.2 µL Taq polymerase (Sigma-Aldrich, D4545); 6 µL of the master mix was added to each well for PCR. The PCR amplification programme used was 96^{0}C for 5 minutes, then 28 cycles of 96^{0}C for 45s, 58.5^{0}C for 45s and 72^{0}C for 180s, followed by 72^{0}C for 10 minutes.

2.8.1.2 Library clean-up

After amplification of all four 96-well plates for a 384-plate sequencing run, 5 µL of each amplified library was pooled across eight 1.5 mL LoBind DNA tubes (Eppendorf, 0030108051). Library clean-up used two AMPure XP SPRI bead (Beckman Coulter, A63881) steps; an initial 0.6X bead concentration to remove unused primers and 1.4X to size select for the expanded *HTT* allele. These steps are described in detail in (Ciosi et al., 2018). Pre- and post-clean-up libraries were run on a Bioanalyser (Agilent) with a high sensitivity DNA chip for QC purposes (see Fig. 2.8).

2.8.2 Sequencing on the MiSeq

Libraries were handed over to the staff at Glasgow University for sequencing on a MiSeq. Libraries were sequenced using a 600-cycle MiSeq v3 reagent kit (Illumina, MS-102-3003), running with 400bp forward and 200bp reverse sequencing. The sequencing parameters are described in (Ciosi et al., 2018).

Illumina i7 adapters

Illumina i5 adapters		N701	N702	N703	N704	N705	N706	N707	N710	N711	N712	N714	N715	N716	N718	N719	N720	N721	N722	N723	N724	N726	N727	N728	N729
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
S502	A																								
S503	B																								
S505	C																								
S506	D																								
S507	E																								
S508	F																								
S510	G																								
S511	H																								
S513	A																								
S515	B																								
S516	C																								
S517	D																								
S518	E																								
S520	F																								
S521	G																								
S522	H																								

Figure 2.7: MiSeq index overview for plates (384 format). Shown are the i5 (S5xx) and i7 (N7xx) primer layouts used for the four 96-well plates making up a 384-plate for MiSeq sequencing.

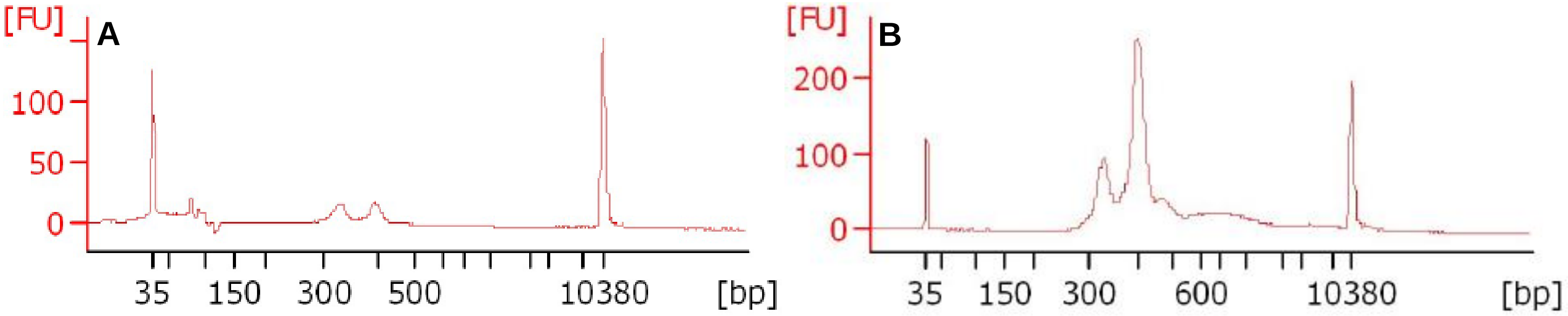


Figure 2.8: Bioanalyser traces of MiSeq libraries pre- and post-clean-up. (A) shows the initial library, pre-clean-up, following PCR amplification, diluted 1:30. Note the primer dimer around ~50-100 bp is so high the Bioanalyser is unable to read these properly (and this shows up as partially negative on the trace). (B) shows post SPRI bead clean-up where the primer dimer has been removed and the expanded *HTT* allele has been selected for, diluted 1:4. Both traces used DNA high sensitivity chips (Agilent, 5067-4626).

2.9 MiSeq data analyses

2.9.1 *HTT* genotyping with Scale-HD

Bioinformatic processing of MiSeq data used the Scale-HD pipeline (v0.322) written by Alastair Maxwell. Our data were demultiplexed and genotyped on Glasgow University's local Galaxy cluster, before being re-analysed locally on the Raven supercomputing cluster (ARCCA, Cardiff University). Scale-HD's installation and usage are described in detail in its documentation (https://scalehd.readthedocs.io/en/latest/) and thus the following is an outline. Scale-HD trims adapters from reads using cutadapt (Martin, 2011). Reads are aligned to a database of 4000 *HTT* structures with varying structures (*i.e.* refseq method) using BWA (Li and Durbin, 2009). Any atypical structures detected undergo further alignment to a separate database of non-canonical *HTT* structures (8000). This process is repeated for the second allele in the sample, and the alignments are used to create a BAM file for each sample. Scale-HD also uses freebayes (Garrison and Marth, 2012) for detection of SNPs occurring in *HTT* alleles. Aligned *HTT* BAM files were manually appraised, visualised with Tablet (Milne et al., 2013). In manual curation, care was taken to examine the entire length of the CAG repeat to identify possible sequence changes/interruptions occurring in the repeat tract.

2.9.2 Estimating instability in MiSeq with Scale-HD

Scale-HD was used to estimate instability of the CAG repeat in MiSeq data. The formulae for somatic mosaicism (forward instability) and slippage (reverse instability) are shown below in Equations 2.3 & 2.4, respectively. More information is available in the Scale-HD documentation (https://scalehd.readthedocs.io/en/latest/).

$$\frac{\left(\sum_{k=1}^{10} n_k\right)}{n_0}$$

Equation 2.3: MiSeq mosaicism (Scale-HD). In the equation, n is the number of reads per peak and k is the distance in peaks from the modal peak (n_0=modal peak).

$$\frac{\left(\sum_{k=-1}^{-2} n_k\right)}{n_0}$$

Equation 2.4: MiSeq slippage (Scale-HD). In the equation, n is the number of reads per peak and k is the distance in peaks from the modal peak (n_0=modal peak).

2.9.3 Redefining onset groups using MiSeq

All samples that were exome sequenced were also sequenced using targeted MiSeq. Two measures were derived for each patient (see Fig. 2.9 for a diagrammatic overview). The first was the ‘uncorrected’ residual, polyQ length – 2, to be directly comparable with lengths obtained from genescan (which assumes a canonical, single interruption at the 3’ end of the CAG repeat, *i.e.* CAG$_n$**CAA**CAG). The second, ‘corrected’ residual, was the length of the pure CAG repeat. The CAC interruption in one *HTT* allele was treated as a CAG for deriving an uncorrected residual. For re-calculation of expected ages at onset using the re-genotyped MiSeq CAGs, we used the Langbehn model ((Langbehn et al., 2004); see 2.4 & equation 2.2), and this was used to recalculate residual age at motor onset. For CAGs 38-40, the median age at motor onset for Registry at that CAG length was used for estimation (62, 62 and 59 years, respectively). Each sample was re-annotated as being early, late or neither based on the uncorrected MiSeq residual. AMO residuals ≤-5 were considered early onset and ≥5 late onset. Note: for two individuals which failed MiSeq (E265 and L301), we instead used an in-house genescan CAG length estimation (see 2.11), and assumed both had canonical alleles (so corrected/uncorrected residuals were the same for these two samples).

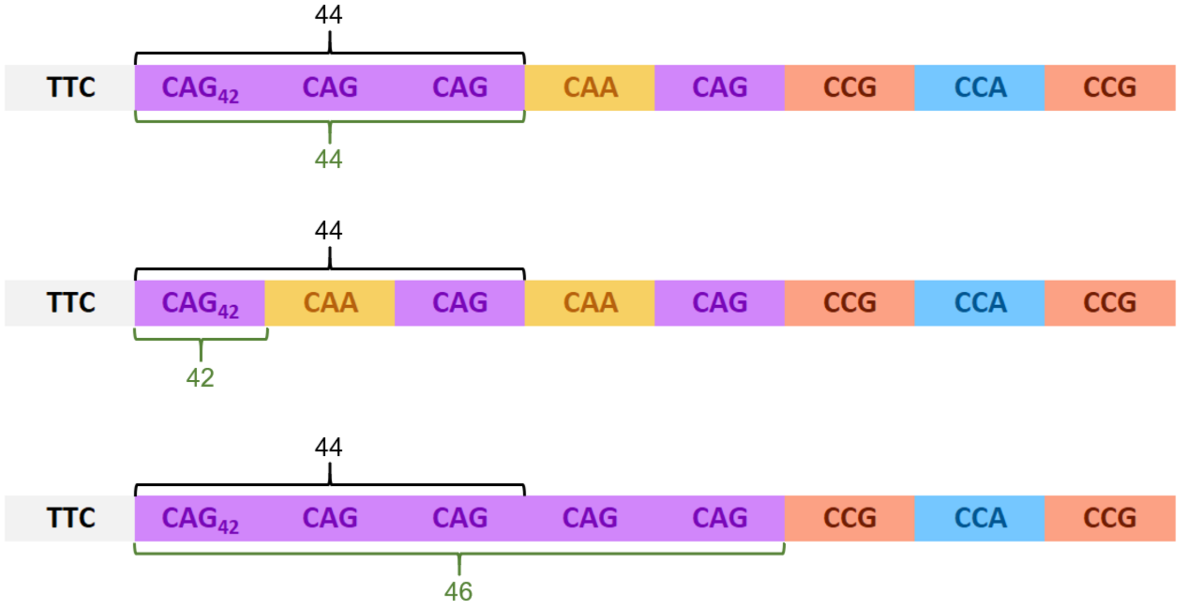


Figure 2.9: Corrected and uncorrected MiSeq genotyping. Shown are differences in sizing between uncorrected (in black) and corrected (in green) genotype lengths for the three most common *HTT* CAG allele structures.

2.9.4 Generalised linear models for MiSeq data

Generalised linear models (GLMs) were used to investigate the effect of interruption structures on onset, regressing residual age at motor onset (both corrected and uncorrected lengths; see 2.9.3) on structural features of *HTT*. Two models were made: in the first, the number of interruptions was as coded as a single value, representing the number of interruptions observed in the allele. A loss-of interruption allele was coded as 0, a normal allele as 1, a duplicated allele (**CAA**CAG)$_2$ or double tandem interruption (**CAA**$_2$CAG) as 2 and a tandem triple interruption (**CAA**$_3$CAG) or non-Q interruption (**CAC**CAG$_3$**CAA**) as 3. The second modelling approach coded interruptions as two separate values, either as having a normal interruption (0, 0), a loss of interruption (0, 1) or an additional interruption of any kind (1, 0). As well as interruptions occurring in the expanded *HTT* allele, the following covariates were also used: interruptions in the wild-type *HTT*, CCG tract interruptions in both expanded and wild-type *HTT* and total pure polyproline (polyP) length in expanded and wild-type *HTT*.

Generalised linear models were also used to investigate the relationship between MiSeq mosaicism with various covariates (including age at onset) and progression (progression measure from (Hensman Moss et al., 2017). MiSeq mosaicism (separately for lymphoblastoid and blood DNA) was regressed on corrected residual age at motor onset (from pure CAG length), interruption status (expanded allele), pure CAG length and age at which the sample was taken from the HD individual. For progression, the progression measure was regressed on MiSeq mosaicism, interruption status, sample age and pure CAG length. Interruption status in both models was the same as the model in the paragraph prior to this, *i.e.* pure CAG allele as 0, canonical as 1, duplication/double tandem as 2 and the non-Q/triple tandem variant as 3.

2.10 Sanger sequencing

2.10.1 Variant confirmation

Amplifications for sanger sequencing were performed using MyTaq™ (Bioline, BIO21127) as per the Bioline's guidelines for PCR amplification. Briefly, a master mix containing 10 µL of 5x MyTaq reaction buffer, 0.5 µL of each fragment (Frag) primer (10 µM) and 26 µL of nuclease free water was made up (multiplied by the number of samples). 2 µL of each DNA template was added (~10-30 ngµL^{-1}). The programme used was 95^{0}C for 2 minutes, followed by 30 cycles of 95^{0}C for 30s, 55^{0}C for 30s and 72^{0}C for 60s, finishing with 72^{0}C for 5 minutes. PCR reactions were cleaned up using QIAquick PCR purification kits (28104, Qiagen) using the provided protocol for obtaining higher concentration purified samples. 1.5 µL of the relevant (single) sequencing (Seq) primer (25 µM) was added to each purified sample. Sequencing was performed externally using the Eurofins Genomics LIGHTRUN service. See Table 2.5 and 2.6 for primer details.

2.10.2 Sanger sequencing of the *HTT* repeat

DNA from individuals with atypical *HTT* alleles was amplified with PCR using TaKaRa LA Taq® DNA polymerase with GC buffer (RR02AG, TaKaRa); the master mix consisted of 0.3

µL deionised water, 5 µL GC buffer (II), 1.6 µL dNTPs (provided), 0.5 µL of each primer (forward and reverse LKH primers from Swami et al., 2009 – see Table 2.6) and 0.1µL of polymerase. 2 µL of each DNA template (~10-30 ngµL^{-1}) was added. The programme used was an initial 94^{0}C for 90s, then 32 cycles of 94^{0}C for 30s, 65^{0}C for 30s and 72^{0}C for 90s. The product was run on a 1.5% agarose gel containing ethidium bromide alongside a 100bp Hyperladder™ (BIO-33029, Bioline). The expanded and wild-type bands were cut out using a clean scalpel and a UV transilluminator. DNA from gel bands was extracted using a QIAquick gel extraction kit (28115, Qiagen) using the provided protocol for obtaining high concentration DNA. 1.5 µL of 25 µM HTT sequencing primer was added to each 9 µL reaction. Sanger sequencing was performed externally with the Eurofins Genomics LIGHTRUN service.

2.10.3 Primer design

Primers for Sanger sequencing were designed using Primer3 (http://primer3.ut.ee/) (Koressaar and Remm, 2007; Untergasser et al., 2012) and cross-checked using the UCSC in-silico PCR tool (https://genome.ucsc.edu/cgi-bin/hgPcr). For initial amplification, the following picking conditions were used: primer size 18-23bp (optimal 20), primer T_m 57-62^{0}C (optimal 59^{0}C) and primer GC% 30-70% (optimal 50%). For sequencing primers: primer size 17-19bp (optimal 18), primer T_m 52-58^{0}C (optimal 56^{0}C) and primer GC% 36-60% (optimal 50%), with a GC clamp on the 3' end (but no more than 3 G/C in a row). No product T_m was specified. Sequencing primers had no more than four identical nucleotides in a row (*e.g.* AAAA).

Target(s): FAN1 M50R (Seq_a), T187fs (Seq_b)		
Primer name	**Length**	**Primer sequence**
FAN1_Frag1F	23	ACTCATGATGTCAGAAGGGAAAC
FAN1_Frag1R	23	ACCATATGTTCAGGAATGCACTC
FAN1_Seq1F_a	19	ACTCATGATGTCAGAAGGG
FAN1_Seq1R_a	17	TAAGCCAACCTGCCCTG
FAN1_Seq1F_b	18	TTGTTTGGGAAGCCTAGC
FAN1_Seq1R_b	19	TATGTTCAGGAATGCACTC

Target(s): FAN1 R377W, L395P		
Primer name	**Length**	**Primer sequence**
FAN1_Frag2F	23	AGGCAAAATCTCATAGTTCTGCA
FAN1_Frag2R	20	CATCATGCCCCAATCAGAGC
FAN1_Seq2F	19	CAATGATATCCCTCACAGC
FAN1_Seq2R	18	TGAAAACAAACACGTGCG

Target(s): FAN1 D498N, R507H, R507C		
Primer name	**Length**	**Primer sequence**
FAN1_Frag3F	21	ACTCCTTTCTGCTCCTGAACT
FAN1_Frag3R	23	CCAGCCTTCTCAATCTAACTACA
FAN1_Seq3F	18	TCCTTTCTGCTCCTGAAC
FAN1_Seq3R	19	CCAGCCTTCTCAATCTAAC

Target(s): FAN1 P654L, R658W		
Primer name	**Length**	**Primer sequence**
FAN1_Frag4F	20	TGGTAGCTGGCTGTGAGAAT
FAN1_Frag4R	22	TCACATGTTTAACGCCATCACA
FAN1_Seq4F	18	TAGCTGGCTGTGAGAATG
FAN1_Seq4R	19	TGTTTAACGCCATCACATC

Target(s): FAN1 D702E, Q717R		
Primer name	**Length**	**Primer sequence**
FAN1_Frag5F	20	CCTCAAAGTCCCTGTCCTGT
FAN1_Frag5R	23	GGTTGGAAGAACACAGACAAAAG
FAN1_Seq5F	17	GGTCCTTGGCCTCATTG
FAN1_Seq5R	17	TGAGTACCGGTTCCAGG

Target(s): FAN1 K794R		
Primer name	**Length**	**Primer sequence**
FAN1_Frag6F	21	ACTTTGTGGTAAGGGAGGTCA
FAN1_Frag6R	20	CTGGGTGCCACAAGAGAAAG
FAN1_Seq6F	18	CTTTTGCTGACCTGAGGC
FAN1_Seq6R	18	CCACAAGAGAAAGCCTGC

Target(s): FAN1 V963W964insL, R969L		
Primer name	**Length**	**Primer sequence**
FAN1_Frag7F	20	CCATTCTCTGTCACGAGGGA
FAN1_Frag7R	19	CGGCCCAAAAGCTCTCAAG
FAN1_Seq7F	17	CGAGGGAAGTGGCTAAC
FAN1_Seq7R	17	CCTACTTGTGGCCTCTG

Target(s): FAN1 D702E + Q717R		
Primer name	**Length**	**Primer sequence**
FAN1_Frag8F	20	CAGTGAGAGAGCAGAAGAGC
FAN1_Frag8R	19	TGGGTGACAGAGCGAGACT
FAN1_Seq8F	18	AGTGAGAGAGCAGAAGAG
FAN1_Seq8R	19	ACCAAATATCCCCAATTCC

Target(s): FAN1 M50R + V77I		
Primer name	**Length**	**Primer sequence**
FAN1_Frag9F	23	ACTCATGATGTCAGAAGGGAAAC
FAN1_Frag9R	20	ATCACTTTGGCCAGGGGTTA
FAN1_Seq9R	19	TGGCCAGGGGTTAAATTTG

Target(s): FAN1 R982C + C1004G		
Primer name	**Length**	**Primer sequence**
FAN1_Frag10F	20	CAGTGAGAGAGCAGAAGAGC
FAN1_Frag10R	22	ACTGTGTGGAATCAATGAGTGT
FAN1_Seq10F	18	AGTGAGAGAGCAGAAGAG
FAN1_Seq10R	19	TGTGTGGAATCAATGAGTG

Target(s): FAN1 T187fs		
Primer name	**Length**	**Primer sequence**
FAN1_Frag11F	20	TGTTTGGGAAGCCTAGCATC
FAN1_Frag11R	23	ACCATATGTTCAGGAATGCACTC
FAN1_Seq11F	18	TTTGGGAAGCCTAGCATC
FAN1_Seq11R	18	TGTTCAGGAATGCACTCT

Target(s): FAN1 M50R + V77I		
Primer name	**Length**	**Primer sequence**
FAN1_Frag12F	20	TCAGAGTTCGCTTTTCCCCT
FAN1_Frag12R	22	CACACTACGATTTCTCAGCTCA
FAN1_Seq12F	19	ACTCATGATGTCAGAAGGG
FAN1_Seq12R	18	TTGCTGAATCACTTTGGC

Target(s): FAN1 T187fs		
Primer name	**Length**	**Primer sequence**
FAN1_Frag13F	20	GGGAAGTAAAGCAGAAGATCAGT
FAN1_Frag13R	22	TTCTCACATTCCCGGGTAGC
FAN1_Seq13F	19	GCTGAGAAATCGTAGTGTG
FAN1_Seq13R	18	GTTCAGGAATGCACTCTTC

Table 2.5: *FAN1* Sanger sequencing primers. Listed are the primers for PCR used for *FAN1* variant confirmation. For PCR amplification, the fragment (frag) primers were used. For the Sanger sequencing reaction, the sequencing (seq) primers were used.

Target(s): *HTT* exon 1		
Primer name	**Length**	**Primer sequence**
LKH-1	22	CCCATTCATTGCCCCGGTGCTG
LKH-5	22	TGGGTTGCTGGGTCACTCTGTC

Target(s): *HTT* exon 1 CAG		
Primer name	**Length**	**Primer sequence**
huHTT_Ex1F SEQ	17	ATTCATTGCCCCGGTGC
huHTT_Ex1R SEQ	19	CTGGGTCACTCTGTCTCTG

Target(s): EXO1 R108H, R121W		
Primer name	**Length**	**Primer sequence**
EXO1_Frag1F	23	TGTTACAGTTTCTTGAGTCAGCC
EXO1_Frag1R	20	GCTTTGGTGAACTTGCCCAT
EXO1_Seq1F	19	ACAGTTTCTTGAGTCAGCC
EXO1_Seq1R	19	TGGTGAACTTGCCCATCTG

Target(s): EXO1 A137S, D143E		
Primer name	**Length**	**Primer sequence**
EXO1_Frag2F	23	CTGAGGCTAGTAATAAAGTGGGT
EXO1_Frag2R	20	GGAGATCCGAGTCCTCTGTA
EXO1_Seq2F	18	AAGTGGGTTTGAAACAGG
EXO1_Seq2R	17	GAGATCCGAGTCCTCTG

Target(s): EXO1 G223V, D249N		
Primer name	**Length**	**Primer sequence**
EXO1_Frag3F	22	AGATAATGACAAAAGTGGCCCT
EXO1_Frag3R	20	GTGCCTCAGTCATTTGCTCC
EXO1_Seq3F	19	GAAATTGATCAAGCTCGGC
EXO1_Seq3R	18	GCCTCAGTCATTTGCTCC

Target(s): EXO1 G274R		
Primer name	**Length**	**Primer sequence**
EXO1_Frag4F	23	GCAGTTAATGTTTCAATCCCTCT
EXO1_Frag4R	23	CGTAGCTTAGTGTTTCAGGATCA
EXO1_Seq4F	19	CCCTGTTCTTTAGTTGCAG
EXO1_Seq4R	18	TTCATAGGCGTTCAGAGG

Target(s): EXO1 E584K, S610G		
Primer name	**Length**	**Primer sequence**
EXO1_Frag5F	20	GCCTCTGGATGAAACTGCTG
EXO1_Frag5R	21	GACTCCTCGCTCTTTAACTGC
EXO1_Seq5F	18	CTCTGGATGAAACTGCTG
EXO1_Seq5R	17	TCAGGCAAAGAGGTGGG

Target(s): EXO1 1:242048615:G:C, G759E, L790R		
Primer name	**Length**	**Primer sequence**
EXO1_Frag6F	20	GCGACAGAGTGAGAGTCCAT
EXO1_Frag6R	21	AGGAAGAGTTGGGAGAAAGGG
EXO1_Seq6F	18	ACAGAGTGAGAGTCCATC
EXO1_Seq6R	18	AGTTGGGAGAAAGGGATG

Target(s): EXO1 A827V		
Primer name	**Length**	**Primer sequence**
EXO1_Frag7F	23	TGTTACAGTTTCTTGAGTCAGCC
EXO1_Frag7R	22	GCTTTGGTGAACTTGCCCAT
EXO1_Seq7F	18	ACAGTTTCTTGAGTCAGCC
EXO1_Seq7R	18	TGGTGAACTTGCCCATCTG

Target(s): BRCA2 Thr1738fs		
Primer name	**Length**	**Primer sequence**
BRCA2_Frag1F	21	AGTCCTGCAACTTGTTACACA
BRCA2_Frag1R	22	AGAGCTAGTCACAAGTTCCTCA
BRCA2_Seq1F	18	TCCTGCAACTTGTTACAC
BRCA2_Seq1R	19	TTACAGTTTGTGGGTATGC

Target(s): BRCA2 Lys3326*		
Primer name	**Length**	**Primer sequence**
BRCA2_Frag2F	20	GACTGCCTTTACCTCCACCT
BRCA2_Frag2R	20	TCTTCTGAACTGGTGGGAGC
BRCA2_Seq2F	18	ATTTGTTTCTCCGGCTGC
BRCA2_Seq2R	17	TCTGAACTGGTGGGAGC

Target(s): MSH3 Tyr462fs		
Primer name	**Length**	**Primer sequence**
MSH3_Frag1F	23	CTCTCTCTTTCTTCAACTTGGGA
MSH3_Frag1R	20	ATCCTCCCCTACCTCAGTCT
MSH3_Seq1F	18	TTCCTCTTCTGGCCAAGA
MSH3_Seq1R	18	GAGCTCTTCTTCTCCCAC

Target(s): MSH3 SPLICEACCEPTORc.2254-1G>C		
Primer name	**Length**	**Primer sequence**
MSH3_Frag2F	20	TGTCCCAAGTAGTGAACCCT
MSH3_Frag2R	22	TCAAGAATGTGGCTACGATGAG
MSH3_Seq2F	18	GTCCCAAGTAGTGAACCC
MSH3_Seq2R	18	TACGATGAGCCCAGTAGC

Target(s): MSH3 SPLICEACCEPTORc.2319-1G>A		
Primer name	**Length**	**Primer sequence**
MSH3_Frag3F	23	GGAATCAGTAGAGTTCAGGACCA
MSH3_Frag3R	20	CCATTCAGCACTGCAGTCAA
MSH3_Seq3F	19	ATCAGTAGAGTTCAGGACC
MSH3_Seq3R	18	CCATTCAGCACTGCAGTC

Table 2.6: Other Sanger sequencing primers (not *FAN1*). Listed are sequencing primers for *HTT* (targeting the CAG repeat) and primers for *EXO1*, *MSH3* and *BRCA2*. Not all primers were used for sequencing but are listed for any future projects that may find them useful. As before, these primers were designed for initial amplification with the frag primers, followed by clean-up and sequencing with the seq primer(s). LKH-1 and LKH-5 primers were taken from (Swami et al., 2009).

2.11 Genescan sizing of *HTT* CAG

Lymphoblastoid DNA for extreme early/late onset individuals (2.4) were genotyped using a genescan. DNA was prepared to 10-30 ngµL^{-1} using PicoGreen™, and the samples were kindly run and CAG lengths determined by Dr Thomas Massey. Briefly, TaKaRa LA Taq® DNA polymerase with GC buffer (RR02AG, TaKaRa) was used for amplification: the master mix consisted of 1.3 µL deionised water, 5 µL GC buffer (II), 1.6 µL dNTPs (provided), 0.5 µL of each primer and 0.1µL of polymerase. 1.0 µL of each DNA template (~10-30 ngµL^{-1}) was added. The forward primer was ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC and the reverse primer GGCGGCTGAGGAAGCTGAGGA. The PCR programme used was the same as in 2.10.2. 2 µL of each PCR product was added to 27.6 µL Hi-Di fomamide (ThermoFisher, 4311320) and 0.4 µL LIZ600 ladder (ThermoFisher, 4366589) was added to a new plate, and these run on a GA3130*x*L Genetic Analyser (Applied Biosystems) per the manufacturer's instructions. GeneMapper software was used to derive pure CAG lengths from acquired fluorescent traces by taking the (PCR product length – 86)/3 (assuming a canonical **CAA**CAG sequence). 86 nucleotide length comes from the forward (30 nt) and reverse (21 nt) primer length, downstream sequence (29) and the canonical **CAA**CAG sequence (6). Dr Thomas Massey also calculated both the expansion and instability indices, and the derivation of these measures is described elsewhere (Lee et al., 2011).

2.12 Detecting copy number variants in select samples

DNA from select lymphoblastoid and blood DNA was prepared to 200 ngµL^{-1} using Qubit™. DNA was given to the MRC core team at Cardiff University who kindly genotyped the samples using a 24 lane Infinium™ global screening array chip (v2.0) (Illumina, 20024444), with an Infinium™ HTS assay. A full set of protocols and workflows are available online at https://support.illumina.com/array/array_kits/infinium-global-screening-array/downloads.html. Briefly, the 200 ngµL^{-1} DNA was amplified using the provided MA1/MA2/MSM buffers and incubated at 37^{0}C overnight. The resultant DNA was fragmented enzymatically (FMS), precipitated and resuspended. The amplified fragmented, resuspended DNA was then loaded onto the global screening array chip. Following washing and staining, the chip was imaged using an Illumina iScan reader. The MRC core team used a standard CNV calling analytical pipeline using PennCNV (Wang et al., 2007) to detect CNVs using the CRCh38 reference assembly. CNVs with <100kb or <10 total SNPs were excluded.

2.13 Computing facilities and analytical tools

Computational work in most of 2.7.1-2.7.3 and 2.9.1 used the Raven supercomputing cluster of the Advanced Research Computing Division (ARCCA) at Cardiff University. The auxiliary analysis in Appendix 15 used the Hawk supercomputer housed at Cardiff University, part of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) *via* the Welsh government. Long-term storage of WES data (FASTQ, BAM, VCF and VDS files) is in a dedicated storage cluster on the ROCKS computing cluster, housed in the Department of Psychological Medicine and Clinical Neurosciences (DPMCN) at Cardiff University.

R (version 3.6.0; R Core Team 2019, https://www.r-project.org) was used for statistical analyses throughout and the plotting of many of the figures throughout this work. Packages used for figure plotting were the ggplot2, rgl, ggpubr, Hmisc, corrplot, gridExtra, GWASTools (part of Bioconductor (BiocManager)), plot3D, plotrix and RColorBrewer packages. The QuantPsyc, dplyr, tidyr Hmisc, doBy, e1071 and reshape2 packages were used for data manipulation. The SKAT package was used for whole-exome SKAT(-O) analyses. The tools/ databases are described in their relevant methods sections, and listed in Tables 2.7 & 2.8.

Database name	Version	Reference(s)
Registry-HD	R3 cut	http://www.ehdn.org/; (Orth et al., 2010)
dbNSFP	3.0	(Liu et al., 2011, 2016)
ExAC	(*)	(Lek et al., 2016)
CADD	(*)	(Kircher et al., 2014; Rentzsch et al., 2019)
Polyphen	(*)	(Adzhubei et al., 2010)
SIFT	(*)	(Ng and Henikoff, 2003; Sim et al., 2012)
gnomAD	2.0.2	(Karczewski et al., 2019)
GENCODE	Release 19	(Frankish et al., 2019)
Gene Ontology database (GO)	Accessed Apr 2017	(Ashburner et al., 2000; The Gene Ontology Consortium, 2019)

Table 2.7: Databases used for analyses. Listed are the databases used for data acquisition and analysis, and their version (accession if unknown). (*) databases were included with dbNSFP and thus dependent on its version (v3.0).

Tool/program name	Version	Reference(s)
R	3.6.0	https://www.r-project.org/
Hail	v0.1, devel-b08433b	https://github.com/hail-is/
Burrows-Wheeler Aligner (BWA)	0.7.5a	(Li and Durbin, 2009)
Genome analysis toolkit (GATK)	3.4	(McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013)
Picard (picard-tools)	1.97	https://github.com/broadinstitute/picard/
SnpEff	4.1l	(Cingolani et al., 2012)
SnpSift	4.1l	(Cingolani et al., 2012)
VerifyBamID	1.1.3	(Jun et al., 2012)
Peddy	0.3.5	(Pedersen and Quinlan, 2017)
PLINK	1.9	(Purcell et al., 2007)
Variant-effect predictor tool (VEP)	89.6	(McLaren et al., 2016)
SAMtools	1.9	(Li et al., 2009; Li, 2011)
BEDTools	2.24.0	(Quinlan and Hall, 2010)
Scale-HD	0.322	https://github.com/helloabunai/ScaleHD/
Tablet	1.19.05.28	(Milne et al., 2013)
Primer3	Accessed June 2018	(Koressaar and Remm, 2007; Untergasser et al., 2012)
UCSC In-Silico PCR	Accessed June 2018	https://genome.ucsc.edu/cgi-bin/hgPcr

Table 2.8: Bioinformatic tools used for analyses. Listed are all the tools/programs used for analysis (including PCR design) and their version numbers. For primer3 and UCSC in-silico PCR, these are websites and thus the specific version number is unknown; as such, the accession is instead given.

Chapter 3: Phenotypical analysis and cohort selection in Registry-HD

3.1 Introduction

Huntington's disease (HD) is a monogenic neurodegeneration whose hallmark are motor disturbances, cognitive decline and behavioural/psychiatric symptoms (Roos, 2010) (1.2). Mental and behavioural changes are especially disabling for the patient and their family (Vamos et al., 2007; Paulsen et al., 2010; Tabrizi et al., 2013), although the extent to which patients experience these symptoms varies substantially. Notably, age at HD onset can vary considerably between individuals. Although well established that the length of the *HTT* CAG repeat tract is responsible for ~50-70% of this variability (Wexler et al., 2004; Lee et al., 2012c), the remaining variation is accounted for by a combination of genetic and environmental factors (Wexler et al., 2004).

Registry-HD was a multi-site, multi-national longitudinal observational study of Huntington's disease that took place in Europe between 2004 and 2017 (Orth et al., 2010). After an initial baseline visit, participants were assessed annually, and biological samples (*e.g.* blood and urine) were collected from those opting to do so. At its completion, Registry-HD comprised data for >10,000 manifest HD individuals, at-risk individuals and controls. Although now superseded by the international Enroll-HD study, which itself contains Registry individuals who rolled over (currently >18,000 participants), Registry still contains a wealth of clinical data, both cross-sectional and longitudinal, collected from an HD population for more than a decade. These data include demographic information, CAG lengths, TMS/TFC scores and distinct phenotypic data for several symptoms.

Of specific interest for this chapter are the phenotypic data describing symptom onset. Determination of symptom onset allows for stratification of individuals by their disease phenotype and can be factored into cohort selection or *post hoc* analyses in genome-wide association (GWA) studies, sequencing studies, regression modelling or polygenic risk score (PRS) derivation. However, genetic studies, including the GeM-HD GWA studies in 2015 and 2019 and the seminal Venezuelan kindred study (Wexler et al., 2004), have often focused on motor onset as the primary outcome measure. This is in large part as motor symptoms are well-described, measurable and mostly specific to HD. Given both the usefulness of age at onset derivation and the destructive nature of the non-motor symptoms in HD, we were interested in deriving symptomatic ages at onset across a range of symptoms in a large HD patient cohort.

The work presented in chapter 3 will derive ages at onset across the three primary HD symptomatic domains: motor, cognitive and psychiatric/behavioural. The usefulness of the clinical characteristics questionnaire (CCQ), routinely gathered in the Registry and Enroll-HD studies, will be assessed and used to derive distinct symptomatic ages at onset. The degree to which CAG length differentially affects symptom onset will then be examined. Inter-symptom correlations, symptom sex stratification and inter-generational anticipation will also be explored using correlation matrices and generalised linear models. Finally, an extreme onset cohort (N=500) will be selected by stratifying with a residual age at motor onset to select individuals with unexplained early or late HD onset. This cohort will be used for genetic analysis in chapters 4 and 5.

3.2 Initial onset determination in Registry-HD

3.2.1 Exploring onset types in Registry

There are two direct onset measures for Registry-HD participants. The first is the clinical rater's best estimate of onset (sxrater), which describes the onset of the first HD symptom experienced. Onset types in Registry are classified as motor, cognitive, psychiatric, oculomotor, other (*e.g.* weight loss or insomnia) or mixed. Analogous onset estimates are given by both the family and participant. Fig. 3.1 shows, strikingly, that while motor onset is the most common onset in Registry, accounting for 50.3% onsets called by the rater, the remaining 49.7% of onsets were mixed or non-motor. The clinician called psychiatric and mixed onset types more frequently than the participant or family.

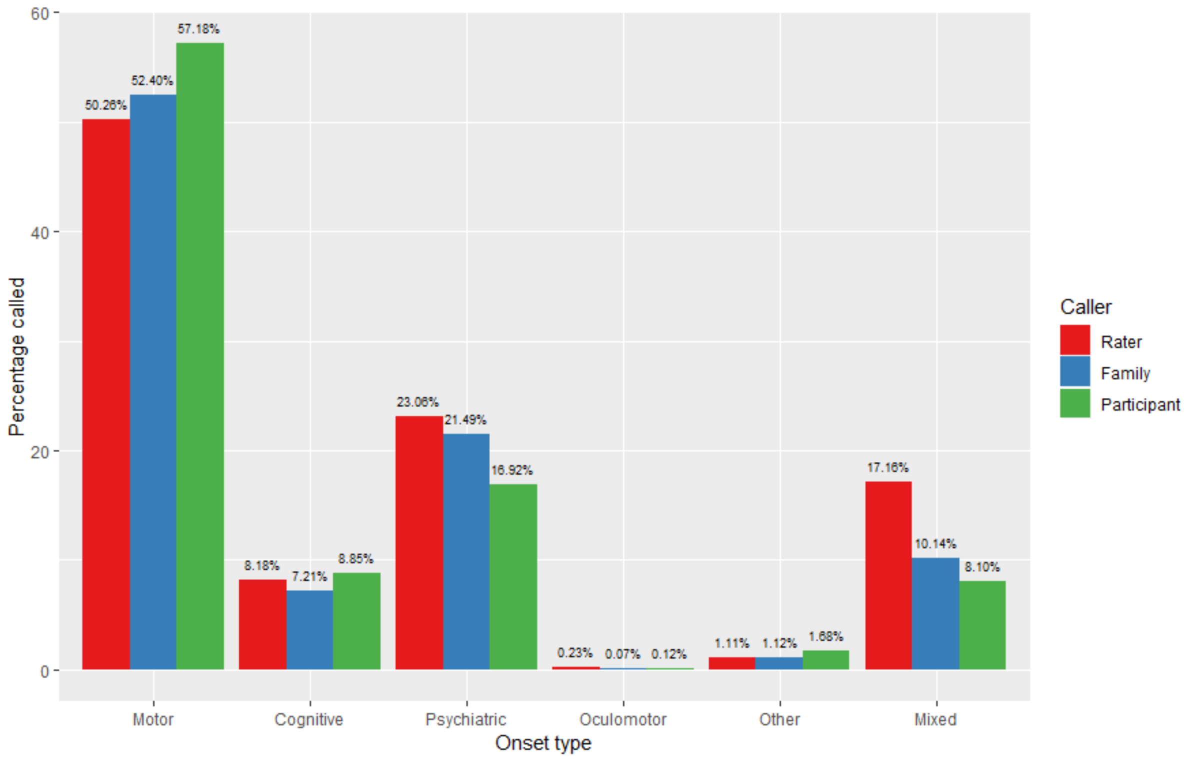


Figure 3.1: HD onset types observed in Registry. Types of onset (first symptom experienced) called by the clinical rater (red), family (blue) and participant (green) in Registry. Rater N=6855; family N=6336; participant N=6365.

In addition to differences between the rater, family and participant, the age at which onset occurs also heavily influences the first symptom experienced (Fig. 3.2). Juvenile HD (JHD) participants (0-20 years) have substantially fewer purely motoric onsets compared to older groups; the oldest onset group, 60-80 years, have motor onset in over two thirds of cases. Cognitive onset is almost three times more likely in JHD cases than any other age group, and both psychiatric and mixed onset types are also more likely in JHD. Psychiatric onset was called at similar levels both in the JHD and 20-40 age groups. Similar patterns emerge between the clinician, family and participant between different age groups.

The second onset estimate from Registry derives from the clinical characteristics questionnaire (CCQ), a retrospective questionnaire completed with the clinician, HD participant and any family present. The CCQ describes the first time a participant has experienced a range of different symptoms including motor, apathy and depressive symptoms (see methods Table 2.2). It is notable that, unlike the rater's estimate of onset, CCQ does not only describe symptoms exclusively caused by HD. Thus, the CCQ can describe symptoms experienced any point in a participant's life before or after clinical diagnosis of HD. This is especially the case for symptoms such as depression which occur at high rates in the general population, making interpretation of the CCQ more difficult. Nonetheless, sxrater itself is quite limited as it only records the first HD symptom onset for each individual, and neglects all later symptoms experienced. We were therefore interested in (1) determining how similar sxrater and CCQ symptom onsets were and (2) if it was possible to use CCQ in conjunction with sxrater for determination of onset across symptomatic domains.

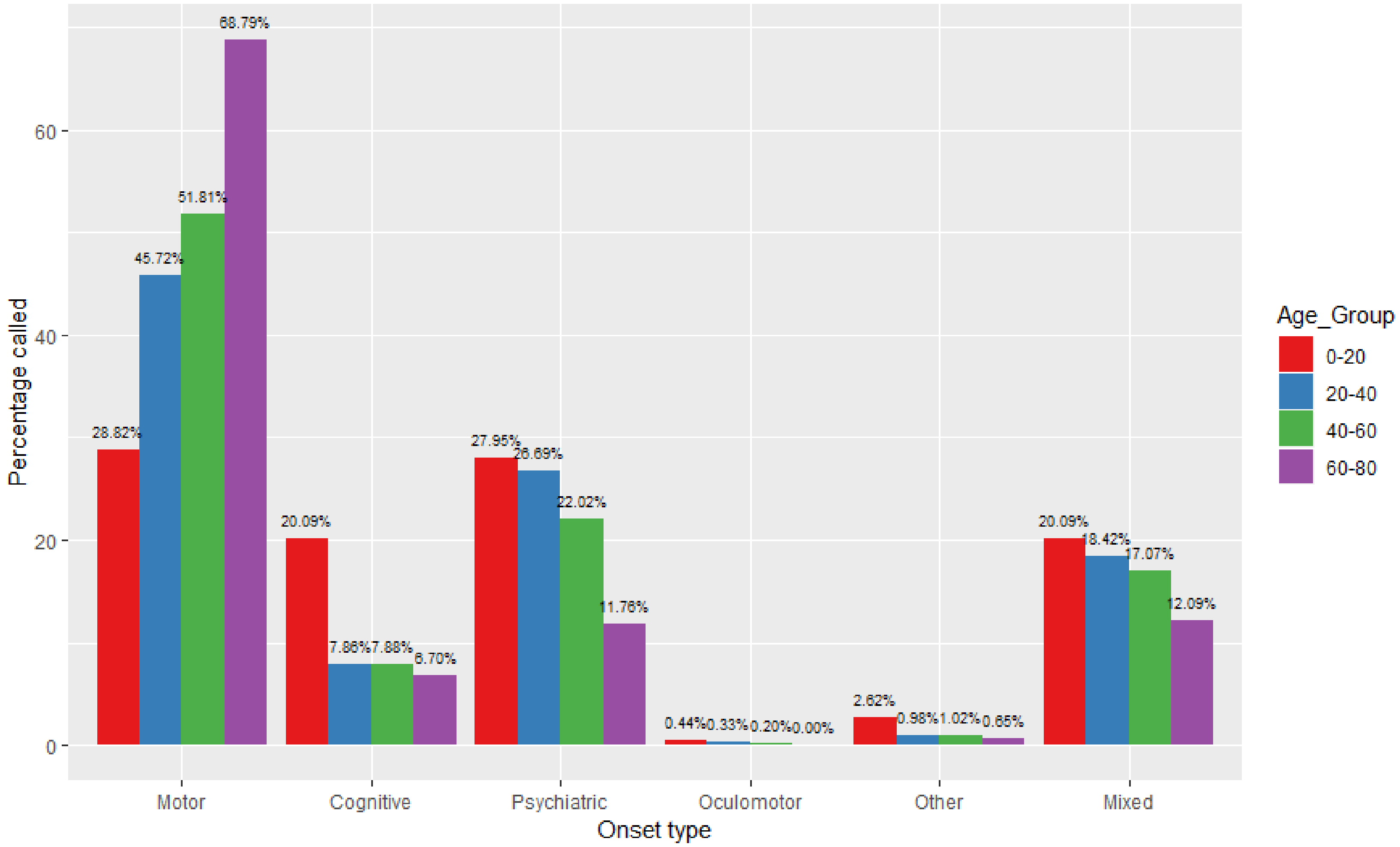


Figure 3.2: HD onset type by age at onset in Registry. The onset types (*i.e.* the first symptom HD patients experienced) as called by the clinician/rater are shown across four age ranges: 0-20 (red; N=229), 20-40 (blue; N=2454), 40-60 (green; N=3515) and 60-80 (purple; N=612) years. Total N=6810.

3.2.2 Rater’s estimate of HD onset

Initially, we wanted to calculate ages at onset determined by the clinical rater, regardless of the onset type. These types of combinatorial approaches to determining ages at onset have been used elsewhere (Illarioshkin et al., 1994; Pekmezovic et al., 2007; Rinaldi et al., 2012), albeit often using different methodologies. Summaries for these analyses can be found in Table 3.1. The mean HD onset was 43.73 years (standard deviation (SD) = 12.50 years).

CAG	N	Mean/yr	Median/yr	Max/yr	Min/yr	SD/yr
36	11	55.67	57.91	78.92	35.42	14.47
37	28	56.81	55.00	73.91	39.00	9.30
38	53	55.38	53.25	84.33	34.00	11.50
39	159	56.16	56.33	84.41	20.42	11.54
40	425	57.18	57.92	86.00	16.67	10.06
41	723	53.48	54.25	78.91	12.42	9.65
42	970	49.95	50.58	74.75	13.25	8.64
43	855	46.15	46.67	75.00	14.92	7.97
44	726	43.05	43.17	67.08	18.00	7.63
45	562	40.21	40.17	58.83	15.41	6.92
46	429	37.48	37.41	67.25	7.50	6.90
47	348	35.50	35.33	64.00	10.75	7.01
48	228	33.29	32.71	71.00	16.17	7.19
49	179	32.04	31.08	68.33	8.91	7.39
50	145	29.81	29.59	62.42	13.59	7.41
51	91	28.20	27.92	44.08	12.33	5.83
52	76	25.91	25.29	49.42	6.41	6.94
53	51	25.45	24.75	40.50	16.33	5.47
54	35	24.94	25.83	37.42	11.59	5.12
55	39	24.99	24.41	38.59	15.84	5.35
56	23	22.55	22.08	30.92	15.75	3.76
57	15	22.73	21.66	45.58	13.92	7.59
58	15	22.12	22.00	41.50	6.92	8.64
59	15	20.39	20.41	30.67	5.17	6.22
60	13	20.96	18.25	33.83	15.33	5.81
61	7	21.55	20.59	31.42	17.08	4.98
62	10	16.60	16.30	24.75	7.92	4.82
63	10	16.57	15.38	28.25	10.84	5.32
64	11	17.45	17.34	26.25	12.75	4.10
65	7	11.12	10.75	15.75	4.58	3.64
66-70	20	15.14	15.21	21.75	7.92	3.89
71-75	9	12.61	7.42	50.51	4.25	14.50
76-80	9	8.17	6.50	24.84	2.75	6.74
>80	7	6.28	6.17	9.83	3.25	2.36

Table 3.1: Onset summaries by CAG (general onset). General onsets (as determined by the rater (sxrater)) are shown. Mean, median, maximum (Max), minimum (Min) and SD are given as years. >65 CAGs have been grouped to avoid identification. SD: Standard deviation

3.2.3 Motor onset derivation (best-estimate)

After calculating ages at onset using the clinician's estimate of onset (sxrater), we wanted to explore specific onset types using both CCQ and sxrater. To calculate an age at motor onset (AMO), the sxrater for motoric onset (motor or oculomotor) was compared to the motor CCQ to derive a 'best-estimate' using the two measures (Fig. 3.3A) – a ≤2y difference was tolerated to account for age estimation. The motor sxrater and CCQ were concordant in 2680 of 2917 cases (91.9%). Discrepancies occurred equally in both directions, with 118 having an earlier motor CCQ and 119 having an earlier sxrater. Onset was resolved through manual curation for 158 individuals (see methods 2.2), resulting in 97.3% of cases with an estimate for AMO. This process was repeated where onset was mixed. In this case, only 787 of 987 participants (79.7%) had ≤2y difference between the mixed sxrater and motor CCQ (Fig. 3.3B). 164 individuals had an earlier sxrater and only 36 had an earlier motor CCQ. Although mixed onset often describes motor symptoms in combination with at least one other symptom, upon investigation we found most discrepant individuals with an earlier sxrater likely had a mixed psychiatric onset phenotype. For these cases, motor CCQ was used for AMO estimation. 172 participants with inconsistent motor onset estimates were subsequently resolved leaving 97.2% with an AMO estimate.

Overall, the motor CCQ agreed with sxrater data where onset was motor, oculomotor or mixed in most cases (~90%), and using a combination of both sxrater and CCQ (*i.e.* best-estimate) resulted in AMO estimates for more individuals (>95%). A best-estimate AMO was then calculated for as many individuals as possible for whom at least either motor CCQ or motor sxrater was available. 6832 participants available had enough data to calculate AMO before quality control (QC) (Fig. 3.4A). Discrepant data were curated as before and those with irresolvable AMO estimates were removed, leaving 6704 participants with an AMO estimate (Fig. 3.4B). The mean AMO for Registry was 44.95 years (SD = 12.47 years). Motor onset ranged from 3 to 87 years.

Of the participants with an AMO, 6520 had CAG information. Plotting AMO against CAG shows the expected inverse correlation (Fig. 3.4C; summarised in Table 3.2). As CAG length increases, generally both the mean AMO and SD decrease (see Fig. 3.4D). Each additional CAG from 40 repeats confers ~2.5-3.5 years earlier onset, although this varies depending on CAG length, and larger CAGs have smaller differences between onsets. On an individual level there still exists a large variability between individuals at the same CAG length. CAGs with partial disease penetrance (36-39 CAGs) have very high SDs, although the numbers of these individuals were quite small. The modal CAG length for Registry was 42 and the median CAG was 43 for CAGs ≥36.

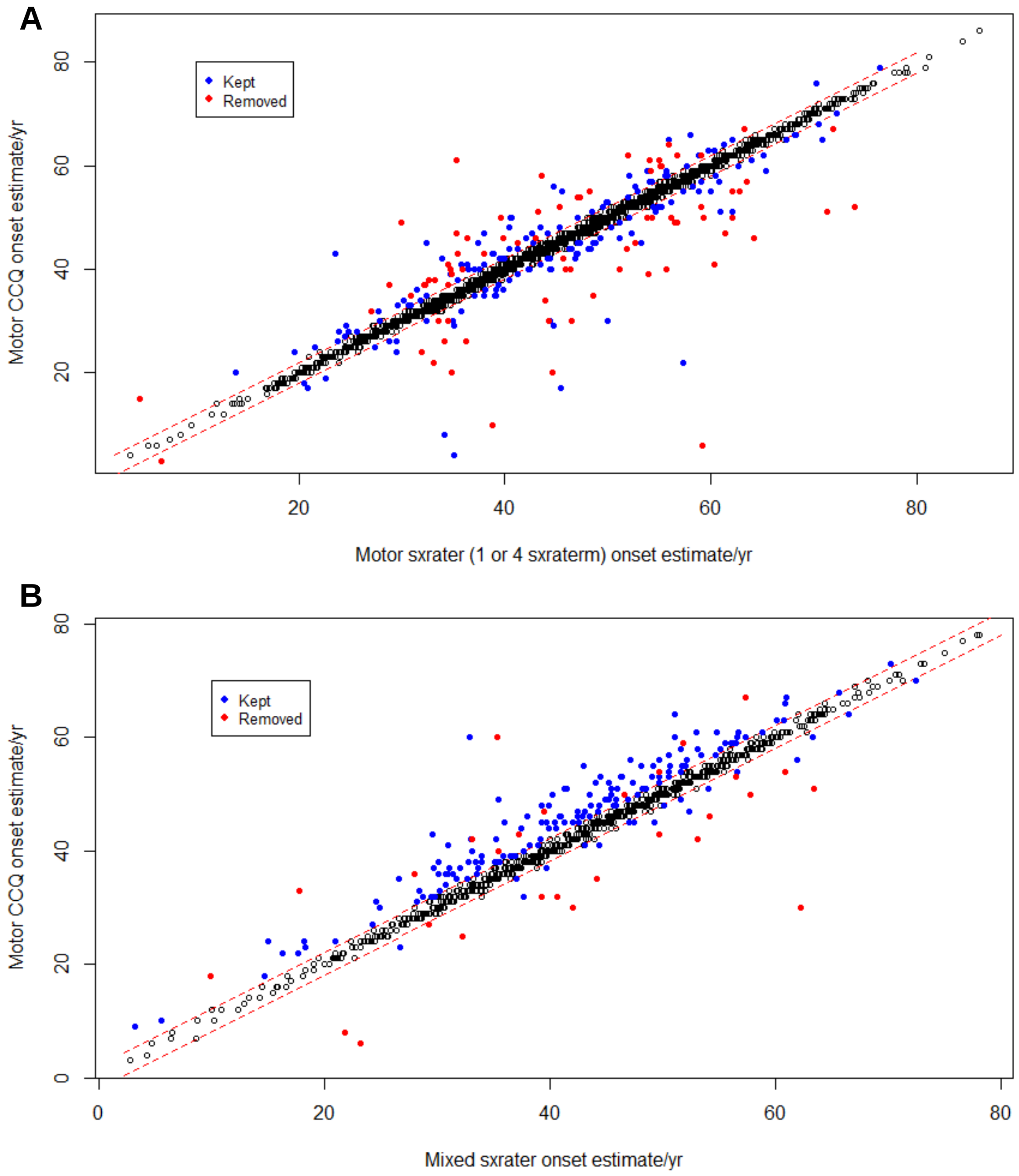


Figure 3.3: Comparison of sxrater and CCQ for motor HD onset. Black circles show participants where the motor CCQ and sxrater are ≤2 years different. The dotted red lines delineate individuals lying outside the ≤2 year range, shown as coloured dots. Blue dots indicate individuals where a best-estimate AMO could be resolved despite the discrepancy, whereas red dots are individuals where AMO proved impossible to estimate with sufficient accuracy. Motor and oculomotor onset types (sxraterm 1 or 4) are shown in (A), N=2917; mixed onset types (sxraterm 6) are shown in (B), N=987. In (A), 39 with earlier sxrater and 40 with earlier motor CCQ could not be resolved. In (B), 12 with earlier sxrater and 16 with earlier motor CCQ could not be resolved.

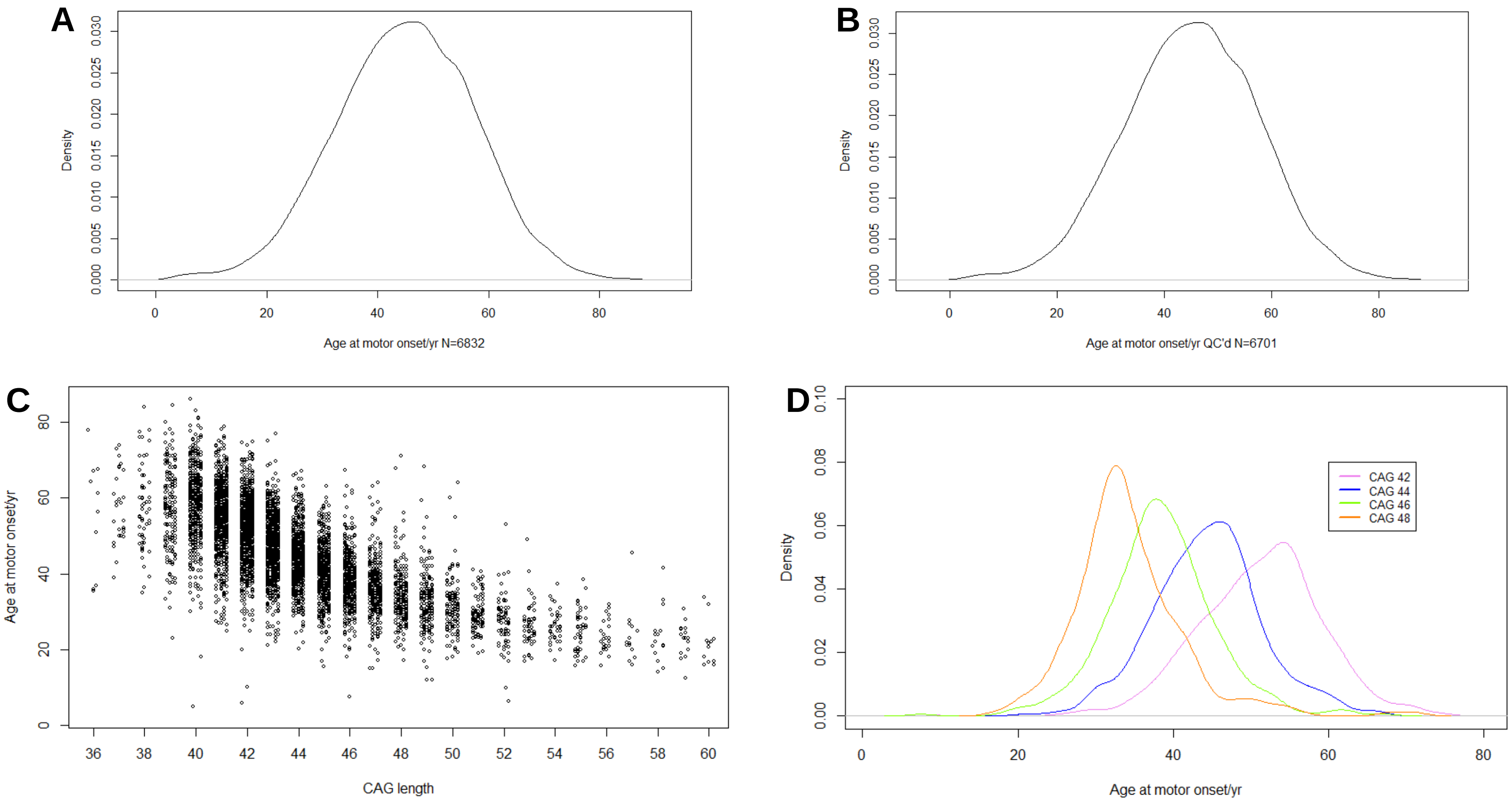


Figure 3.4: Motor onset in Registry across different ages and CAG repeat lengths. Indicated in (A) and (B) are the motor onset distributions in Registry before and after quality control (see Fig. 3.3), respectively (N=6832 and N=6701). Ages at onset calculated from (B) are then plotted against CAG length, where possible, for CAGs 36-60 in plot (C), N=6520. The points have been jittered to improve data point visibility. (D) shows the AMO distribution for CAGs 42, 44, 46 and 48.

CAG	N	Mean/yr	Median/yr	Max/yr	Min/yr	SD/yr
36	10	55.36	58.79	78.00	35.42	15.08
37	28	57.39	56.00	73.91	39.00	9.29
38	55	57.34	56.17	84.00	35.00	11.03
39	173	57.28	57.25	84.41	23.00	10.82
40	459	58.14	59.16	86.00	5.00	9.78
41	767	54.79	55.17	78.91	25.00	8.87
42	1016	51.13	51.88	74.75	6.00	8.10
43	887	47.27	47.67	77.00	21.17	7.67
44	737	44.36	44.50	67.08	20.17	7.04
45	573	40.98	41.00	63.00	15.41	6.95
46	441	38.41	38.00	67.25	7.50	6.99
47	359	36.35	36.00	64.00	15.00	7.05
48	231	34.25	33.08	71.00	17.25	7.15
49	180	32.83	32.21	68.33	12.00	7.43
50	143	30.99	30.42	64.00	17.67	6.77
51	93	29.28	29.00	41.00	19.08	4.76
52	70	27.24	27.71	53.00	6.41	6.76
53	52	26.42	25.96	49.00	17.00	5.67
54	36	26.17	25.88	37.42	17.17	4.33
55	41	25.10	25.50	38.59	15.84	5.50
56	23	23.56	22.83	32.00	15.75	4.33
57	14	23.63	21.75	45.58	16.00	7.26
58	15	23.63	22.00	41.50	14.00	7.26
59	16	22.62	23.04	30.67	12.42	4.45
60	12	21.56	21.12	33.83	16.00	5.87
61	7	21.63	20.59	32.00	17.08	5.17
62	11	18.31	18.00	24.75	13.00	3.70
63	11	15.76	14.25	28.25	8.00	5.69
64	10	17.97	17.75	25.00	13.33	3.64
65	7	11.56	14.33	15.75	4.58	4.66
66-70	20	15.96	15.63	21.75	12.00	3.01
71-75	8	15.04	9.79	50.51	4.25	14.82
76-80	8	11.47	7.00	50.00	2.75	15.75
>80	7	7.30	8.00	10.00	4.00	2.15

Table 3.2: Age at motor onset summary statistics. Mean, median, maximum (Max), minimum (Min) and SD are given as years. CAGs larger than 65 have been grouped to avoid identification. N=6520. SD: Standard deviation.

3.2.4 Cognitive onset determination (best-estimate)

We employed a similar strategy for deriving best-estimate age at cognitive onset (ACO). As before, we first wanted to assess how well the cognitive CCQ tracked the rater's estimate for cognitive onset. We examined individuals where onset was classed as cognitive by the rater, although this reduced the number of participants substantially as only 8.2% had cognitive onset as the first HD symptom. Cognitive CCQ and sxrater were the same (≤2 years different) in 285 of 373 cases (76.4%; Fig. 3.5). Discrepant data occurred unevenly with 77 having an earlier sxrater and only 11 having an earlier cognitive CCQ. 78 of these discrepant cases were resolved, many by simply using the CCQ for cognitive symptoms when this was concordant with other CCQ data (see methods 2.2), giving an ACO for 97.3% of cases.

Overall, the sxrater and cognitive CCQ did not track as well for ACO estimation (*i.e.* >2 year difference in more cases) as they did for motor symptoms. Notably, sxrater tended to be slightly earlier than cognitive CCQ. Despite this, ACO estimates from CCQ/sxrater were still within 2 years in approximately three quarters of cases. ACOs were then calculated for every individual possible, with most estimates using cognitive CCQ. 3853 individuals had ACOs before QC and 3788 following QC (Fig. 3.6A and 3.6B). Plotting these gives a near-normal distribution as before (SD = 13.53 after QC), although with a broader distribution than seen with AMO. Mean ACO was 47.1 years. CAG length data was available for N=3552 between CAG 36-60 (Fig. 3.6C); a summary is available in Table 3.3.

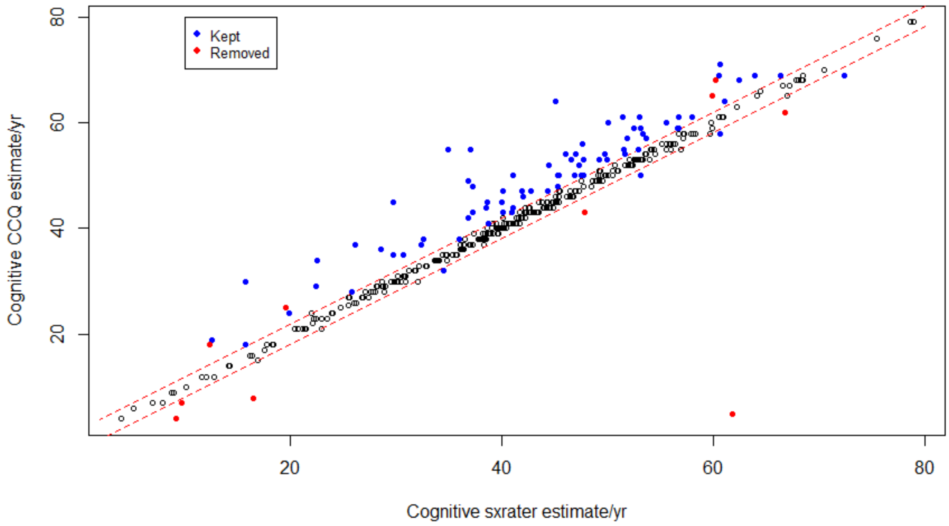


Figure 3.5: Comparison of sxrater and CCQ for cognitive HD onset. Indicated by black circles are participants where cognitive sxrater and CCQ are ≤2 years of each other. Coloured dots lying outside the dashed red lines indicate a >2 year difference. Blue dots were participants where an ACO was possible to calculate, whereas red dots were individuals excluded from the analysis. N=373 (only individuals with cognitive HD onset). 6 individuals with an earlier sxrater and 4 with an earlier CCQ couldn't be resolved.

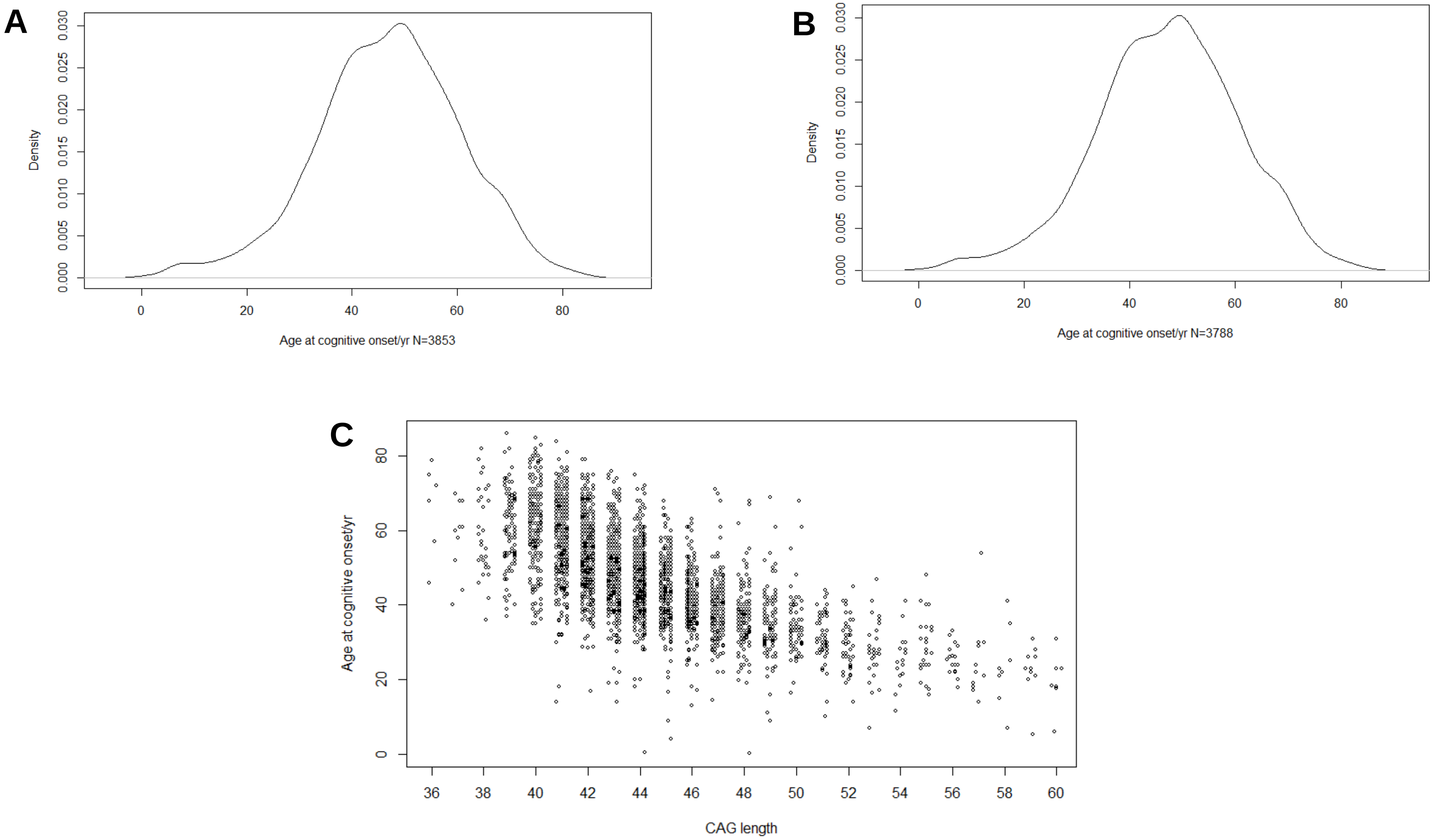


Figure 3.6: Cognitive onset in Registry across different ages and CAG repeat lengths. (A) and (B) show frequency distributions of ACO across age ranges before and after QC, respectively (N=3853 and N=3788). The ACO is plotted against CAG length for CAGs 36-60 in (C), N=3552.

CAG	N	Mean/yr	Median/yr	Max/yr	Min/yr	SD/yr
36	6	66.15	70.00	78.92	46.00	12.40
37	10	58.20	60.50	70.00	40.00	10.12
38	29	59.57	57.00	82.00	36.00	11.88
39	93	60.22	60.00	86.00	37.00	10.43
40	234	61.64	62.29	85.00	35.00	10.56
41	397	57.36	58.00	84.00	14.00	10.29
42	557	54.19	55.00	79.00	16.91	8.70
43	503	50.21	50.00	76.00	14.00	9.00
44	420	47.02	47.00	75.00	0.33	8.78
45	297	44.00	44.00	68.00	4.00	8.70
46	243	41.38	41.00	63.00	13.00	7.37
47	219	39.10	39.00	71.00	14.34	7.45
48	133	36.58	36.00	68.00	0.08	8.68
49	107	35.19	35.00	69.00	8.91	8.63
50	78	34.68	33.46	68.00	16.33	7.96
51	52	31.82	31.00	44.08	10.00	6.89
52	43	29.49	30.00	45.00	14.00	6.82
53	32	28.12	28.00	47.00	7.00	7.67
54	16	25.39	24.83	41.00	11.59	7.34
55	27	28.68	27.00	48.00	16.00	7.75
56	18	25.75	25.71	33.00	17.75	4.29
57	11	25.27	22.00	54.00	14.00	10.96
58	8	23.61	22.50	41.00	6.92	10.67
59	12	22.60	23.00	31.00	5.17	6.30
60	7	19.54	18.25	31.00	6.00	7.60
61	3	27.33	24.00	38.00	20.00	9.45
62	7	17.42	17.00	27.00	7.92	6.26
63	6	16.22	16.00	24.00	10.00	5.38
64	7	17.71	18.25	21.00	12.75	3.22
65	3	12.58	13.00	14.00	10.75	1.67
66-70	15	16.91	16.00	25.00	10.08	4.60
71-75	6	7.67	8.00	10.00	5.00	1.66
76-80	7	14.95	7.00	59.00	6.00	19.49
>80	6	6.47	6.96	10.00	3.00	2.60

Table 3.3: Age at cognitive onset summary statistics. Mean, median, maximum, minimum and SD are given as years. Very large CAGs have been grouped together to avoid identifying data. N=3612. SD: Standard deviation.

3.2.5 Psychiatric onset determination (best-estimate)

Best-estimate age at psychiatric onset (APO) was calculated differently than AMO and ACO. Firstly, CCQ for psychiatric and behavioural symptoms is split into six separate symptoms: apathy (APT), depression (DEP), irritability (IRB), violent or aggressive behaviour (VAB), perseverative or obsessive behaviour (POB) and psychosis (PSY). In contrast, the sxrater does not distinguish between specific psychiatric/behavioural phenotypes in psychiatric onset types. We first wanted to derive an onset for the first psychiatric/behavioural symptom experienced, regardless of specific phenotype, to be equivalent to the best-estimate AMO/ACO measures previously calculated. Plotting these data shows that, as expected, psychiatric CCQ data are often earlier than the sxrater's estimate for APO where onset was classed as psychiatric (Fig. 3.7A). This is likely due to the CCQ capturing symptoms occurring earlier in an individual's life, which may or may not be related HD pathology. Calculating the age when the first psychiatric symptom was experienced (Fig. 3.7B) for the entire dataset (using the earliest of psychiatric sxrater or psychiatric CCQ) results in slight negative skew (skew = -0.060), although with a pronounced incidence of psychiatric symptoms around 15-25 years. The mean and SD were 41.88 and 14.13 years, respectively.

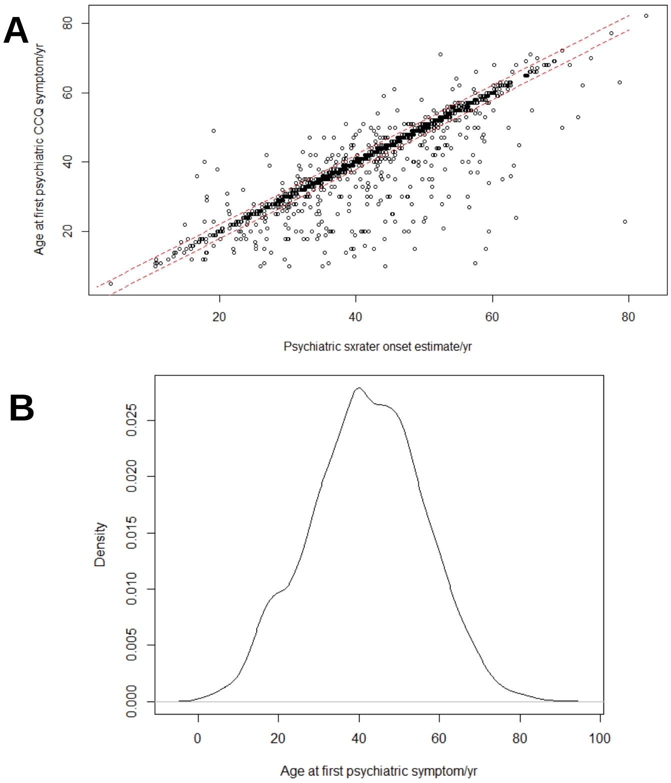


Figure 3.7: Initial APO estimation (before adjustment with sxrater). (A) Psychiatric sxrater and the age of the first psychiatric CCQ symptom. The dashed red line indicates a >2 year between sxrater and CCQ. Only individuals with psychiatric onset types are included, N=1285. (B) shows the distribution of individuals using the age at first psychiatric onset, either sxrater or psychiatric CCQ, whichever is earliest. N=6179.

Trying to disentangle which psychiatric symptoms are part of HD pathology and which are from other psychiatric disorders is very difficult and is discussed in more detail later (3.8.3). For our purposes, we used the first psychiatric symptom which occurred <2 years earlier or at any time later than sxrater, to select for symptoms occurring around or after HD diagnosis only. For individuals where sxrater was earliest and onset type psychiatric, this was instead used. Adjusting for these data result in a more normal Gaussian function (skew = -0.052) with a mean APO at 45.39 years and SD of 12.84 years (Fig. 3.8A). Again, plotting onsets against CAG lengths produces a graph with an inverse correlation (Fig. 3.8B). A full summary by CAG length is available in Table 3.4.

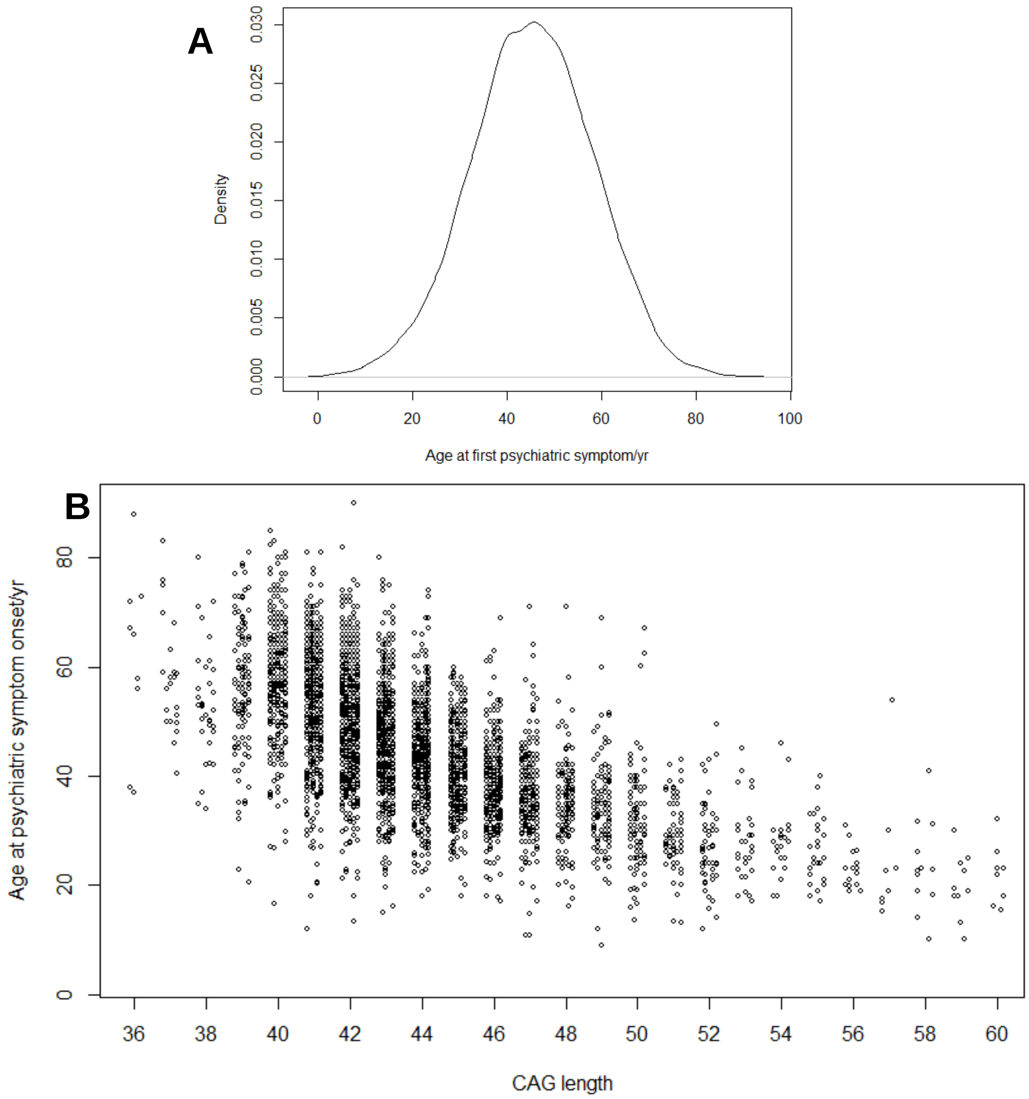


Figure 3.8: Initial APO estimation (after adjustment with sxrater). (A) Distribution of first psychiatric symptom after adjustment for sxrater (only using CCQ data ≤2 years earlier than sxrater); N=4836. (B) Age at first psychiatric after adjusting for sxrater across 36-60 CAGs; N=4563.

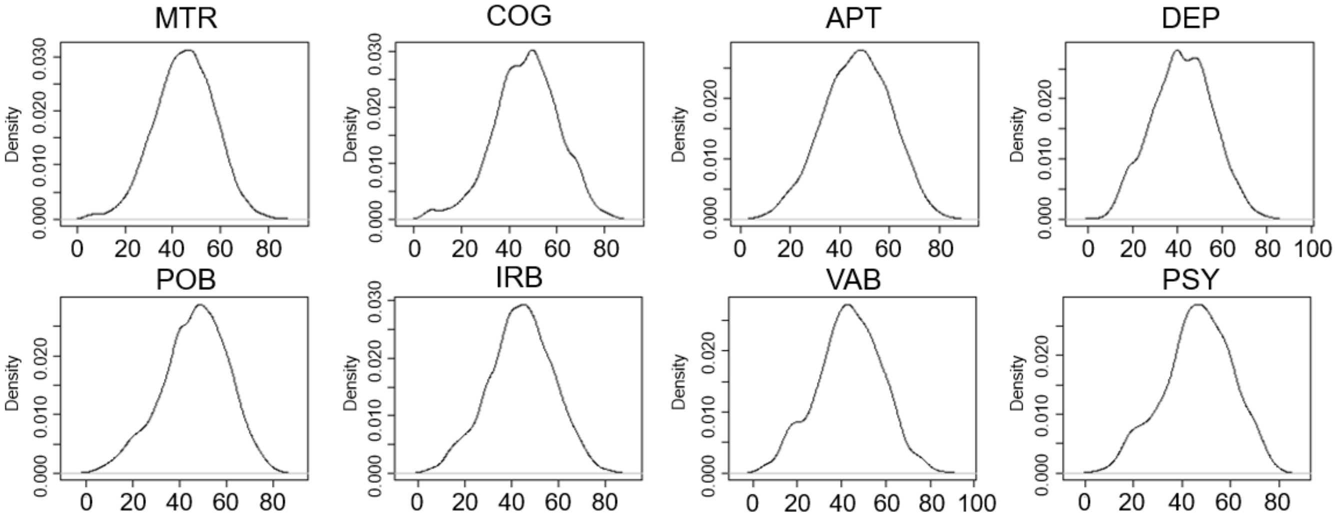
CAG	N	Mean/yr	Median/yr	Max/yr	Min/yr	SD/yr
36	9	61.66	66.00	88.00	37.00	16.55
37	22	58.22	57.50	83.00	40.50	10.64
38	35	52.90	52.00	80.00	34.00	10.34
39	117	56.18	56.08	81.00	20.42	11.70
40	302	58.86	59.34	85.00	16.67	10.83
41	527	54.71	56.00	81.00	12.00	10.41
42	722	51.60	52.00	90.00	13.25	9.70
43	634	47.95	48.00	80.00	14.92	9.26
44	539	44.75	45.00	74.00	18.00	8.47
45	421	42.13	42.00	60.00	18.00	7.32
46	322	39.21	39.00	69.00	17.00	7.49
47	267	36.97	37.00	71.00	10.75	7.72
48	158	34.97	34.92	71.00	16.17	7.32
49	130	34.28	34.00	69.00	9.00	7.91
50	102	31.57	30.00	67.00	13.59	8.77
51	66	29.30	29.00	43.00	13.00	6.31
52	52	27.71	26.84	49.42	12.00	7.81
53	33	27.64	27.00	45.00	17.00	6.76
54	23	27.71	28.00	46.00	18.00	6.46
55	27	26.11	25.00	40.00	17.00	6.23
56	16	23.46	23.00	31.00	19.00	3.81
57	8	24.78	20.83	54.00	15.09	12.70
58	12	23.46	22.71	41.00	10.00	8.32
59	11	20.36	19.42	30.00	10.00	5.73
60	8	21.91	22.50	32.00	15.33	5.54
61	3	26.14	24.00	31.42	23.00	4.60
62	7	16.86	17.00	21.00	12.00	2.90
63	5	18.60	21.00	23.00	12.00	4.51
64	10	17.48	15.76	26.25	13.00	4.75
65	5	13.70	13.00	16.00	12.00	1.72
66-70	13	18.08	19.00	27.00	10.58	4.77
71-75	6	11.22	12.00	17.00	6.00	4.49
76-80	7	8.69	7.00	24.84	3.00	7.49
>80	6	8.50	8.50	14.00	3.00	3.83

Table 3.4: Age at first psychiatric symptom summary statistics. Mean, median, maximum (Max), minimum (Min) and SD are given as years. Very large CAGs have been grouped together to avoid identifying data. N=4625. SD: Standard deviation.

3.3 Deriving symptom onset using the CCQ

3.3.1 Adjusting CCQ for earlier symptom onset

The CCQ holds more descriptive data than sxrater as (1) data are available regardless of onset type, and (2) CCQ captures different psychiatric phenotypes. Simply using CCQ without any modification (Fig. 3.9) results in negative skew (longer left-hand tails) for all psychiatric symptoms, especially between 15-25 years of age. This effect is particularly pronounced for perseveration (-0.34 skew), VAB (-0.14) and psychosis (-0.28). Apathy has a more symmetrical distribution (-0.11 skew) and the lowest standard deviation (13.58 years). Depression has the smallest skew of all the symptoms (-0.024), the second lowest standard deviation (13.60 years) and the earliest symptom onset (42.06 years). A breakdown by CAG length is available in Appendix 1.



Symptom	N	Mean/yr	Median/yr	SD/yr
MTR	6324	44.92	45	12.50
COG	3692	47.27	48	13.61
APT	3413	46.92	47	13.58
DEP	4528	42.06	42	13.60
POB	2325	46.44	47	14.02
IRB	4018	43.97	44	13.80
VAB	2062	43.42	44	14.62
PSY	762	46.46	47	13.99

Figure 3.9: Distribution of psychiatric symptoms in Registry (before sxrater adjustment). Mean, median and SD are given as years. MTR: Motor (N=6324); COG: cognitive (N=3692); APT: apathy (N=3413); DEP: depression (N=4528); POB: perseverative/ obsessive behaviour (N=2325); IRB: irritability (N=4018); VAB: violent/aggressive behaviour (N=2082); PSY: psychosis (N=762).

As already observed, CCQ can capture bias from symptoms potentially unrelated to actual HD pathogenesis, especially for depression, which has a ~10% lifetime prevalence in the general population (Salk et al., 2017; Lim et al., 2018), and psychosis, which has a ~0.5-0.8% lifetime prevalence (Messias et al., 2007; Moreno-Küstner et al., 2018). To partially overcome this limitation, we implemented a similar methodology to APO estimation where CCQ data would be removed if it occurred much earlier than the sxrater. Figure 3.10 shows 2, 5 and 10 year cut-offs and how they affect which data are kept or removed for each symptom. A breakdown for each CAG for the three cut-offs is available in Appendix 1. The most stringent (≤2 year) cut-off was chosen to minimise symptoms occurring that may be unrelated to HD pathology, although this cut-off probably removes useful data (*i.e.* symptoms occurring before motor onset due to HD) as well; the limitations of this approach are discussed later (3.8.3).

As shown in Fig. 3.11, adjusting for sxrater using a 2 year cut-off produces more normally distributed psychiatric onset data, and for all phenotypes decreased the SD and increased the average onset (between 0.2 to 3.8 years later). Motor and cognitive symptoms are the most unchanged symptom following adjustment by sxrater. Mean apathy symptom onset is the most unchanged of the psychiatric/behavioural symptoms following adjustment (1.43 years later). Again, this is likely as apathy innately captures more HD-specific pathology. In contrast, average depression onset is the most changed following modification using the 2 year sxrater cut-off (3.82 years later), reflecting that many individuals had depressive symptoms before HD onset, although it is unclear whether these depression symptoms are due to HD pathology ahead of onset or for other unrelated reasons. Despite the sxrater modification of the CCQ, depression is still the earliest psychiatric symptom experienced by most HD patients. Comparatively, irritability and VAB occur on average half a year later than depression symptoms with apathy, perseveration/obsessive behaviour (POB) and psychosis occurring the latest. Looking across CAG lengths (Appendix 1), psychosis in particular has substantial variability across CAG length, especially for larger CAGs (~50+ CAGs) – it is unclear whether this is a true effect or simply due to small sample size.

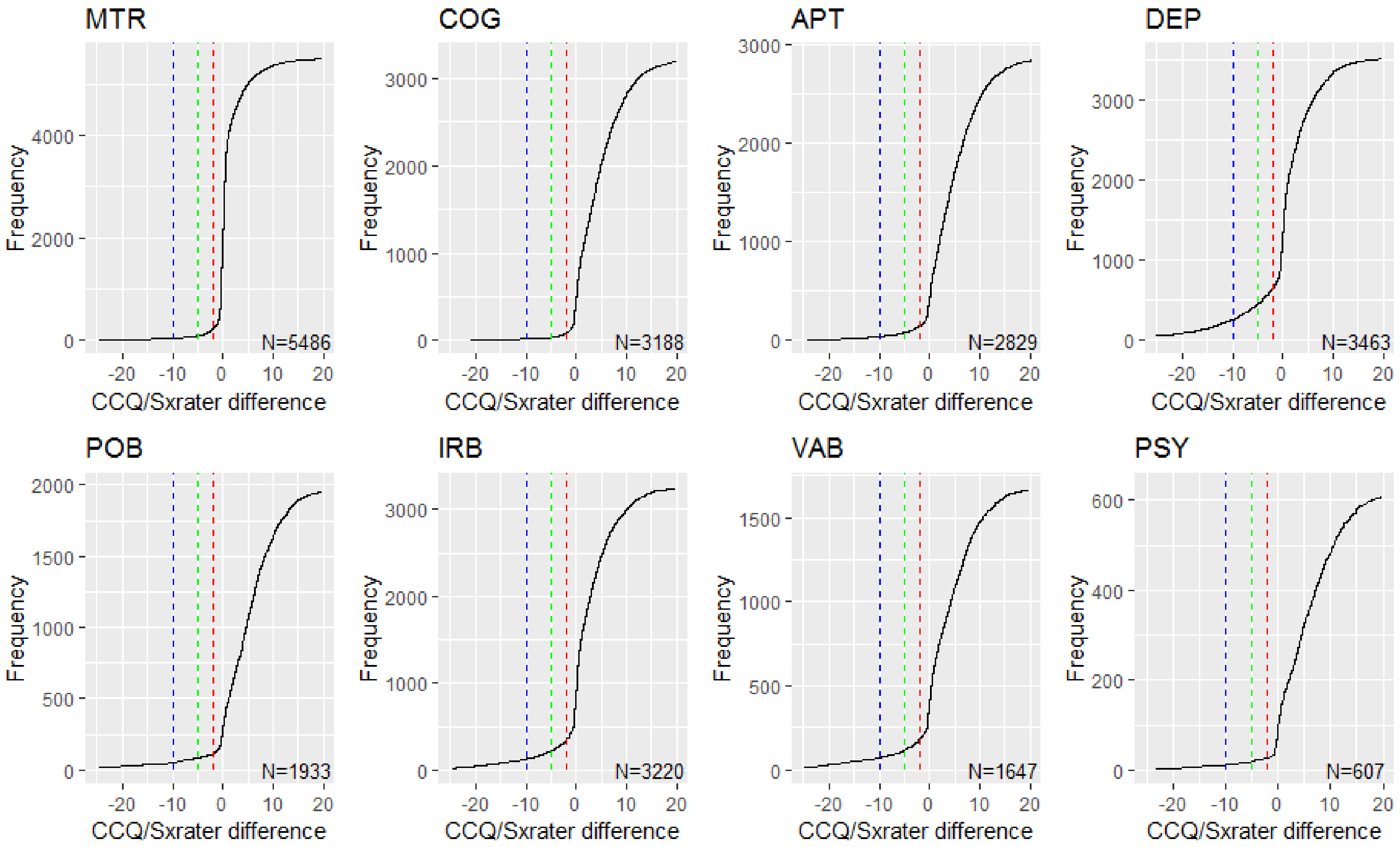
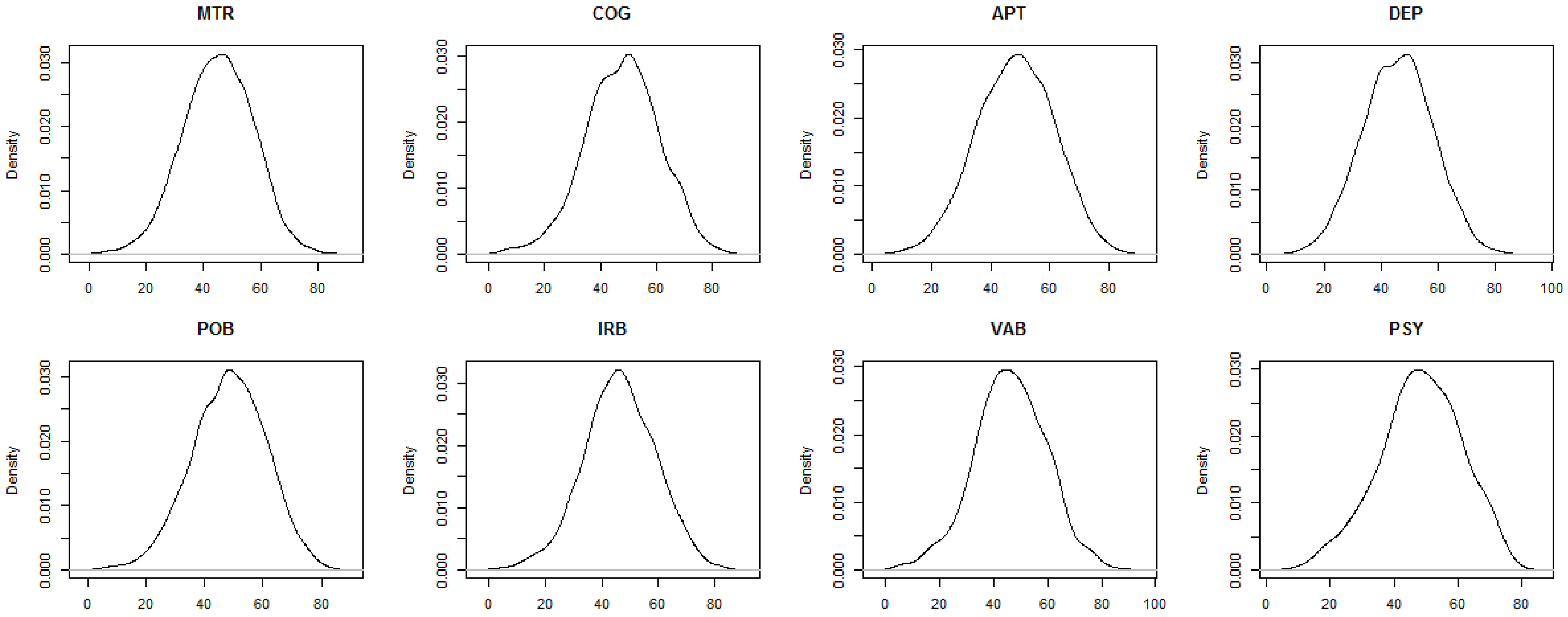


Figure 3.10: Comparing onset estimation between sxrater and CCQ. The red, green and blue dashed lines show the 2, 5 and 10 year CCQ/sxrater cut-offs, respectively, with the x-axis showing CCQ-Sxrater difference (in years). 0 on the graph represents the clinician's estimate of onset (sxrater). MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis. Axis limited to -25y and 20y which removed MTR (32), COG (54), APT (40), DEP (70), POB (51), IRB (38), VAB (34), PSY (22). Individuals had to have a CCQ age at onset, sxrater and a known CAG length 36-99.



Symptom	N	Mean/yr	Median/yr	SD/yr	Symptom	N	Mean/yr	Median/yr	SD/yr
MTR	5471	45.13	45	12.22	POB	1940	48.34	49	12.63
COG	3271	47.84	48	13.21	IRB	3029	46.30	46	12.75
APT	2815	48.35	49	12.97	VAB	1566	46.18	46	13.29
DEP	2984	45.88	46	12.22	PSY	627	48.36	49	12.95

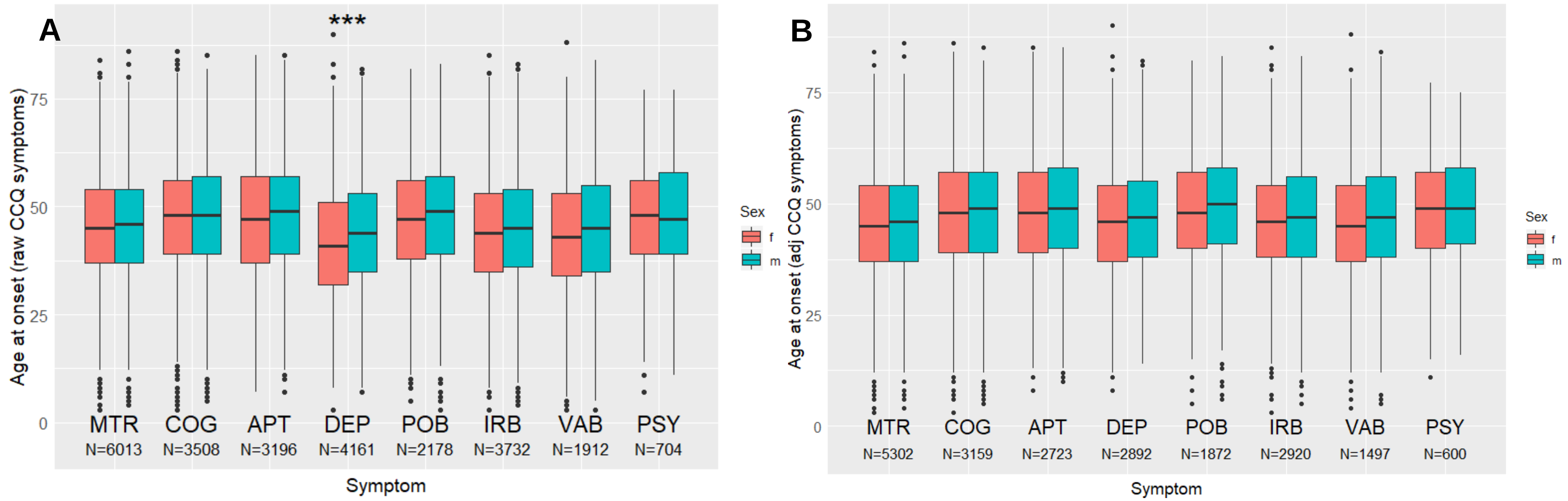
Figure 3.11: Distribution of psychiatric symptoms in Registry (after sxrater adjustment). CCQ data plotted following adjustment with sxrater, which removes any CCQ data that is >2 years earlier than sxrater, with summary statistics in the table (mean, median and SD given as years). MTR: Motor (N=5471); COG: cognitive (N=3271); APT: apathy (N=2815); DEP: depression (N=2984); POB: perseverative/obsessive behaviour (N=1940); IRB: irritability (N=3029); VAB: violent/aggressive behaviour (N=1566); PSY: psychosis (N=627).

3.3.2 The relationship between age at onset of different symptoms

We next wanted to examine whether there were sex-dependent differences in symptom ages at onset. To do so, generalised linear models (GLMs) were constructed for each symptom onset regressing sex on CCQ onset data before and after sxrater adjustment (Table in Fig. 3.12). Using unadjusted CCQ data shows only depression has a significant difference between males and females after multiple testing correction (*p*=6.97E-08; *p*<6.25E-03 Bonferroni *p*-value threshold); POB has nominal significance. This is illustrated in Fig. 3.12A which shows a mean 2.52 year difference between males and females for depression and 1.49 years for POB. Following adjustment of CCQ data using the sxrater, however, no symptom remains significant for sex. Hence sex differences in depression onset are likely driven by earlier onset depression in some women, and this fits with general population observations (Salk et al., 2017).

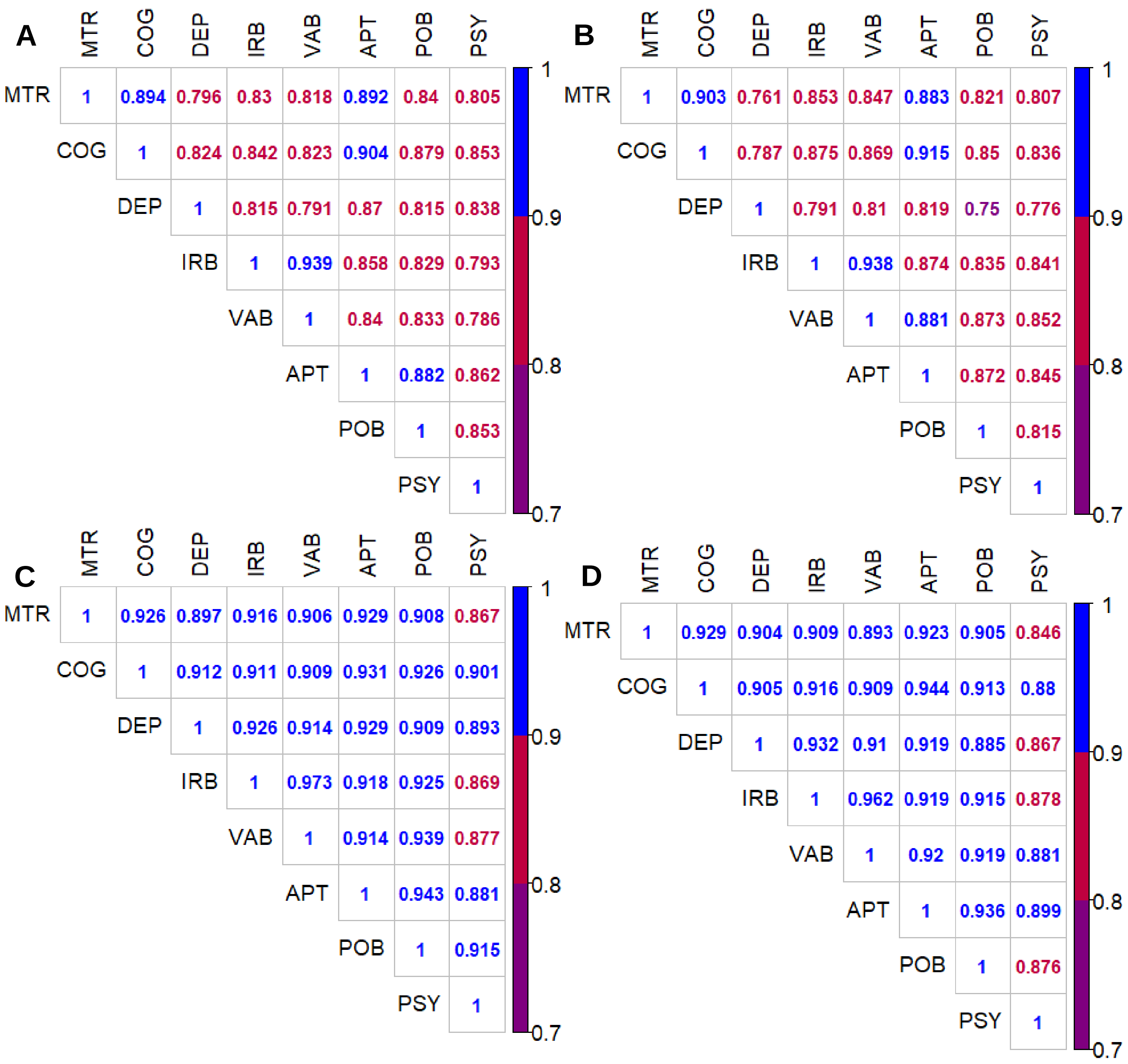
Correlation matrices were produced using a pairwise deletion method for males and females using sxrater adjusted and unadjusted CCQ data. Pairwise deletion minimises loss of data caused by missingness, and only removes missing data in the model being evaluated. All correlations were very high (0.75-0.97) as the differences in ages at different symptom onsets were often small. Starting with the correlations for unadjusted data (Fig. 3.13A-B), the lowest association is between POB and depression for females, likely reflecting that (1) perseveration typically occurs later in disease course and (2) depression is common in women in the general population. Depression correlates more highly with other symptom onsets in male unadjusted data. Irritability and VAB have the highest correlations with each other, which is unsurprising given VAB could be an extreme manifestation of irritability. Furthermore, cognitive, apathy and motor symptoms also have high inter-symptom correlations, likely as these are more distinct to HD or related to the same degenerative pathology. This observation mirrors what was seen in the symptom distribution data.

Applying the <2 year sxrater cut-off (Fig. 3.13C-D) unsurprisingly increases the correlation between all symptoms. Psychosis, and to a lesser degree depression, however, are the least associated with other symptoms. The association between irritability and VAB ages at onset is extremely high (0.96-0.97), and the motor/apathy/cognitive triad association remains high. Adjusting for different sxrater cut-offs (2, 5 and 10 years, see Appendix 2) does not greatly modify the correlations observed, although correlations are slightly lowered at less stringent cut-offs. This indicates CCQ data driving the significant sex differences in depression and POB onset before sxrater adjustment probably occur much earlier than HD onset (*i.e.* >10 years earlier).



	GLM(Sex~CCQ_raw+CAG)		GLM(Sex~CCQ_adj+CAG)	
	CCQ_raw	CAG	CCQ_adj	CAG
MTR	2.59E-01	8.73E-02	3.47E-01	9.09E-02
COG	2.76E-01	1.67E-01	5.16E-01	2.82E-01
APT	5.14E-02	9.59E-01	2.19E-01	5.40E-01
DEP	**6.97E-08**	6.38E-01	3.02E-01	1.11E-01
POB	*4.38E-02*	9.49E-01	3.59E-01	5.53E-01
IRB	6.46E-01	*3.43E-02*	8.95E-01	*2.65E-02*
VAB	5.60E-01	1.78E-01	6.57E-01	1.38E-01
PSY	4.73E-01	3.34E-01	9.87E-01	6.06E-01

Figure 3.12: Sex differences for ages at onset for symptoms in Registry. (A) Before adjustment with sxrater (CCQ_raw); (B) after adjustment with sxrater (2 year cut-off) (CCQ_adj). Unadjusted data in the table refers to raw CCQ data; adjusted data uses an sxrater adjustment (2 year cut-off). Significant values are emboldened; nominally significant values are italicised. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.



A

	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.894	0.796	0.83	0.818	0.892	0.84	0.805
COG		1	0.824	0.842	0.823	0.904	0.879	0.853
DEP			1	0.815	0.791	0.87	0.815	0.838
IRB				1	0.939	0.858	0.829	0.793
VAB					1	0.84	0.833	0.786
APT						1	0.882	0.862
POB							1	0.853
PSY								1

B

	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.903	0.761	0.853	0.847	0.883	0.821	0.807
COG		1	0.787	0.875	0.869	0.915	0.85	0.836
DEP			1	0.791	0.81	0.819	0.75	0.776
IRB				1	0.938	0.874	0.835	0.841
VAB					1	0.881	0.873	0.852
APT						1	0.872	0.845
POB							1	0.815
PSY								1

C

	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.926	0.897	0.916	0.906	0.929	0.908	0.867
COG		1	0.912	0.911	0.909	0.931	0.926	0.901
DEP			1	0.926	0.914	0.929	0.909	0.893
IRB				1	0.973	0.918	0.925	0.869
VAB					1	0.914	0.939	0.877
APT						1	0.943	0.881
POB							1	0.915
PSY								1

D

	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.929	0.904	0.909	0.893	0.923	0.905	0.846
COG		1	0.905	0.916	0.909	0.944	0.913	0.88
DEP			1	0.932	0.91	0.919	0.885	0.867
IRB				1	0.962	0.919	0.915	0.878
VAB					1	0.92	0.919	0.881
APT						1	0.936	0.899
POB							1	0.876
PSY								1

Figure 3.13: Correlation plots for ages at onset comparing between sexes. (A and B) Correlation matrices for unadjusted CCQ symptoms for males (A) and females (B). (C and D) Correlation matrices for adjusted CCQ symptoms (2 year cut-off) for men (C) and women (D). Total male N=4140; total female N=4753. See Appendix 2 for more sxrater cut-offs. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/ obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

3.4 Symptom prevalence in HD

3.4.1 Initial modelling of symptom prevalence

Having calculated ages at onset for motor, cognitive and psychiatric/behavioural domains using the CCQ, we were next interested in investigating how variables such as sex and the duration of the disease affected symptom prevalence in HD. In doing so, we also wanted to further investigate the degree to which CCQ captures symptoms occurring outside of HD disease course, especially for irritability, VAB and depression.

CCQ binary responses (0 = no symptom reported, 1 = symptom experienced at some point in life) were used for all individuals for whom sxrater data was available. Both unadjusted and sxrater adjusted CCQ data were used (adjusted data removed CCQ observations occurring >2 years earlier than sxrater). Initially, we investigated the frequency of all eight symptoms in Registry (Table 3.5). Motor symptoms were by far the most common in Registry, with >98% of all symptomatic participants having a positive response for motor symptoms. The least common symptom was psychosis, with 10.9% of individuals reporting having experienced psychotic symptoms during their HD disease course. Women had ~10% higher prevalence of depression symptoms compared with men, even after removing people with depression onset >2 years earlier than sxrater. A simple chi-square test shows this difference is highly significant (before adjustment *p*=7.67E-21; after sxrater adjustment *p*=2.44E-14). Furthermore, both irritability and VAB were found to have significantly higher prevalence in in men both before (chi-square, *p*=2.88E-06 irritability; *p*=3.35E-10 VAB) and after adjustment (chi-square *p*=5.64E-05 irritability; *p*=3.76E-08 VAB).

There is a small (non-significant) difference in CAG length between males and females in Registry (CAG mean = 44.08 males vs 44.14 females; *p*=0.573, Welch two sample t-test). To account for potential confounders and investigate both the direction and size of effect, binary CCQ symptoms were used to construct multivariate logistic generalised linear regression models. GLMs allow for possible confounders to be included as covariates. We included CAG length, HD onset age (defined by the clinical rater (sxrater), regardless of onset type) and disease duration as covariates in the first set of models. Disease duration was available for many participants (N>4500) and was used an approximation for disease stage (see 2.3.1 for calculation). Only manifest individuals with a known sxrater were included. The outputs of these models are shown in Table 3.6-3.7.

Note that as most manifest HD individuals in Registry reported a positive motor CCQ (~98-99%), the motor models are predicated on a small number of individuals who reported no motor symptoms (~1%). This is demonstrated by the (hugely) inflated standardised coefficients (β coefficients), and consequently limits the usefulness of the motor models. All symptoms are significantly associated with disease duration, and this observation is especially true for cognitive impairment. The association with disease duration becomes stronger after sxrater adjustment in all cases. Presumably this association is capturing that the longer an individual has HD, the more likely they are to experience any given symptom. Sex remains significantly associated with symptom prevalence for depression, irritability and VAB in the GLMs at a similar significance level. Interestingly, depression, irritability and VAB are all associated with earlier onsets (sxrater), suggesting that these psychiatric symptoms have a higher prevalence in younger HD adult individuals. This observation is similar to that seen with symptom onset type in 3.2.1. Finally, CAG length also is associated with symptom prevalence for cognitive impairment, depression and irritability. However, for cognitive symptoms CAG length is positively associated (*i.e.* a larger CAG is associated with a higher log odds ratio) and for depression/irritability it is negatively associated.

	Males				Females				
A	Y	N	Unknown	Freq.	Y	N	Unknown	Freq.	Chi-Square
MTR	2768	30	369	98.93%	2943	28	353	99.06%	6.22E-01
COG	1633	1162	372	58.43%	1742	1220	362	58.81%	7.66E-01
APT	1497	1297	373	53.58%	1542	1414	368	52.17%	2.83E-01
DEP	1631	1165	371	58.33%	2080	885	359	70.15%	**7.67E-21**
POB	1034	1760	373	37.01%	1076	1882	366	36.38%	6.19E-01
IRB	1759	1032	376	63.02%	1687	1274	363	56.97%	**2.88E-06**
VAB	978	1817	372	34.99%	810	2154	360	27.33%	**3.35E-10**
PSY	331	2462	374	11.85%	337	2624	363	11.38%	5.78E-01

	Males						Females						
B	Y	N	Missing	Filtered	Unknown	Freq.	Y	N	Missing	Filtered	Unknown	Freq.	Chi-Square
MTR	2638	30	27	103	369	98.88%	2798	28	29	116	353	99.01%	6.28E-01
COG	1547	1162	40	46	372	57.11%	1655	1220	47	40	362	57.57%	7.29E-01
APT	1394	1297	36	67	373	51.80%	1420	1414	35	87	368	50.11%	2.07E-01
DEP	1368	1165	32	231	371	54.01%	1613	885	33	434	359	64.57%	**2.44E-14**
POB	950	1760	30	54	373	35.06%	988	1882	24	64	366	34.43%	6.21E-01
IRB	1523	1032	38	198	376	59.61%	1504	1274	34	149	363	54.14%	**5.64E-05**
VAB	845	1817	19	114	372	31.74%	721	2154	17	72	360	25.08%	**3.76E-08**
PSY	308	2462	7	16	374	11.12%	317	2624	8	12	363	10.78%	6.80E-01

Table 3.5: Prevalence of HD symptoms in men and women. CCQ data before (A) and after (B) adjustment with sxrater (>2 year cut-off), using binarily coded CCQ data (0 = no symptom, 1 = symptom experienced). Significant *p*-values are emboldened (*p*<6.25E-03). Only individuals with an sxrater are included, and this includes all CAG lengths. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis. Response abbreviations: Y: Yes (symptom experienced); N: No (symptom not experienced); Missing (in (B)): Some data missing (*e.g.* sxrater or CCQ age); Filtered (in (B)): CCQ age occurred >2 years earlier than sxrater; Unknown: No response (blank).

	MTR (N=4624)				COG (N=4615)				APT (N=4609)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.002	0.014	0.337	9.94E-01	-0.038	-0.038	0.061	5.37E-01	0.026	0.026	0.060	6.61E-01
CAG	0.382	19.402	0.058	**4.99E-11**	0.045	0.423	0.010	**1.37E-05**	-0.002	-0.017	0.009	8.44E-01
Duration	0.085	5.472	0.026	**1.23E-03**	0.089	1.046	0.006	**6.53E-47**	0.053	0.610	0.006	**9.37E-21**
Onset	0.167	23.057	0.019	**4.12E-18**	0.006	0.148	0.004	1.08E-01	-0.003	-0.086	0.003	3.20E-01

	DEP (N=4618)				POB (N=4612)				IRB (N=4610)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	-0.589	-0.617	0.063	**1.47E-20**	-0.017	-0.018	0.063	7.86E-01	0.285	0.290	0.061	**2.82E-06**
CAG	-0.079	-0.761	0.010	**1.05E-14**	0.019	0.182	0.010	*4.96E-02*	-0.040	-0.376	0.010	**2.49E-05**
Duration	0.040	0.487	0.006	**5.04E-11**	0.058	0.705	0.006	**7.04E-25**	0.030	0.357	0.006	**8.79E-08**
Onset	-0.026	-0.684	0.004	**2.77E-12**	-0.001	-0.026	0.004	7.85E-01	-0.019	-0.470	0.004	**1.84E-07**

	VAB (N=4615)				PSY (N=4613)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.410	0.445	0.065	**3.36E-10**	0.044	0.069	0.094	6.43E-01
CAG	-0.007	-0.073	0.010	4.63E-01	-0.002	-0.026	0.014	8.98E-01
Duration	0.053	0.666	0.006	**2.64E-20**	0.073	1.328	0.007	**1.79E-22**
Onset	-0.018	-0.477	0.004	**3.95E-06**	-0.007	-0.260	0.005	2.23E-01

Table 3.6: Unadjusted CCQ data GLMs for symptom prevalence. Generalised linear models (GLMs) for unadjusted CCQ symptom onset data are shown. Only individuals CAG 36-99, known sex and known sxrater were included in models. Significant values are emboldened that pass multiple correction (8 tests Bonferroni threshold *p*=6.25E-03); nominally significant values are italicised. *B* = unstandardised coefficient; β = standardised coefficient; *SE* = standard error. Sex is the effects for males vs females. 'Onset' here is a general onset defined by the clinician (sxrater) regardless of onset type. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

	MTR (N=4407)				COG (N=4479)				APT (N=4438)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.010	0.053	0.338	9.77E-01	-0.043	-0.043	0.062	4.89E-01	0.038	0.038	0.061	5.29E-01
CAG	0.383	19.102	0.058	**5.43E-11**	0.043	0.395	0.010	**4.61E-05**	-0.002	-0.019	0.009	8.25E-01
Duration	0.085	5.334	0.026	**1.27E-03**	0.090	1.051	0.006	**9.82E-47**	0.055	0.634	0.006	**1.21E-21**
Onset	0.167	22.327	0.019	**5.95E-18**	0.005	0.112	0.004	2.22E-01	-0.004	-0.103	0.004	2.44E-01

	DEP (N=4042)				POB (N=4491)				IRB (N=4288)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	-0.504	-0.514	0.066	**2.52E-14**	-0.006	-0.007	0.064	9.21E-01	0.255	0.257	0.062	**4.55E-05**
CAG	-0.076	-0.731	0.011	**6.71E-13**	0.023	0.225	0.010	*1.98E-02*	-0.039	-0.365	0.010	**6.20E-05**
Duration	0.053	0.625	0.006	**2.10E-16**	0.061	0.754	0.006	**2.28E-26**	0.035	0.413	0.006	**1.29E-09**
Onset	-0.028	-0.702	0.004	**2.23E-12**	0.000	-0.006	0.004	9.50E-01	-0.019	-0.483	0.004	**1.38E-07**

	VAB (N=4446)				PSY (N=4577)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.390	0.434	0.069	**1.28E-08**	0.036	0.058	0.097	7.14E-01
CAG	-0.004	-0.045	0.010	6.70E-01	0.001	0.014	0.015	9.50E-01
Duration	0.059	0.768	0.006	**2.30E-23**	0.077	1.445	0.008	**7.46E-24**
Onset	-0.018	-0.488	0.004	**1.12E-05**	-0.006	-0.223	0.006	3.24E-01

Table 3.7: Adjusted CCQ data GLMs for symptom prevalence. Generalised linear models (GLMs) for adjusted CCQ symptom onset data are shown (filtering symptoms occurring <2 years earlier than sxrater onset). Only individuals CAG 36-99, known sex and known sxrater were included in models. Significant values are emboldened that pass multiple testing correction (8 tests Bonferroni threshold *p*=6.25E-03); nominally significant values are italicised. *B* = unstandardised coefficient; β = standardised coefficient; *SE* = standard error. Sex is the effects for males vs females. 'Onset' here is a general onset defined by the clinician (sxrater) regardless of onset type. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

3.4.2 Extending the symptom GLMs with additional covariates

We next wanted to extend our generalised linear modelling approach with more covariables to adjust for potential confounders, using the same binary CCQ data as in 3.4.1. We based the covariates used in this section from the Dale study (Dale et al., 2016), which found no difference between depression in men and women in HD disease course when using the Hospital Anxiety and Depression Scale (HADS; see 3.4.3/3.4.4). The covariates we added were alcohol use, tobacco use, years in education, total functional capacity (TFC) score and total motor score (TMS), from the latest visit data available. We did not include medication, as the Dale study did, as these data were not available for most of our sample. The disease duration covariate we used in the extended models is explicitly from the visits that were used to derive TFC. Furthermore, as in the Dale study, we did not include very advanced HD patients with TFC score=0 (TFC score of 0 indicates total reliance on others), and only included individuals with *HTT* CAGs between 39-55. Juvenile HD cases (JHD) were removed (sxrater onset <20 years). This left N=~1500 individuals for regression modelling (Tables 3.8-3.9).

As before, the motor models are predicated on a small number of individuals who did not report motor symptoms. These models have highly inflated βs consequently, and their usage is somewhat underpowered; although TMS does have a small amount of significance with motor symptom prevalence. Onset (sxrater, regardless of onset type) is nominally and positively associated with motor symptoms. TFC score is significantly associated with all other symptoms, especially cognitive impairment, with a lower TFC score being associated with a higher log odds ratio risk. TFC likely captures disease stage better than simply disease duration which is only nominally significant for VAB after sxrater adjustment. Sex-dependent effects remain significant for depression in both unadjusted and adjusted data, however irritability and VAB are only nominally significant adjusting for multiple testing correction. Age of onset (sxrater) is significantly and negatively associated with cognitive impairment, depression, irritability and VAB, and nominally significant for psychosis, indicating younger adults are more likely to experience psychiatric/cognitive symptoms in their HD disease course compared to older groups. This was also noted for several symptoms in the simpler models. Psychosis was found to be significantly and negatively associated with the number of education years. Finally, CAG length was found to be negatively associated with depression and irritability, and nominally so for cognitive symptoms and VAB.

	MTR (N=1494)				COG (N=1494)				APT (N=1494)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.921	5.379	0.723	2.03E-01	-0.189	-0.190	0.117	1.06E-01	-0.016	-0.016	0.109	8.83E-01
CAG	-0.010	-0.363	0.138	9.44E-01	-0.067	-0.427	0.029	*2.17E-02*	-0.010	-0.061	0.027	7.23E-01
Duration	-0.103	-6.511	0.080	1.98E-01	0.013	0.144	0.013	3.07E-01	0.003	0.032	0.012	8.03E-01
Alcohol	-0.030	-2.615	0.039	4.44E-01	0.023	0.340	0.009	*7.92E-03*	0.001	0.021	0.007	8.44E-01
Tobacco	0.097	11.188	0.073	1.87E-01	0.004	0.084	0.006	4.78E-01	0.010	0.200	0.006	7.05E-02
Education	-0.117	-4.899	0.104	2.60E-01	0.011	0.080	0.017	5.05E-01	-0.029	-0.206	0.015	6.18E-02
TFC	0.097	4.191	0.179	5.88E-01	-0.287	-2.139	0.026	**1.57E-27**	-0.153	-1.136	0.024	**1.22E-10**
TMS	0.130	34.313	0.045	**3.62E-03**	-0.005	-0.217	0.004	2.53E-01	-0.006	-0.256	0.004	1.39E-01
Onset	0.098	13.783	0.048	*4.33E-02*	-0.024	-0.580	0.008	**2.92E-03**	-0.008	-0.190	0.007	2.91E-01

	DEP (N=1496)				POB (N=1493)				IRB (N=1496)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	-0.536	-0.553	0.113	**1.89E-06**	-0.080	-0.087	0.118	4.95E-01	0.326	0.330	0.110	**3.06E-03**
CAG	-0.124	-0.812	0.028	**1.21E-05**	-0.051	-0.354	0.029	8.16E-02	-0.080	-0.514	0.027	**3.49E-03**
Duration	0.056	0.624	0.013	**2.47E-05**	0.043	0.508	0.013	**5.99E-04**	0.050	0.552	0.013	**8.39E-05**
Alcohol	-0.012	-0.186	0.007	1.04E-01	0.010	0.165	0.008	1.81E-01	0.014	0.207	0.008	9.45E-02
Tobacco	0.013	0.263	0.006	*3.04E-02*	0.003	0.068	0.006	5.90E-01	0.017	0.351	0.006	**2.81E-03**
Education	-0.017	-0.127	0.016	2.79E-01	-0.004	-0.031	0.017	8.13E-01	-0.011	-0.077	0.015	4.90E-01
TFC	-0.122	-0.933	0.025	**7.60E-07**	-0.126	-1.019	0.026	**7.35E-07**	-0.089	-0.668	0.024	**2.14E-04**
TMS	-0.015	-0.702	0.004	**1.82E-04**	-0.006	-0.305	0.004	1.26E-01	-0.009	-0.411	0.004	*2.14E-02*
Onset	-0.039	-0.984	0.008	**5.66E-07**	-0.016	-0.433	0.008	*4.10E-02*	-0.030	-0.731	0.008	**8.04E-05**

	VAB (N=1496)				PSY (N=1485)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.280	0.311	0.122	*2.14E-02*	0.225	0.414	0.198	2.57E-01
CAG	-0.037	-0.261	0.030	2.15E-01	-0.098	-1.155	0.050	*4.98E-02*
Duration	0.061	0.735	0.013	**2.37E-06**	0.024	0.479	0.020	2.20E-01
Alcohol	0.007	0.110	0.008	3.97E-01	0.009	0.255	0.012	4.47E-01
Tobacco	0.012	0.267	0.006	*4.00E-02*	-0.002	-0.091	0.010	8.07E-01
Education	-0.025	-0.197	0.017	1.54E-01	-0.108	-1.425	0.030	**2.82E-04**
TFC	-0.129	-1.064	0.026	**1.02E-06**	-0.170	-2.323	0.043	**8.59E-05**
TMS	-0.006	-0.303	0.004	1.49E-01	-0.004	-0.312	0.007	5.66E-01
Onset	-0.029	-0.766	0.008	**7.11E-04**	-0.024	-1.085	0.013	6.92E-02

Table 3.8: Unadjusted CCQ data extended GLMs. Generalised linear models (GLMs) showing binary unadjusted CCQ data for individuals with an onset ≥20 (sxrater), >TFC score=0, CAGs 39-55. *B* = unstandardised coefficient; β = standardised coefficient; *SE* = standard error. ‘Onset’ here is a general onset defined by the clinician (sxrater) regardless of onset type. Significant values are emboldened that pass multiple testing correction (8 tests Bonferroni threshold *p*=6.25E-03); nominally significant values are italicised. Note that for psychosis, individuals with a co-morbid diagnosis of schizophrenia, schizotypal disorder or schizoaffective disorder were removed (see 2.3.1). MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

	MTR (N=1416)				COG (N=1450)				APT (N=1434)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.981	5.582	0.726	1.77E-01	-0.181	-0.182	0.119	1.26E-01	-0.002	-0.002	0.111	9.83E-01
CAG	0.008	0.290	0.138	9.54E-01	-0.066	-0.426	0.029	*2.36E-02*	-0.018	-0.115	0.028	5.10E-01
Duration	-0.001	-0.083	0.079	9.86E-01	-0.008	-0.087	0.014	5.60E-01	-0.004	-0.046	0.013	7.41E-01
Alcohol	-0.030	-2.547	0.039	4.48E-01	0.023	0.340	0.009	*8.59E-03*	0.002	0.030	0.008	7.84E-01
Tobacco	0.103	11.757	0.075	1.70E-01	0.004	0.083	0.006	4.88E-01	0.009	0.179	0.006	1.12E-01
Education	-0.115	-4.662	0.104	2.72E-01	0.014	0.102	0.017	3.97E-01	-0.026	-0.185	0.016	1.01E-01
TFC	0.124	5.207	0.186	5.07E-01	-0.285	-2.115	0.027	**1.40E-26**	-0.155	-1.150	0.024	**1.72E-10**
TMS	0.131	33.280	0.045	**3.75E-03**	-0.005	-0.220	0.004	2.49E-01	-0.005	-0.246	0.004	1.65E-01
Onset	0.109	13.787	0.050	*2.89E-02*	-0.024	-0.546	0.008	**2.88E-03**	-0.012	-0.260	0.008	1.28E-01

	DEP (N=1302)				POB (N=1461)				IRB (N=1384)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	-0.435	-0.439	0.119	**2.42E-04**	-0.066	-0.073	0.121	5.86E-01	0.267	0.268	0.114	*1.86E-02*
CAG	-0.137	-0.870	0.030	**5.13E-06**	-0.055	-0.389	0.030	7.05E-02	-0.091	-0.576	0.028	**1.32E-03**
Duration	0.023	0.258	0.014	9.98E-02	0.024	0.286	0.014	8.05E-02	0.020	0.218	0.014	1.37E-01
Alcohol	-0.010	-0.156	0.008	1.90E-01	0.012	0.202	0.008	1.11E-01	0.013	0.184	0.009	1.28E-01
Tobacco	0.010	0.195	0.006	1.11E-01	0.004	0.083	0.006	5.32E-01	0.016	0.305	0.006	*9.27E-03*
Education	-0.020	-0.145	0.016	2.31E-01	0.000	0.000	0.017	9.99E-01	-0.009	-0.069	0.016	5.50E-01
TFC	-0.127	-0.957	0.026	**8.60E-07**	-0.131	-1.071	0.026	**6.35E-07**	-0.085	-0.636	0.025	**5.97E-04**
TMS	-0.013	-0.584	0.004	**2.50E-03**	-0.006	-0.295	0.004	1.54E-01	-0.007	-0.330	0.004	7.31E-02
Onset	-0.042	-0.957	0.008	**3.51E-07**	-0.016	-0.406	0.008	5.18E-02	-0.033	-0.751	0.008	**2.06E-05**

	VAB (N=1441)				PSY (N=1472)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.258	0.296	0.129	*4.51E-02*	0.242	0.467	0.210	2.49E-01
CAG	-0.070	-0.508	0.032	*2.67E-02*	-0.104	-1.284	0.053	*4.86E-02*
Duration	0.029	0.355	0.014	*4.24E-02*	0.000	0.005	0.021	9.91E-01
Alcohol	0.005	0.079	0.009	5.86E-01	0.015	0.445	0.012	2.11E-01
Tobacco	0.009	0.205	0.006	1.47E-01	-0.018	-0.673	0.012	1.51E-01
Education	-0.014	-0.113	0.018	4.52E-01	-0.108	-1.488	0.031	**5.87E-04**
TFC	-0.133	-1.140	0.028	**1.96E-06**	-0.175	-2.512	0.046	**1.38E-04**
TMS	-0.003	-0.148	0.004	5.20E-01	-0.002	-0.210	0.007	7.26E-01
Onset	-0.036	-0.939	0.009	**4.50E-05**	-0.028	-1.238	0.014	*4.44E-02*

Table 3.9: Adjusted CCQ data extended GLMs. Generalised linear models (GLMs) showing binary adjusted CCQ data (CCQ data removing >2 years earlier than sxrater were filtered) for individuals with an onset ≥20 (sxrater), >TFC score=0, CAGs 39-55. *B* = unstandardised coefficient; β = standardised coefficient; *SE* = standard error. 'Onset' here is a general onset defined by the clinician (sxrater) regardless of onset type. Significant values are emboldened that pass multiple testing correction (8 tests Bonferroni threshold *p*=6.25E-03); nominally significant values are italicised. Note that for psychosis, individuals with a co-morbid diagnosis of schizophrenia, schizotypal disorder or schizoaffective disorder were removed (see 2.3.1). MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

3.4.3 Using the HADS-SIS scales to investigate phenotype

We were interested whether the independent hospital anxiety and depression score (HADS) with the additional Snaith's irritability scale (SIS; together HADS-SIS) would produce similar results to the binary CCQ data. A similar but smaller study using the HADS was carried out (Dale et al., 2016) using individuals from Registry, many of whom are likely also in this study. The HADS asks 14 questions (methods Table 2.3), 7 of which relate to depression and 7 to anxiety (ANX), with scores ranging from 0-3 for each answer. The SIS similarly asks the participant a series of 8 questions (methods Table 2.4) related to irritability with scores 0-3 for each response. Total depression scores (TDS), anxiety scores (TAS) and irritability scores (TIS) are then derived from the responses, with higher scores being more associated with more of the related phenotype (Snaith et al., 1978; Zigmond and Snaith, 1983).

Using the HADS-SIS scores to construct a generalised linear model (Table 3.10) shows there is nominal significant of TDS with sex, however this does not survive multiple testing correction. The direction of effect is opposite to what was expected (males have slightly but non-significant higher scores on average than women, mean 6.96 vs 6.64). Tobacco usage is significantly associated with TDS, and education years is nominally significant. TDS is most strongly associated with TFC, with a higher TFC score inversely associated with TDS, and this agrees with the association seen between TFC and depression CCQ earlier (Table 3.9). TAS only shows nominal significance with CAG, education years and age. The TIS on the other hand shows a significant association with CAG length, in the same direction as TAS, where longer CAGs are associated with less irritability after correcting for other variables. This is similar to what was seen in several of the psychiatric symptoms when using the binary CCQ (Table 3.9). In addition, TIS is also associated with the onset (sxrater), with younger individuals having more irritability, again matching what was seen in several of the psychiatric binary CCQ models before. There is a small but non-significant difference in TIS between males and females (6.51 vs 6.09 male/female, *p*=0.23, Welch two sample t-test). Neither TIS nor TAS were significantly associated with TFC.

A second modelling approach transformed TDS, TAS and TIS data into binary variables with scores of 0-7 being "normal" (0) and >8 being a "case" (1) to be more methodologically comparable to the binary CCQ GLMs. We chose >8 as TDS/TAS of 8-10 were considered possible cases by Zigmond and Snaith, and scores of >11 as definite cases (Zigmond and Snaith, 1983). This method mostly had the same findings as numerical TDS/TAS/TIS data but with higher *p* values. The granularity offered by the full TDS/TAS/TIS scales, therefore, improved the predictive power of the GLMs.

	TDS (N=766)				TAS (N=762)				TIS (N=767)			
A	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.656	0.072	0.324	*4.36E-02*	0.179	0.020	0.329	5.87E-01	0.495	0.051	0.352	1.60E-01
CAG	-0.105	-0.072	0.082	2.04E-01	-0.178	-0.124	0.083	*3.28E-02*	-0.315	-0.201	0.089	**4.30E-04**
Duration	-0.064	-0.074	0.039	9.62E-02	-0.095	-0.111	0.039	1.44E-02	-0.085	-0.092	0.041	3.99E-02
Alcohol	-0.011	-0.016	0.025	6.51E-01	-0.008	-0.012	0.025	7.39E-01	-0.017	-0.023	0.027	5.28E-01
Tobacco	0.048	0.097	0.018	**6.96E-03**	0.032	0.066	0.018	7.33E-02	0.032	0.062	0.019	9.02E-02
Education	-0.103	-0.080	0.046	*2.63E-02*	-0.121	-0.096	0.047	*9.91E-03*	-0.099	-0.073	0.050	*4.85E-02*
TFC	-0.386	-0.288	0.071	**7.91E-08**	-0.118	-0.089	0.072	1.04E-01	-0.076	-0.053	0.077	3.26E-01
TMS	-0.010	-0.045	0.012	4.02E-01	-0.010	-0.046	0.012	4.10E-01	-0.017	-0.071	0.013	1.95E-01
Onset	-0.006	-0.015	0.023	7.79E-01	-0.048	-0.119	0.023	*3.55E-02*	-0.101	-0.230	0.024	**4.09E-05**

	TDS (N=766)				TAS (N=762)				TIS (N=767)			
B	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
Sex	0.051	0.052	0.036	1.49E-01	0.024	0.025	0.035	5.04E-01	0.031	0.033	0.035	3.75E-01
CAG	-0.006	-0.035	0.009	5.40E-01	-0.025	-0.163	0.009	*4.93E-03*	-0.023	-0.151	0.009	**8.52E-03**
Duration	-0.004	-0.038	0.004	3.94E-01	-0.009	-0.099	0.004	2.98E-02	-0.006	-0.064	0.004	1.51E-01
Alcohol	-0.002	-0.025	0.003	4.89E-01	-0.003	-0.035	0.003	3.46E-01	-0.002	-0.027	0.003	4.62E-01
Tobacco	0.004	0.068	0.002	6.15E-02	0.003	0.061	0.002	9.94E-02	0.002	0.046	0.002	2.15E-01
Education	-0.005	-0.033	0.005	3.73E-01	-0.007	-0.053	0.005	*1.59E-01*	-0.006	-0.048	0.005	*1.98E-01*
TFC	-0.039	-0.267	0.008	**8.26E-07**	-0.012	-0.086	0.008	1.16E-01	-0.013	-0.091	0.008	9.69E-02
TMS	-0.001	-0.051	0.001	3.44E-01	-0.002	-0.076	0.001	1.72E-01	-0.003	-0.134	0.001	1.54E-02
Onset	-0.001	-0.023	0.002	6.73E-01	-0.006	-0.148	0.002	*9.15E-03*	-0.007	-0.169	0.002	**2.66E-03**

Table 3.10: GLMs for the HADS-SIS scale. Generalised linear models (GLMs) are shown. Both sets of tables use the same individuals, and only considered individuals with CAGs 39-55, TFC scores >0, sxrater>=20. *B* = unstandardised coefficient; β = standardised coefficient; *SE* = standard error. (A) is the raw TDS/TAS/TIS scales, table (B) transforms the binary TDS/TAS/TIS table where scores of 0-7 were considered normal (0), and scores >7 were considered cases (1). Significant values are emboldened that pass multiple testing correction (3 tests Bonferroni *p*=1.67E-02), nominal values are italicised. TDS: Total depression score; TAS: Total anxiety score; TIS: Total irritability score.

3.4.4 Correlations for all covariates, HADS-SIS and binary CCQ

Correlation matrices were constructed for all adjusted binary CCQ symptoms, HADS-SIS scores and covariates used for modelling purposes using a pairwise deletion method (Fig. 3.14; full numerical table in Appendix 3).

Firstly, as previously established, motor symptoms were reported in most individuals in Registry who had clinical onset as defined by the rater (sxrater), thus it has quite weak associations with other symptoms (although is weakly correlated with TMS/TFC and onset). All three of the HADS-SIS scales are highly correlated with one another, especially the TAS and TIS scales (0.681). Unsurprisingly TFC score is negatively associated with most other variables except CAG length and disease duration. Both disease duration and TMS have similar inverse patterns of correlation compared to TFC, although these are weaker than TFC itself. Hence, disease duration may be a useful metric in cases where TFC is not available. Psychosis binary CCQ is poorly correlated with other symptoms, analogous to the observation that psychotic symptom onset is least correlated with other symptoms. TIS is most strongly associated with both irritability and VAB CCQ, which themselves are correlated. TAS correlates slightly more strongly with depression CCQ (0.220) than does TDS (0.213); TDS has the highest correlation with reporting of apathy CCQ (0.252). Of the demographic information collected, education years has the strongest correlation with apathy, depression and POB CCQ, and these independent correlations are significant. Importantly, these correlations, although useful, do not account for potential confounders.

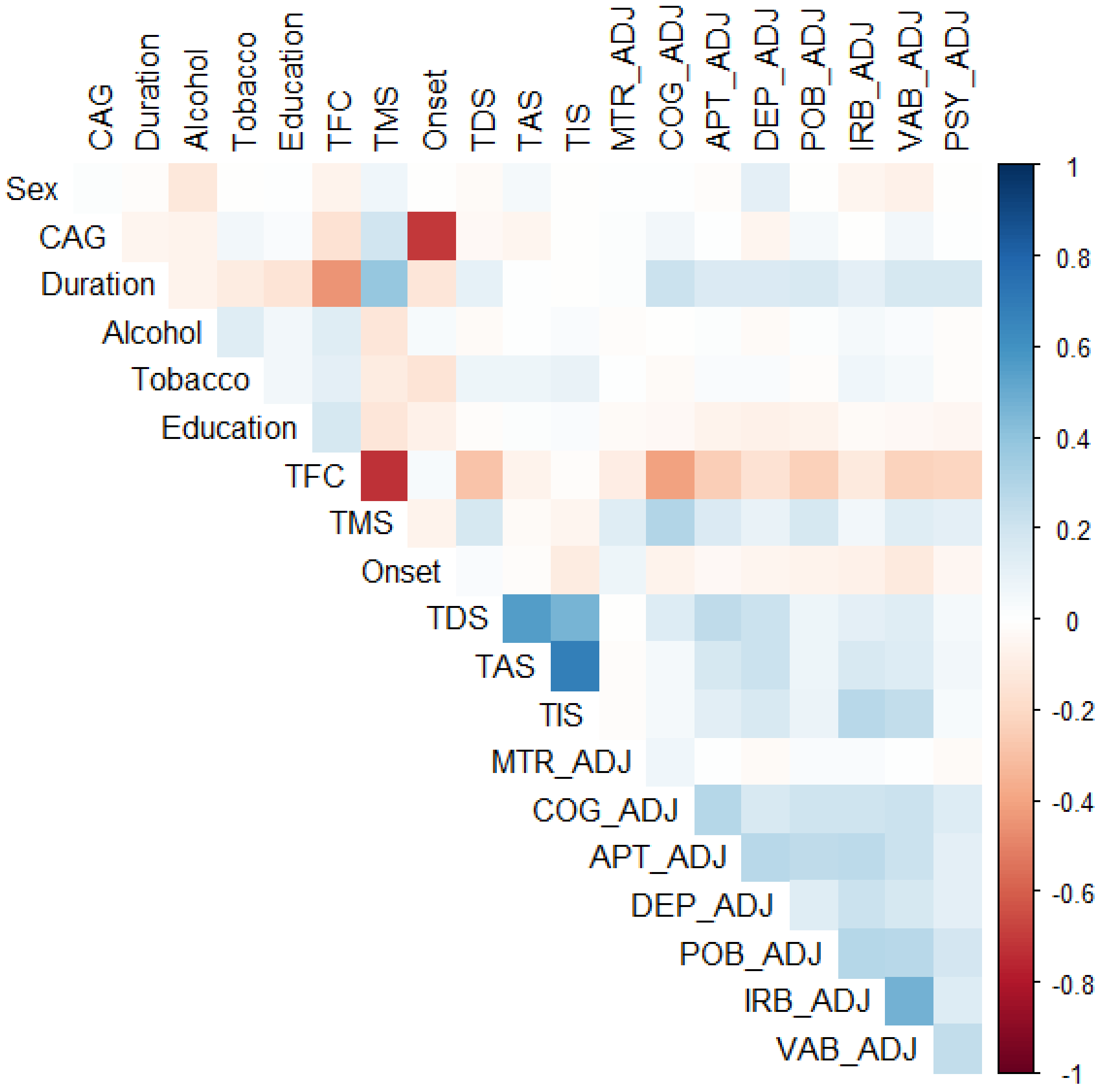


Figure 3.14: Correlation plot for binary-derived CCQ data, HADS-SIS and covariates. Constructed using N=6303 individuals for whom sxrater was known. Full numerical form of this figure (with r, N and p) is in Appendix 3. Symptoms shown here are adjusted (ADJ) binary CCQ responses, removing symptoms occurring >2 years earlier than sxrater. Duration: disease duration; Alcohol: alcohol use in units per week; Tobacco: cigarettes per day; TFC: total functional capacity score; TMS: total motor score; onset: onset defined by sxrater; TDS: total depression score; TAS: total anxiety score; TIS: total irritability score; MTR_ADJ: Motor CCQ; COG_ADJ: cognitive CCQ; APT_ADJ: apathy CCQ; DEP_ADJ: depression CCQ; POB_ADJ: perseverative/obsessive behaviour CCQ; IRB_ADJ: irritability CCQ; VAB_ADJ: violent/aggressive behaviour CCQ; PSY_ADJ: psychosis CCQ.

3.5 Quantifying variation in symptom onset explained by CAG length

CAG length is the primary modifier for age at onset in HD, accounting for ~60% of age at motor onset modification (Wexler et al., 2004; Lee et al., 2012c). However, most estimates for the influence of CAG length on HD onset derive entirely from motor onset data. Hence, we were interested in quantifying the variation explained by CAG length across different symptoms experienced by individuals in Registry. To do so, age at onset data were logarithmically transformed and plotted against CAG length using a linear model. Only individuals with known sex, ages at symptom onset >3 years and CAGs 36-90 were considered. Linear models were assessed by plotting leverages against standardised residuals to find potential influential cases; one outlying individual with a high standardised residual and leverage was identified. Investigating this individual revealed an aberrantly early sxrater, >30 years earlier than all other symptoms called by the family, patient and CCQ. This data point was removed to improve all model fits. Representative regression and leverage plots are shown for apathy in Fig. 3.15, and are very similar for other symptoms.

All symptoms were highly significantly associated with CAG length (p<2E-16), however the R^2 values, *i.e.* the measure of how much variability in age at onset is explained by CAG length, varies for each symptom. Starting with R^2 derived using unadjusted CCQ data (Table 3.12A), motor symptoms have the highest R^2 with CAG accounting for about 61% of age at onset variation (similar to other studies, discussed in 3.8.4). This is followed by cognition and apathy which have R^2 of 0.551 and 0.460, respectively. POB, irritability and VAB all have similar R^2 (0.377, 0.367 and 0.350, respectively) whilst psychotic and depressive symptoms have the lowest R^2 (0.315 and 0.236, respectively). In all cases, using adjusted CCQ data increased the R^2 between CAG size and symptom onset. Motor symptom onset remains the most associated with CAG length (R^2 = 0.659). Cognitive R^2 is nearly as high (0.644). VAB, irritability, POB and apathy have similar R^2 (between 0.581 and 0.618). Consistent with their relationship before adjustment, both depression and psychosis remain the least associated with CAG length with R^2 <0.50. 5 and 10 year CCQ/sxrater data cut-offs do not greatly change the R^2, although R^2 is (as expected) slightly lower using more tolerant cut-offs (Table 3.13). As indicated in Table 3.13A, the best-estimate measures for motor and cognitive symptoms decrease the R^2 by ~0.035, likely reflecting the reduced stringency of these estimates these estimates as these measures attempted to calculate age at onset for as many people as possible. Running a generalised linear model regressing ages at onset on both expanded CAG length and wild-type CAG length shows no significant role for the wild-type *HTT* CAG length for any of the onset types derived from CCQ data (Table 3.11).

	MTR (N=4555)				COG (N=2660)				APT (N=2258)			
A	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
WT CAG	0.025	0.007	0.034	4.76E-01	-0.019	-0.005	0.051	7.10E-01	0.052	0.014	0.056	3.52E-01
Exp CAG	-1.889	-0.728	0.026	**<2E-16**	-1.899	-0.710	0.037	**<2E-16**	-2.023	-0.700	0.043	**<2E-16**

	DEP (N=2461)				POB (N=1535)				IRB (N=2460)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*	*B*	β	*SE*	*p*
WT CAG	0.059	0.018	0.051	2.46E-01	-0.019	-0.005	0.066	7.79E-01	0.002	0.001	0.054	9.73E-01
Exp CAG	-2.009	-0.659	0.046	**<2E-16**	-1.741	-0.688	0.047	**<2E-16**	-1.898	-0.678	0.041	**<2E-16**

	VAB (N=1221)				PSY (N=500)			
	B	β	*SE*	*p*	*B*	β	*SE*	*p*
WT CAG	-0.036	-0.009	0.079	6.50E-01	0.064	0.018	0.129	6.19E-01
Exp CAG	-1.841	-0.692	0.055	**<2E-16**	-1.723	-0.612	0.100	**<2E-16**

Table 3.11: GLMs examining expanded and wild-type CAGs on age at symptom onset. Generalised linear models (GLMs) for age at onset regressing on wild-type CAG (WT CAG) and expanded CAG (Exp CAG) are shown. Significant values are emboldened. Symptoms are adjusted for the rater's estimate of onset, 2 year cut-off. *B* = unstandardised coefficient; β = standardised coefficient; *SE* = standard error. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

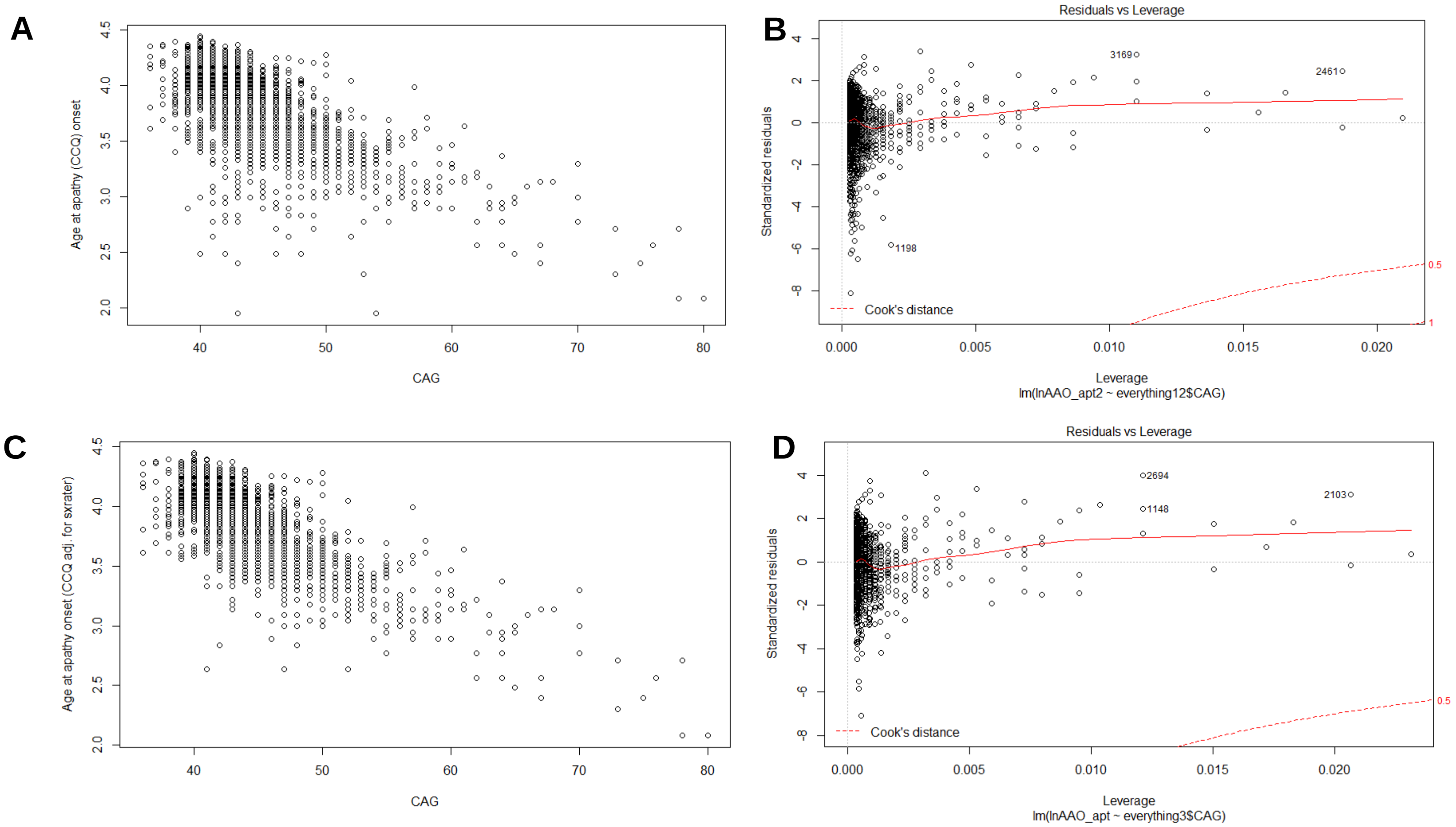


Figure 3.15: Regression for apathy against CAG length. (A-B) Regression analysis for raw CCQ data for apathy with associated leverage plot after standard QC had been applied. (C-D) Similar regression analyses following adjustment of CCQ with sxrater data, sxrater cut-off 2yr. In (B) and (D), the x axes are leverage (how far away data is from other observations) and the y axes standardised residuals. Cook's distance (indicated by the red dashed line) is used as a cut-off for identifying greatly influential data points – in this case there are no obvious outliers. Note that the age at onset has been logarithmically transformed (LN(symptom onset age)).

	Male			Female			Both		
A	r	R^2	N	r	R^2	N	r	R^2	N
MTR	-0.779	0.606	2911	-0.780	0.609	3102	-0.779	0.607	6013
COG	-0.718	0.515	1689	-0.767	0.588	1819	-0.742	0.551	3508
APT	-0.677	0.458	1556	-0.680	0.462	1640	-0.678	0.460	3196
DEP	-0.500	0.250	1780	-0.475	0.226	2381	-0.486	0.236	4161
POB	-0.632	0.399	1059	-0.595	0.354	1119	-0.614	0.377	2178
IRB	-0.593	0.351	1877	-0.618	0.381	1855	-0.606	0.367	3732
VAB	-0.601	0.361	1043	-0.580	0.337	869	-0.592	0.350	1912
PSY	-0.558	0.312	351	-0.567	0.322	353	-0.562	0.315	704

	Male			Female			Both		
B	r	R^2	N	r	R^2	N	r	R^2	N
MTR	-0.821	0.674	2579	-0.803	0.645	2724	-0.812	0.659	5303
COG	-0.794	0.630	1520	-0.812	0.659	1641	-0.803	0.644	3161
APT	-0.761	0.579	1351	-0.773	0.597	1372	-0.768	0.589	2723
DEP	-0.687	0.472	1327	-0.712	0.506	1565	-0.703	0.494	2892
POB	-0.773	0.598	921	-0.783	0.614	952	-0.778	0.606	1873
IRB	-0.751	0.563	1470	-0.773	0.597	1450	-0.762	0.581	2920
VAB	-0.792	0.627	811	-0.779	0.607	686	-0.786	0.618	1497
PSY	-0.664	0.441	295	-0.680	0.462	306	-0.671	0.450	601

Table 3.12: Variance explained by CAG length for symptoms across sexes. (A) shows R^2 for unadjusted CCQ; (B) shows sxrater adjusted CCQ data R^2. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

	Male			Female			Both		
A	r	R^2	N	r	R^2	N	r	R^2	N
C_MTR	-0.793	0.628	3161	-0.784	0.615	3350	-0.788	0.621	6511
C_COG	-0.771	0.594	1746	-0.790	0.624	1858	-0.780	0.609	3604
PSYCH	-0.740	0.540	2234	-0.735	0.555	2390	-0.745	0.548	4624
RTR	-0.784	0.614	3079	-0.774	0.599	3221	-0.779	0.606	6300

	Male			Female			Both		
B	r	R^2	N	r	R^2	N	r	R^2	N
MTR	-0.820	0.672	2638	-0.800	0.641	2793	-0.810	0.655	5431
COG	-0.796	0.634	1547	-0.809	0.655	1664	-0.803	0.644	3211
APT	-0.758	0.575	1386	-0.767	0.588	1407	-0.763	0.582	2793
DEP	-0.689	0.475	1405	-0.709	0.503	1675	-0.702	0.493	3080
POB	-0.782	0.611	931	-0.774	0.600	970	-0.778	0.606	1901
IRB	-0.741	0.549	1542	-0.771	0.594	1497	-0.756	0.572	3039
VAB	-0.783	0.613	849	-0.759	0.576	715	-0.772	0.597	1564
PSY	-0.676	0.458	299	-0.679	0.462	309	-0.677	0.459	608

	Male			Female			Both		
C	r	R^2	N	r	R^2	N	r	R^2	N
MTR	-0.818	0.669	2661	-0.800	0.639	2819	-0.808	0.653	5480
COG	-0.798	0.637	1553	-0.807	0.650	1675	-0.802	0.643	3228
APT	-0.752	0.565	1400	-0.761	0.579	1435	-0.757	0.573	2835
DEP	-0.675	0.455	1470	-0.700	0.490	1808	-0.691	0.477	3278
POB	-0.761	0.579	944	-0.758	0.574	988	-0.760	0.577	1932
IRB	-0.729	0.531	1590	-0.764	0.584	1542	-0.746	0.557	3132
VAB	-0.772	0.596	875	-0.753	0.566	734	-0.763	0.583	1609
PSY	-0.644	0.415	305	-0.674	0.455	312	-0.657	0.432	617

Table 3.13: CAG length on less stringent CCQ-derived data. (A) R^2 data derived using CCQ-sxrater best-estimate data (C_MTR = best-estimate motor onset, C_COG = best-estimate cognitive onset and PSYCH = age at first psychiatric symptom) and sxrater-derived onset (RTR); (B) Adjusted CCQ data using a >5 year cut-off for sxrater; (C) Adjusted CCQ data using a >10 year cut-off for sxrater. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

3.6 Estimating anticipation in Registry

In HD, onset tends to become earlier in each generation due to intergenerational CAG repeat expansions – this is known as clinical anticipation. Anticipation tends to be higher when HD is passed paternally as spermatogenesis is more prone to repeat expansion (see 1.3). Retrospective data for parental HD onset were available in Registry which we used to estimate anticipation. A summary of these data is available in Table 3.14 – only Registry individuals for whom the sxrater was known were included in this analysis. Parental onset status was known in 94.8% of cases. Approximately 7.8% of HD cases were *de novo*, and 6 individuals had parents who both had HD.

In addition to prevalence data, parental ages at onset HD were available for 3075 individuals (Fig. 3.16). When HD was inherited maternally, age at onset in the proband was on average 2.88 years earlier (SD = 9.85). As expected, when HD was inherited paternally anticipation was much higher in the proband with an average of 7.21 years earlier with a higher standard deviation (SD = 10.75). The difference between maternal and paternal anticipation is significant (*p*=2.05E-30, Welch two sample t-test). Interestingly at the two extremes of the distribution in Fig. 3.16C, both paternal and maternal anticipation is very similar. It is possible some of this is due to retrospective error at the two extremes, although in the <20 year group this may also represent floor effects. HD onset in mothers (mean=45.1 years) was significantly earlier than HD onset in fathers (mean=47.0 years) (*p*=9.64E-06, Welch two sample t-test). Of the JHD cases (HD sxrater onset <20), 64.1% originated paternally.

Parent HD	Freq.	Percentage
Father	2802	45.55%
Mother	2862	46.52%
Both	6	0.10%
Neither	482	7.83%
Unknown	339	N/A

Table 3.14: Parental HD onset summary information. Indicated are the frequencies (Freq.) of HD in parents of the individuals in Registry. Only individuals with an sxrater were used for calculations. Both refers to both parents having an HD onset; neither means neither parent was reported to have HD; unknown refers to data not filled in.

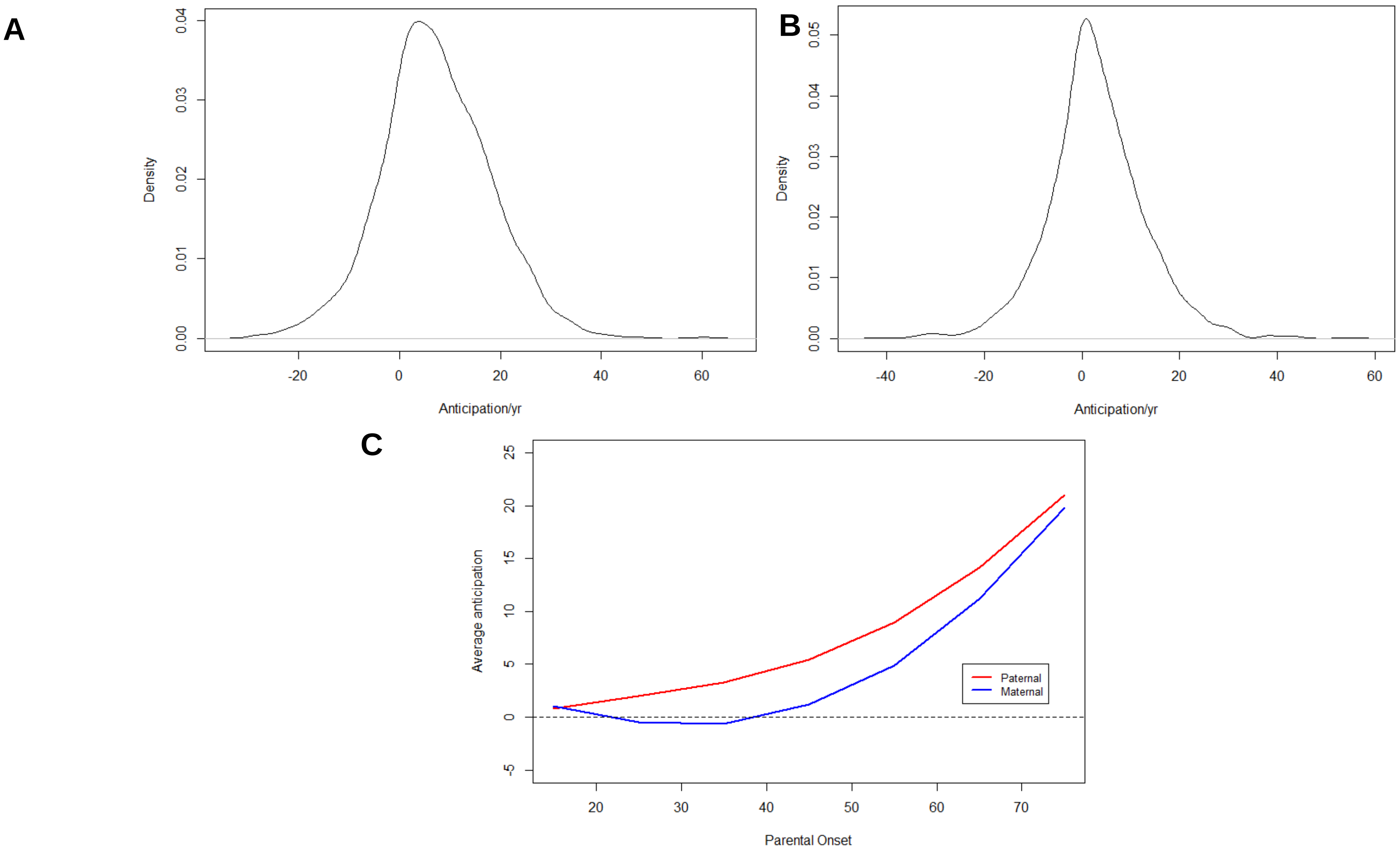


Figure 3.16: Anticipation in Registry. Paternal (A; N=1443) and maternal (B; N=1632) patterns of anticipation. (C; N=3075) demonstrates maternal (blue) and paternal (red) anticipation across age groups (data is averaged in 7 parental onset groups; <20, 20-30 , 30-40, 40-50, 50-60, 60-70 and >70 years).

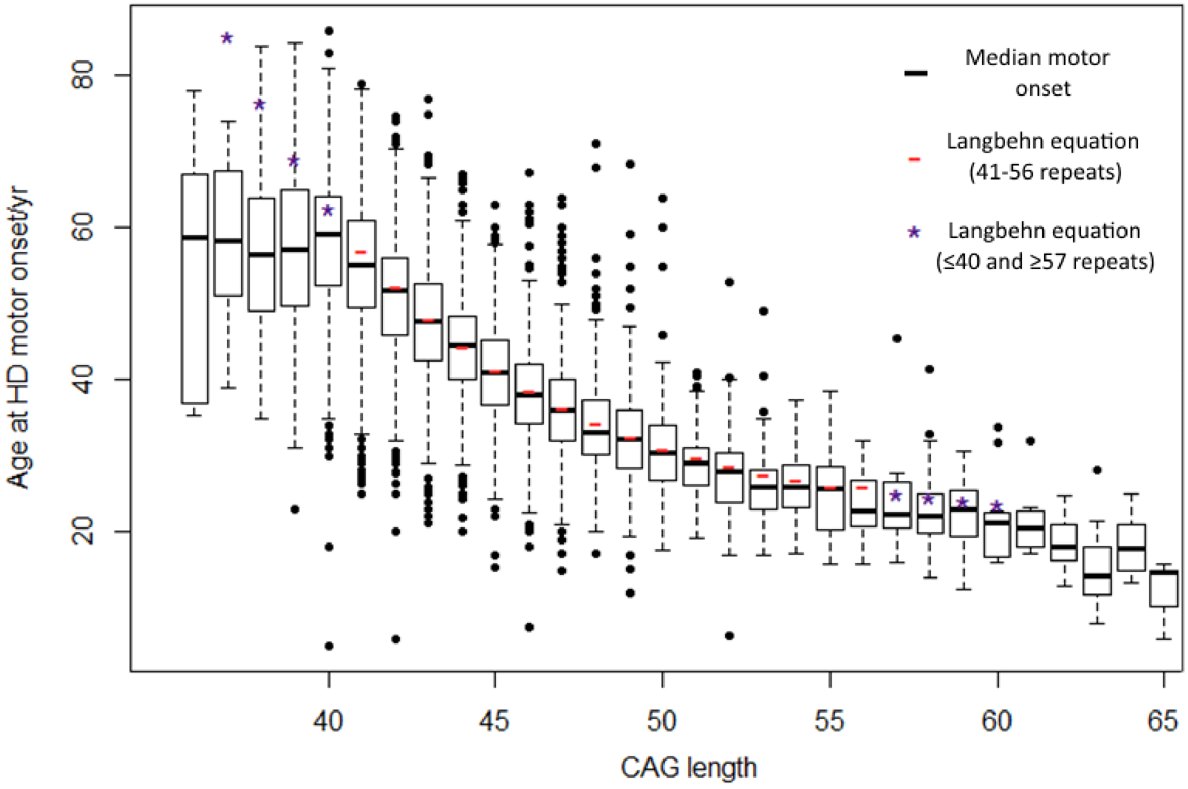
3.7 Selecting an extreme onset population of Huntington’s disease patients

3.7.1 Selecting early and late onset HD patients using AMO

In order to carry out whole-exome sequencing in a subset of the Registry participants (Chapter 4), it was first necessary to select an appropriate population. As we were financially limited in whom could be sequenced (N=500), it made theoretical sense to select patients from the extremes of the onset distribution to try to enrich for coding variants of potentially large effect on onset. We elected to use a similar methodology as the GeM-HD consortium (GeM-HD Consortium, 2015) wherein (1) patients possessing 40-55 CAG repeats were considered for selection, as outside this range the relationship between observed and expected age at onset is less clear and may be influenced by floor effects, and (2), a residual age at motor onset was calculated taking the best-estimate age at motor onset in 3.2.2 and subtracting the expected AMO for that CAG repeat length. The expected AMO at each CAG length was calculated using equation 2.2 (Langbehn et al., 2004).

The Langbehn et al. study used a patient cohort with 41-56 CAGs, and defined age at onset as permanent HD neurological symptoms. As shown in Fig. 3.17 and its accompanying table, the Langbehn model tracks our data quite well for most repeat sizes, and is <1 year different to the mean age at onset observed in our sample for CAGs 43-51. However, at the extremes the model begins to deviate from our observed values. For instance, at 41 CAGs there was a 2.3 year difference between our mean motor onset and the model’s predicted onset. This is especially true when extrapolating outside of the model’s intended range; noticeably for the purposes of this study CAG 40 has a difference of 4.5 years. For other repeat sizes (41-55) the Langbehn model was used for estimation of expected age at onset, and at CAG=40, the median best-estimate age at motor onset at this CAG length (59 years) was used to calculate an AMO residual.

For selection purposes, it is notable that due to the decreasing standard deviation of larger CAG lengths, selection purely based on residual biases the selection towards smaller repeat lengths. A percentile-based system could have been introduced to reduce this bias but was not for several reasons. (1) larger repeats have less variability in onset, and any errors in calculating onset will be more pronounced at larger CAG sizes. Selection of predominantly smaller repeats minimises this effect. (2) smaller repeat sizes are more common and therefore more representative of the HD patient population. (3) modifiers have the largest effect at smaller repeat lengths, possibly due to more variable CAG-related toxicity. Hence, we proceeded using the raw residual age at motor onset for selection purposes.



CAG	Mean AMO	Langbehn	Difference	CAG	Mean AMO	Langbehn	Difference
36	55.36	95.24	+39.88	51	29.28	29.79	+0.51
37	57.39	85.23	+27.84	52	27.24	28.67	+1.43
38	57.34	76.58	+19.24	53	26.42	27.70	+1.28
39	57.28	69.10	+11.82	54	26.17	26.86	+0.69
40	58.14	62.64	+4.50	55	25.10	26.14	+1.04
41	54.79	57.06	+2.27	56	23.56	25.51	+1.95
42	51.13	52.23	+1.10	57	23.63	24.97	+1.34
43	47.27	48.06	+0.79	58	23.63	24.51	+0.88
44	44.36	44.46	+0.10	59	22.62	24.11	+1.49
45	40.98	41.35	+0.37	60	21.56	23.76	+2.20
46	38.41	38.66	+0.25	61	21.63	23.46	+1.83
47	36.35	36.33	-0.02	62	18.31	23.20	+4.89
48	34.25	34.32	+0.07	63	15.76	22.97	+7.21
49	32.83	32.59	-0.24	64	17.97	22.78	+4.81
50	30.99	31.08	+0.09	65	11.56	22.61	+11.05

Figure 3.17: A comparison of the Langbehn model in Registry cohort (AMO). The figure shows boxplots for age at motor onset (AMO) in Registry across 36-65 CAGs. Dots are individuals outside the expected range for a given CAG (>1.5*IQR). Red lines represent expected ages of onset calculated using the Langbehn model (Langbehn et al., 2004). Purple asterisks represent expected ages using the Langbehn model extrapolating outside its 41-56 CAG range. The table shows the mean AMOs for Registry compared to the expected onsets given the Langbehn model and the difference between the two.

Upon initial selection of sequencing candidates, it was identified that the early cohort were enriched for individuals with no sxrater or known onset type (sxraterm) (~5x more than expected), although the absolute numbers of these individuals were still reasonably small (N~20). This was not observed in the late cohort which seemed equally distributed between individuals and onset types. As we were missing data for individuals without sxrater, the CCQ alone was used to calculate AMO. This could have been capturing non-HD symptoms or may have been otherwise erroneous given no sxrater onset data was available for comparative purposes. As it was important to select individuals for sequencing who had high quality data, we filtered out individuals who were missing sxraterm data (no onset type), excepting one individual missing sxraterm who was sequenced. >95% of the individuals chosen for sequencing had an sxrater. Furthermore, each individual had to be manually assessed by two raters, including one clinician, for inclusion in the sequencing cohort – those with highly discrepant CCQ/sxrater data, large amounts of missing data or discordant clinical notes were excluded. As alluded to, individuals selected were biased towards smaller disease-causing repeat sizes (Table 3.15). Late onset individuals had slightly larger CAG repeat sizes which was significant (means=42.6 vs 43.6, Welch two sample t-test *p*=2.68E-6).

Early			Late		
CAG	Freq.	Expected	CAG	Freq.	Expected
40	30	19.91	40	29	19.91
41	64	31.35	41	30	31.35
42	51	40.54	42	40	40.54
43	40	37.19	43	34	37.19
44	26	30.12	44	36	30.12
45	13	23.49	45	23	23.49
46	11	18.56	46	18	18.56
47	5	14.36	47	16	14.36
48	3	9.25	48	11	9.25
49	4	7.12	49	7	7.12
50	2	5.83	50	4	5.83
51	0	3.52	51	0	3.52
52	1	2.86	52	1	2.86
53	0	2.29	53	1	2.29
54	0	1.74	54	0	1.74
55	0	1.86	55	0	1.86

Table 3.15: Selected individual CAGs for exome sequencing. The CAG lengths for individuals selected using an extreme AMO selection method for the early and late arms. Expected refers to the numbers of each CAG length that would be expected if 250 individuals were randomly selected from our Registry data.

Fig. 3.18 shows individuals selected for sequencing based on their AMO residual. Individuals were selected in two tranches (marked as red and blue) of roughly equal size (tranche 1 = 287, tranche 2 = 213) as originally fewer people were planned to be sequenced, with the first tranche including the most extreme individuals. The residual AMO for the first tranche ranges between +36.7 to +12.5 years for the late onset arm and -30.6 to -14.0 years. The second tranche contains several individuals with higher AMO residuals for whom data were not available during the initial selection; nonetheless the second tranche primarily includes less extreme individuals, ranging from +35.8 to +12.5 years for the late onset and -31.2 to -12.4 years for the early onset arm. The early onset arm contained 112 males/138 females (N=250), and the late arm 114 males/135 females/1 unknown (N=250). A small number (N=15) of juvenile HD (JHD) cases (AMO<20 years) were included in the early onset arm.

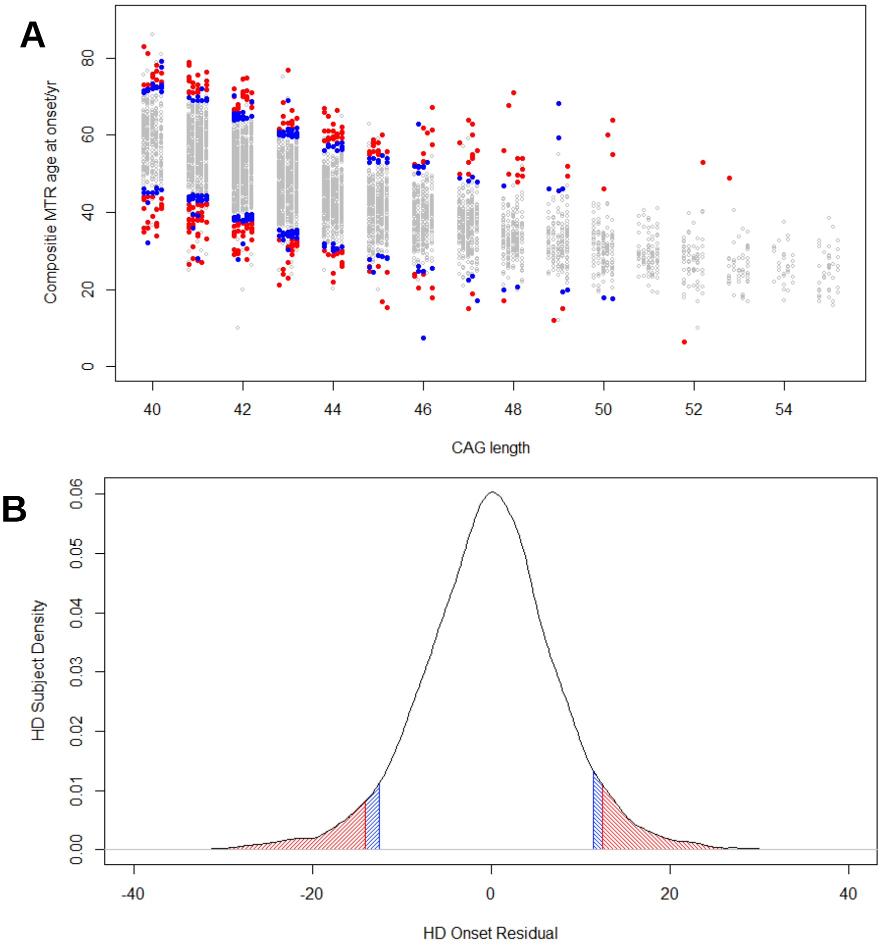


Figure 3.18: Extreme AMO cohort selection for sequencing. (A) Scatterplot showing individuals (red and blue dots) chosen based on age at motor onset (AMO) residual. Red dots are individuals from the first tranche and blue from the second. Smaller, grey dots are those not chosen for sequencing. Note the points on the graph have been jittered to improve legibility. Individuals missing sxraterm were not chosen for sequencing (barring one person). N=5902. (B) A normal distribution plot showing AMO residual. The first (red) and second (blue) tranches of sequencing are shown. No sxraterm individuals not included, N=5902.

3.7.2 Comparing the extreme cohort across symptoms

Although the extreme cohort used an AMO residual for selection purposes, we also wanted to examine other symptoms experienced by the group. Fig. 3.19 shows individuals selected on the basis of their AMO for all eight symptoms, using the best-estimate onset measures calculated in 3.2.2 and 3.2.3 for motor/cognitive symptoms, and sxrater adjusted CCQ data (2 year cut-off) for all other symptoms. The average ages for symptom onset are shown in Table 3.16A-B. For the early onset arm, Motor symptoms are first followed by irritable, depressive and psychotic symptoms. Cognitive impairment is the latest symptom on average. Comparatively, the first symptom in the late onset group interestingly was depression by 1.93 years, followed by motor symptoms and irritability. The latest symptom, again, is cognitive impairment.

Investigating further, we examined whether there was a difference between the frequency of symptom onsets between the early and late arms. As summarised in Table 3.16C, there are significant differences in frequency between depression, irritability and VAB, all of which are more frequent in the early onset participants. These remain significant after covarying for CAG length in a generalised linear model, which also further reinforces the small (~1 CAG) but significant difference in CAG length between early/late onset HD individuals. There is no significant difference for motor, cognitive, apathy, POB or psychosis symptoms between the two early/late arms.

A

E	Yes	No	Missing	Filtered	Unknown	Mean (Adj.)
MTR	214	2	7	5	22	34.23
COG	117	102	5	3	23	39.18
APT	103	115	4	5	23	38.52
DEP	134	67	4	22	23	36.59
POB	63	153	5	6	23	38.19
IRB	128	77	7	16	22	36.48
VAB	68	144	4	11	23	37.69
PSY	22	202	2	2	22	36.73

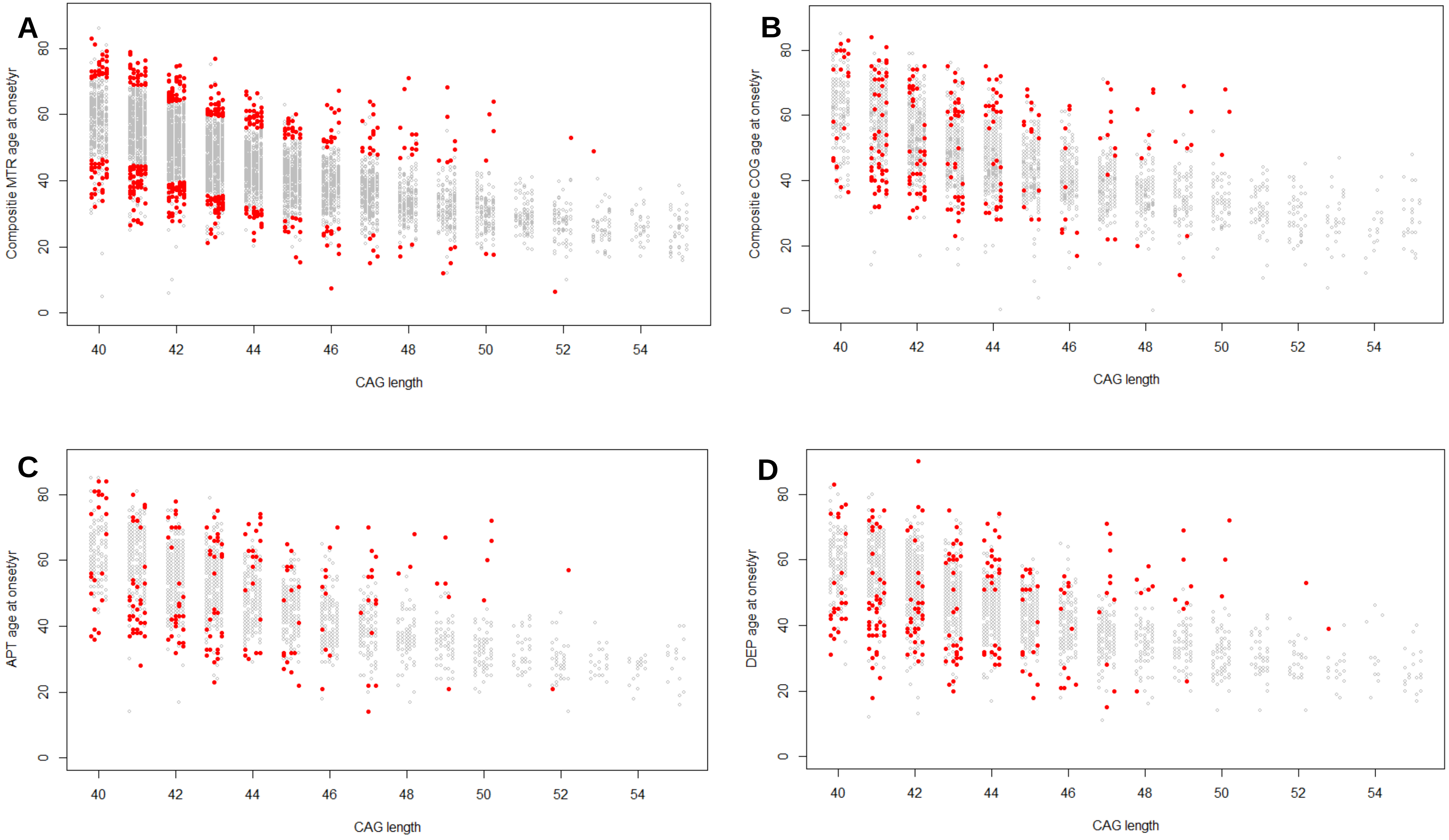
B

L	Yes	No	Missing	Filtered	Unknown	Mean (Adj.)
MTR	206	0	17	6	21	62.65
COG	105	107	15	2	21	65.61
APT	98	120	9	2	21	65.00
DEP	98	99	13	19	21	60.72
POB	67	154	5	3	21	65.04
IRB	99	119	5	6	21	62.93
VAB	42	181	3	2	22	64.31
PSY	15	211	2	1	21	64.73

C

Symptom	N	Chi-Square *p*	GLM(EorL~CCQ_Adj+CAG+Sex)		
			CCQ_Adj	CAG	Sex
MTR	422	1.66E-01	9.81E-01	**8.34E-07**	6.86E-01
COG	431	4.18E-01	4.15E-01	**1.40E-06**	4.53E-01
APT	436	6.31E-01	2.89E-01	**1.80E-06**	7.59E-01
DEP	398	**6.20E-04**	**1.57E-04**	**8.38E-07**	8.20E-01
POB	437	7.93E-01	9.82E-01	**4.73E-06**	5.48E-01
IRB	423	**4.49E-04**	**1.80E-04**	**4.06E-06**	3.67E-01
VAB	434	**1.63E-03**	**4.48E-04**	**1.08E-06**	5.61E-01
PSY	449	2.24E-01	1.21E-01	**2.18E-06**	4.25E-01

Table 3.16: Frequencies of symptoms in the extreme AMO cohort (CCQ). Frequencies of symptoms in early (A) and late (B) onset arms. Data used is CCQ data adjusting for sxrater using a 2 year cut-off (adj.). Yes = symptom present, No = symptom not experienced, Filtered = symptom experienced in lifetime, but >2 years earlier than the rater's estimate of HD onset. (C) shows *p* values for the same data using a chi-square test and a generalised linear model regressing early (1) or late (0) (EorL) on adjusted (Adj) CCQ data and CAG length. Note CCQ data was unavailable for 6 earlies and 15 lates, and one late individual had an unknown gender who was also excluded. Significant data are emboldened that pass multiple testing correction (8 tests Bonferroni *p*=6.25E-03), nominal values are italicised. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.


A
Compositie MTR age at onset/yr
0
20
40
60
80
40
42
44
46
48
50
52
54
CAG length
B
Compositie COG age at onset/yr
0
20
40
60
80
40
42
44
46
48
50
52
54
CAG length
C
APT age at onset/yr
0
20
40
60
80
40
42
44
46
48
50
52
54
CAG length
D
DEP age at onset/yr
0
20
40
60
80
40
42
44
46
48
50
52
54
CAG length

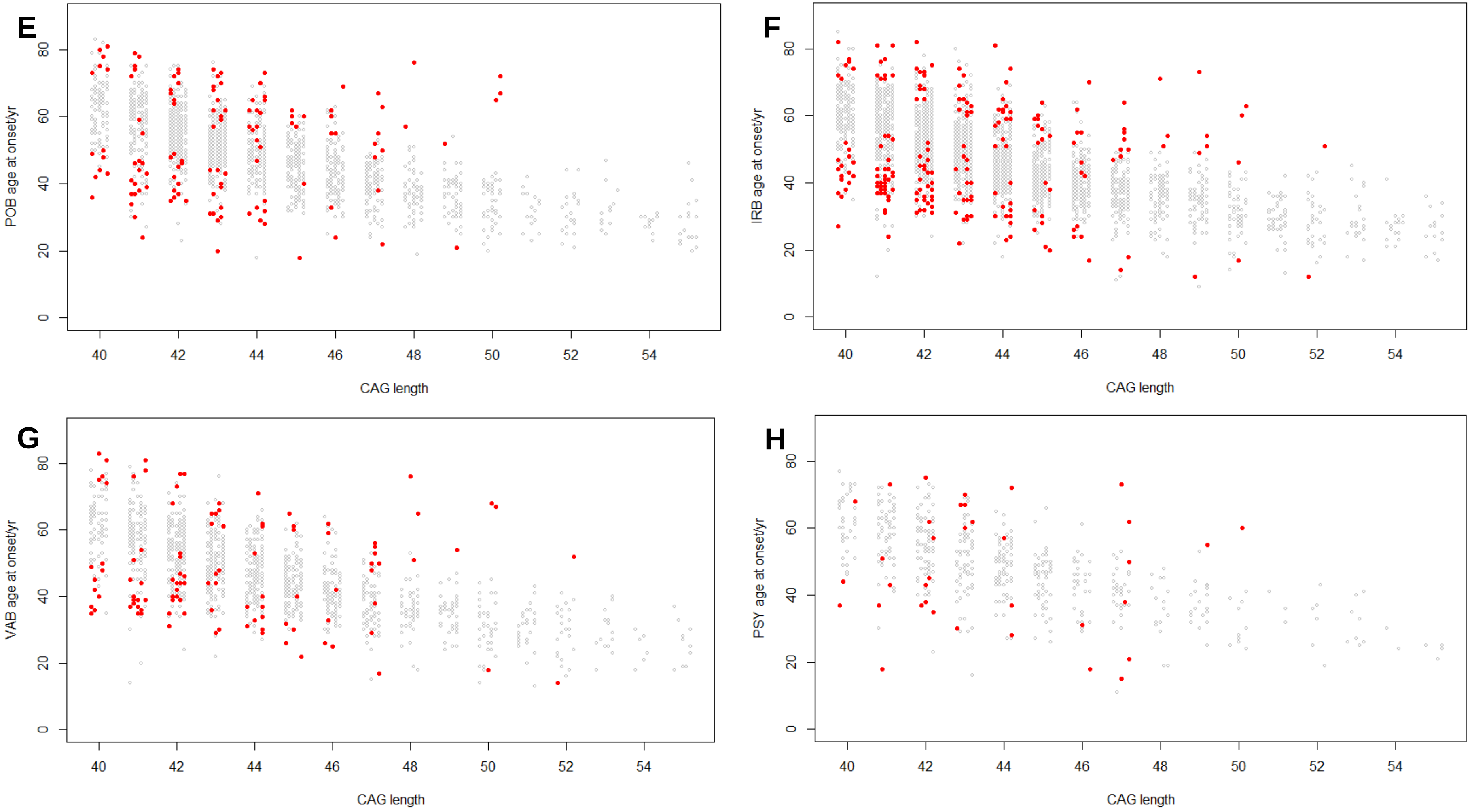


Figure 3.19: Ages at onset for other symptoms in the extreme AMO cohort. Extreme AMO onset individuals are shown as red circles for best-estimate motor (A) N=5585+500; best-estimate cognitive (B) N=3114+244; apathy (C) N=2335+201; depression (D) N=2504+232; POB (E) N=1623+130; irritability (F) N=2508+227; VAB (G) N=1288+110; and psychosis (H) N=525+37. Unsequenced individuals correspond to the initial number, sequenced numbers by the second number. Data is for CAGs 40-55. Motor/cognitive onset values calculated in 3.2.2 and 3.2.3.

3.8 Discussion

3.8.1 Overview of results

This chapter explored the use of the CCQ for HD onset age derivation, and, using these data, determined onset across multiple domains in a large cohort of HD patients (Registry-HD). Multivariate modelling investigated sex-differences in symptoms and found that depression is significantly more prevalent in women, and irritability/VAB symptoms are more prevalent in men, although irritability/VAB were only nominally significant in an extended model. There was no significant difference in the ages at onset of these symptoms between males and females after filtering CCQ data removing >2 years earlier than the clinician's estimate of onset (sxrater). Variation in symptom onset explained by CAG length (R^2) was calculated for each symptom. We find that expanded CAG has the poorest association with depressive and psychotic symptom onsets, although CAG still accounted for a large and significant proportion of onset variation for these symptoms. Anticipation was also estimated for Registry. Finally, an extreme onset cohort was selected using a residual age at motor onset calculated based on the Langbehn model (Langbehn et al., 2004). These individuals went on to be sequenced, as detailed in the next two chapters.

3.8.2 Best-estimate symptom onset estimation

The clinician's estimate of onset is commonly used for in patient studies (Orth and Schwenke, 2011; Aziz et al., 2018). Hence, it was chosen as the primary metric for deriving age at onset in the best-estimate onset calculations, supplemented by the CCQ. Symptom onsets were estimated for motor, cognitive and psychiatric onsets (age at first psychiatric symptom) in as many individuals as possible. Comparison of onset estimates between CCQ and sxrater for motor onset showed a high degree of concordance (91.9%). Sxrater and CCQ for cognition was less consistent, but still agreed in over three quarters of cases (76.4%). One reason for this slight disparity between CCQ/sxrater is the cognitive CCQ specifically asks when cognitive impairment begins to affect daily life (see Table 2.2). This differs for all other CCQ symptoms which ask the participant/family whether a symptom has been experienced at all in lifetime rather than assigning any symptom severity. The slight distinction in definition may have contributed towards the larger disparity between sxrater and CCQ for cognitive onsets, especially given cognitive CCQ tended to be later than sxrater-derived estimates.

The best-estimate onset estimation method used has several disadvantages outside of subtle (and potentially systematic) differences between the way the CCQ and sxrater are called. Best-estimate ages at onset were determined for as many individuals as possible, including those where only one of either sxrater or CCQ was available. As no comparison between CCQ/sxrater was possible in these cases, some onset estimates may be less reliable. Furthermore, the best-estimate age at onset determination relies on manual curation of some data where sxrater and CCQ differ. Although we attempted to be as systematic as possible when deriving these estimates, differences between data inclusion could arise between research groups. Finally, it is notable we did not factor in the rater's confidence level (sxestcfd) into our analysis, which can either be given as 'high' or 'low', as (1) this was not available for many individuals and (2) we were unsure how accurate the metric was.

3.8.3 Symptom onset determination using CCQ

The CCQ was used to determine the onset of eight symptoms: motor, cognitive, apathy, depression, perseveration/obsessive behaviour (POB), irritability, violent/aggressive behaviour (VAB) and psychosis. Symptom onset calculated using unadjusted, raw CCQ was found to have a variable distribution and some individuals had much earlier symptom onsets than expected given their sxrater. This stems from the CCQ capturing symptoms arising prior to clinical HD diagnosis. It is known significant neurodegeneration occurs prior to traditional neurological signs and HD diagnosis (Aylward et al., 1997, 2004; Paulsen et al., 2008). Furthermore, it has been reported that pre-manifest HD gene carriers experience increased rates of apathy, depression and other neuropsychiatric disturbances than non-carriers (Folstein et al., 1983b, 1983a; Julien et al., 2007; Klöppel et al., 2010; Tabrizi et al., 2013; Martinez-Horta et al., 2016) (see also 1.2.1). However, psychiatric symptoms, and especially depressive symptoms and major depression, occur at high rates in the general population (Kessler et al., 2005; Bromet et al., 2011; Ferrari et al., 2013; Salk et al., 2017). Experiencing the deterioration or hospitalisation of a parent with HD may also the risk for depression in an HD family. Hence it is extremely difficult to distinguish between symptoms that are directly the result of HD neurological changes and those that may originate from indirect environmental or general population effects. There may also be interaction effects between these factors, although this is beyond the scope of our study.

In our approach, we filtered symptom data occurring >2 years earlier than the clinician's estimate of onset (sxrater). We purposely chose the most stringent cut-off here to contrast with the raw CCQ-derived symptom data, however we are likely removing some symptoms that are the result of early, pre-manifest HD pathology. Although a comparison of 2, 5 and 10 year cut-offs only showed minor differences, suggesting most symptoms removed in this analysis occurred much earlier than HD diagnosis (>10 years earlier). This stringency does, however, afford the removal of many symptoms that may be unrelated to HD, and may be important to consider for the interpretation of symptom data (especially for sex-dependent effects, see 3.8.6). Accordingly, we present both unadjusted and adjusted CCQ data in many of our downstream analyses with the caveat the stringent adjusted CCQ data is probably removing some pre-manifest HD symptoms that we cannot disentangle from population-based effects using CCQ alone.

We found for several symptoms, especially those more, although not uniquely, characteristic of HD such as apathy and motor symptoms, little data was filtered in the sxrater adjusted CCQ data. However, this sxrater adjustment approach does preclude individuals without both an sxrater and known age at CCQ symptom. A further disadvantage of this method is that only one date is recorded by the CCQ; therefore, if someone reports depression earlier in life, all subsequent depressive episodes are not recorded. Possibly (or even logically) individuals with earlier symptom episodes may be more likely to experience those symptoms in their HD, hence this methodology likely underestimates symptom prevalence. In future observational studies, collecting data on more than one episode of each symptom in the CCQ may allow for a better estimation of symptom prevalence and onset for as many participants as possible.

Overall, we find that the CCQ is a useful tool that can complement the rater's estimate of onset, and supplementation with sxrater may be useful to consider in some contexts. Using the CCQ without any adjustment has the danger of capturing symptoms that may be unrelated to HD, although, as discussed, disentangling the origin of these symptoms is not possible with CCQ data alone. Additionally, there are several further limitations to consider when using CCQ data. CCQ is an imprecise instrument as its data is retrospectively derived and thus subject to participant recall bias and error. It is essentially a pseudo-cross-sectional measure as it only captures whether an individual has experienced a symptom and at what age this occurred for the first time. CCQ also does not consider symptom severity or frequency, and thus relies on other clinically gathered data such as TMS or TFC, or the problem behaviours assessment (PBA). Considering other instruments, such as the HADS or PBA, may be useful in combination with the CCQ for capturing longitudinal changes in HD patients.

Additionally, CCQ lacks granularity. For symptomatic individuals with an sxrater, almost all participants in Registry (~98-99%) reported a positive motor CCQ making the predictive models for motor symptom data less powerful. CCQ, too, is unlikely to pick up subtle cognitive or psychiatric/behavioural changes that occur well before motor conversion, although it may for some individuals. Further investigation of CCQ sensitivity is warranted in future study. However, unlike the best-estimate onset measures derived in this chapter, CCQ onset data is derived entirely systematically, ensuring estimates between different research sites and groups are consistent. Though, adjusting for sxrater does preclude some individuals for whom CCQ is not available (this includes the older R2 cut of the Registry dataset). CCQ, as with all clinically gathered data, is open to inter-rater variability and possibly inter-site variability.

3.8.4 The effect of CAG length on symptom onset

Motor onset was calculated for a total of 6520 individuals for whom CAG lengths were known. To our knowledge, this is the largest study of CAG length on symptom onset in Huntington's disease, with the largest previously having been the study from Lee et al. in 2012 with 4068 manifest motor-onset participants (Lee et al., 2012c). We find that of the eight symptoms investigated, motor symptoms track most consistently with CAG length, ranging from R^2 = 0.621 to 0.659 depending on the stringency of the cut-offs and derivation used (N=6511 and 5303, respectively, see Table 3.12). These results are similar, although marginally higher for the most stringent cut-off, to the study from Lee et al. which found R^2 ranging from 0.637 to 0.653 using a similar methodology in a large cohort of HD patients, including Registry-HD participants.

Our AMO results are in-line with many other studies (Andrew et al., 1993; Duyao et al., 1993; Snell et al., 1993; Illarioshkin et al., 1994; Kieburtz et al., 1994; Rosenblatt et al., 2001; Aylward et al., 2004; Wexler et al., 2004; Rinaldi et al., 2012), however it is important to note the robustness of the CAG and age at onset association is highly dependent on the original data derived. Studies using onset data from the Venezuelan kindred, a closely followed and assessed cohort of HD participants, many of whom are related, have reported CAG length R^2 between 0.67-0.73 based on a predominantly motor onset (Wexler et al., 2004; Andresen et al., 2007b, 2007a). HD cohorts may also have varying onset estimates, which may stem from differences in population (Ramos et al., 2012b), relatedness, clinical assessment and study design, and these may also account for the differences seen in the Venezuelan cohort compared to our study. We found the mean AMOs in our study track the frequently used Langbehn et al., 2004 model very strongly for most repeat lengths for AMO. Disparity between our data and the Langbehn model could be due to slight under-reporting at smaller CAG lengths (40-41 repeats). It may also be due to ceiling effects; individuals with smaller CAG lengths may die from other causes before developing HD and coming to clinical attention, whereas the Langbehn model is based on an HD population with complete (or nearly complete) disease penetrance (41-56 CAGs).

In addition to AMO, we also determined CAG length association with age at onset of seven other symptoms in HD: cognition, apathy, depression, POB, irritability, VAB and psychosis. Our results show that after removing CCQ data occurring much earlier than the clinician's estimate of HD onset, moderate cognitive impairment as described by the CCQ tracks CAG length very strongly (R^2=0.644) (Table 3.12). Apathy and behavioural changes irritability, POB and VAB have similar R^2 ranging from 0.581-0.618. Notably both psychosis and depression had the lowest association with CAG length (0.450 and 0.494, respectively), although the correlation is still quite substantial and significant.

Most studies investigating CAG length on symptom onset in HD have used motor onset as the primary outcome measure; several studies have used a combinatorial approach to generate an age at first symptom onset across symptomatic domains, however these do not distinguish between onset types (Illarioshkin et al., 1994; Brandt et al., 1996; Pekmezovic et al., 2007; Rinaldi et al., 2012). Andrew et al., 1993 distinguished between motor symptoms (chorea), dementia and psychiatric abnormalities, although only had small numbers of the latter two (N=39 for dementia, N=84 for psychiatric signs). They found psychiatric symptoms had the weakest correlation compared to dementia and motor symptoms, however psychiatric symptoms were not further defined (*e.g.* depression, apathy, *etc.*). Consistent with our results, dementia onset, a form of advanced cognitive impairment, had a higher correlation with CAG length than psychiatric symptoms, but this was lower than the correlation of motor symptoms in their study. Vassos et al. (Vassos et al., 2008) found similar results; CAG length accounted for a highly significant but lower proportion of the variation of first psychiatric symptom onset (n=49, R^2=0.51) compared to motor features (n=66,R^2=0.78).

The finding that psychosis and depression had smaller (although still highly significant) associations with CAG length than other symptoms was interesting, although interpretation is not straightforward. HD psychopathology can be misattributed to other syndromes (*e.g.* major depression or schizophrenia (SZ)), especially before development of characteristic neurological signs (Kumar and Jog, 2011; Martino et al., 2013; Pascu et al., 2015), particularly those without a family history of HD. HD patients can also have co-morbidity with other disorders which may be distinct from HD (Sipilä et al., 2016) such as SZ, although we have attempted to reduce this bias in our study by removing SZ or similar disorders as defined by ICD-10 for psychotic symptoms in our extended model. Furthermore, CCQ data is retrospective in nature from patients/family, and while adjusting for the age at onset determined by the rater as described likely improves onset estimation, it introduces its own biases. These problems can make onset estimates less reliable and affect the R^2 values calculated in our data. Interestingly, psychotic symptoms in HD tend to have a familial predisposition (Lovestone et al., 1996; Tsuang et al., 2000), potentially suggesting additional or separate mechanisms by which psychosis arises in HD patients, and several gene candidates have been proposed by Tsuang and colleagues (Tsuang et al., 2018). Possibly, then, some of the unexplained variability not captured by CAG length may be explained by these additional genetic factors. In support of this, a recent study showed a degree of genetic overlap between certain HD psychiatric phenotypes and other neurological diseases (including schizophrenia) (Ellis et al., 2019).

CAG length was found to correlate almost as highly for CCQ-derived cognitive onset as it did for motor symptoms. It is known that mild cognitive impairment, as well as apathy, can predate HD diagnosis (Paulsen et al., 2008; Duff et al., 2010) and these worsen over time until disease diagnosis (Baake et al., 2017) (see 1.2). Although cognitive CCQ probably lacks the sensitivity to capture subtle changes to cognition, the CCQ instead focuses on a disease milestone well into HD pathology where cognitive impairment begins to affect every day functioning. It would be interesting to see at what point cognitive CCQ conversion occurred when comparing to the symbol digit modalities test (SDMT) or more robust cognitive batteries such as those included in the UHDRS, *e.g.* Stroop word test, both of which are strong predictors of HD cognitive phenotype (Tabrizi et al., 2013; Braisch et al., 2019). Further, finding that cognitive, motor and apathy symptoms all have similar levels of variation explained by the CAG length is consistent with the same underlying mechanism driving pathology for these symptoms. Cognitive, motor and apathy symptoms all track HD clinical progression (Thompson et al., 2012; Fritz et al., 2018) and are associated with the atrophy of subcortical brain structures (Aylward et al., 1997, 2004; Misiura et al., 2017, 2019; Baake et al., 2018). Behavioural symptoms such as POB and irritability/VAB have a large amount of their variance explained by CAG length in our study as well.

We found no significant effect of the wild-type *HTT* CAG length on HD onset (CCQ; Table 3.11) after accounting for the expanded CAG length. There have been inconsistent findings about the role of the wild-type CAG length and whether it has an effect on disease (Farrer et al., 1993; Snell et al., 1993; Warner et al., 1993; Djoussé et al., 2003; Aziz et al., 2009); however, our findings replicate more recent studies (Klempíř et al., 2011; Lee et al., 2012c) that also find no role for wild-type CAG length on onset. Our study, however, does not necessarily preclude a role of the wild-type CAG in clinical severity nor progression which have been reported to be linked to wild-type *HTT* CAG length (Aziz et al., 2009). Additionally, we did not explore whether a second expanded allele affected phenotype, as has been reported elsewhere (Squitieri et al., 2003).

3.8.5 Anticipation in Registry

HTT alleles are unstable during intergenerational (vertical) transmission, leading to the phenomenon of anticipation where offspring tend to have longer CAG lengths and, thus, an earlier age at HD onset (Teisberg, 1995; McInnis, 1996) (1.3). Using retrospectively derived onset data for parents, we calculated an average anticipation for maternally inherited alleles as 2.88 years and paternal alleles as 7.21 years. It has been well-established that anticipation is larger in paternally transmitted alleles (Ridley et al., 1988; Duyao et al., 1993; Wheeler et al., 2007; Aziz et al., 2011; Ramos et al., 2012a), probably arising during spermatogenesis (Yoon et al., 2003; Wheeler et al., 2007; Simard et al., 2014; Neto et al., 2017; Jamali et al., 2018). Our estimates for anticipation are similar to several other studies (Ranen et al., 1995; Margolis et al., 1999; Demetriou et al., 2018), although slightly lower than others (Cannella et al., 2004). Some differences may stem from using retrospective non-clinical estimates for age at onset for parents. Having single nucleotide polymorphism (SNP) genotype data for these individuals, many of which are available (GeM-HD Consortium, 2015, 2019), could allow for an examination of genetic liability for expansions to occur vertically using a GWAS. Any identified factors could be similar to those that affect onset as reported elsewhere (GeM-HD Consortium, 2015, 2019), although both the retrospectively-derived anticipation data and the stochastic nature of anticipation could make identification of genetic loci difficult. As others have found, approximately 2/3 of JHD cases arise paternally (Ridley et al., 1988; Myers et al., 1993; Ranen et al., 1995). We also find *de novo* HD cases account for ~7.8% of HD cases.

3.8.6 Sex differences in HD symptoms

We found that there was a ~10.5% higher prevalence of reported depressive symptoms in women than men in Registry after removing symptoms occurring >2 years earlier than the rater’s estimate of HD onset. This was confirmed using logistic generalised linear modelling which found significant sex differences between depression, irritability and VAB symptoms. An extended generalised linear model with more stringent inclusion criteria and more covariates, although fewer participants, found depression remained significant, and irritability/VAB retained nominal significance between males and females. In the general population, depression is ~1.5-2.0x more common in women than men (Piccinelli and Wilkinson, 2000; Ferrari et al., 2013; Salk et al., 2017), however it has been recently reported that there is no sex difference in the likelihood of depression in HD using the HADS (Dale et al., 2016). We do, however, replicate the Dale study finding that TFC score is strongly and negatively associated with depression. TFC score was also negatively associated with all other symptoms we derived in symptomatic HD patients, with the notable exception of motor symptoms where our model was underpowered, demonstrating psychiatric disturbances as reported by the CCQ tend to increase over disease course.

The HADS scores, total depression score (TDS) and total anxiety scores (TAS), were then used as in the Dale et al. study with the same selection criteria and covariates, excepting medication, and gave similar results to their study. Notably, the TDS was strongly associated with TFC and negatively nominally significant with education years. We find a nominally significant difference for sex using TDS; however, this is in the opposite direction than expected, with males having slightly higher TDS than females. TAS had no significant correlation with any of the variables explored, and only had a nominally significant association with CAG length, education years and age at onset. We also expanded on the Dale study by considering the total irritability score (TIS; derived from SIS). We found that TIS shows significant and negative associations for both CAG length and age, and these results are very similar to those seen with CCQ for irritability. Although there was no significant difference for TIS and sex, men tended to have higher scores than women, so it would be very interesting to see the TIS repeated in a larger cohort of HD patients.

The reason for the differences between the TDS and depression CCQ are unclear. The Dale et al. study used the HADS in a moderately sized cohort of HD patients (N=453) who also originate from Registry as those in our study do, and there is likely substantial overlap between the participants in our two studies. The HADS is widely used across a range of psychiatric disorders and is considered robust (Bjelland et al., 2002); in HD, the HADS has been used as a short-form test in HD patient cohorts (De Souza et al., 2010; Dale et al., 2016) and is a recommended scale for depression by Mestre et al. (Mestre et al., 2016). We used our strictest cut-off for CCQ measure to limit any bias originating from symptoms occurring before HD onset, so it is unlikely these are affecting the results. Apathy CCQ had the strongest association with TDS and depression has the strongest association with TAS, although these associations are somewhat small (0.252 and 0.220, respectively). CCQ and TDS may simply be capturing depression differently; whereas TDS is a thought-based measure of depression from the patient (*e.g.* I feel cheerful), the CCQ is a binary yes or no, completed with input from the participant, family and clinician.

Irritability and VAB are only nominally significantly associated with sex following multiple testing correction, although this may be due to reduced numbers in the extended model. The literature has been somewhat mixed with some studies finding no difference for the sexes in HD irritability/violence (Pflanz et al., 1991; Shiwach and Patel, 1993; Reedeker et al., 2012) and others finding that male HD patients exhibit more violent behaviour (Tyler et al., 1983). One study reported an increase in crime in only male HD patients (Jensen et al., 1998), although violent crime on its own was not significantly higher in HD patients, possibly due to a small sample size (N=99). Some of these differences in observation likely arise from (1) large ranges in N and (2) significant differences in study design and outcome measure. There is no single measure that is used to assess behavioural changes in HD (reviewed in (Mestre et al., 2016)). Some scales include the neuropsychiatric inventory (NPI) (Paulsen et al., 2001), the PBA (Craufurd et al., 2001; Kingma et al., 2008) and the irritability scale (IS) (Klöppel et al., 2010; Reedeker et al., 2012). Thus, while we find that there is a significant difference in irritability/VAB prevalence between males and females, we would be interested to see how other behavioural metrics are associated with the CCQ.

Strikingly we also find that age at onset (sxrater, regardless of onset type) is independently and significantly negatively associated with symptom prevalence for cognition, depression, irritability and VAB, and nominally so for psychosis in adult HD, *i.e.* younger manifest HD adults are more likely to report cognitive impairment and psychiatric symptoms. This is further shown in the types of onsets called in Registry which are more likely to be called as psychiatric, mixed or cognitive in younger age groups. A similar effect was reported by Rocha et al. (Rocha et al., 2018) for psychotic symptoms. Age being a risk factor for psychiatric symptoms may have clinical relevance for the management of HD in younger individuals.

Education was identified as being negatively associated with HD psychotic symptoms, with individuals having a higher number of education years less likely to report psychiatric symptoms in their HD disease course. This is reminiscent of schizophrenia where educational attainment is negatively associated with schizophrenia risk (Okbay et al., 2016; Bansal et al., 2018; Escott-Price et al., 2019). This observation further supports the previous finding that psychotic symptoms are least associated with CAG length in our data, suggesting risk factors for psychosis may, in part, be associated with common variation in the general population. The largest study of psychosis in HD that we know of is the (Rocha et al., 2018) study encompassing 2303 manifest HD individuals, with 248 having psychosis in their lifetime. They did not factor educational attainment into their model, however they did find significant negative associations with onset and CAG length, as we also find at nominal significance after multiple testing correction. They did not find TFC to be significant in their model (although it is approaching significance), however they do find part B of the trail making test (TMT-B) significant, and this may be capturing a similar or related effect.

3.8.7 Selection of an extreme onset cohort

A total of 500 individuals were chosen for whole-exome sequencing by stratifying the cohort by residual AMO. AMO was chosen as the primary selection criteria as (1) AMO has been used for cohort selection previously and successfully (GeM-HD Consortium, 2015, 2019); (2) genotype data was available for many of the calculated AMO individuals through the GeM-HD GWA study (147 in GWA3, 337 in GWA4); (3) further progression data was available for a subset of these individuals (the 147 in GWA3) (Hensman Moss et al., 2017) (4) AMO tracked CAG length most strongly of all symptoms. Given only 500 individuals could be sequenced, we elected to use motor onset as it was the most strongly associated with CAG length, easily identifiable clinically (especially in an at-risk population), readily present in most manifest HD patients and reasonably HD-specific. Although other symptom onset information could be factored in *post hoc*. It would be interesting to compare our selection method to others in the literature, such as the Braisch et al. (Braisch et al., 2017) extreme motor onset cohort or the Braisch et al. (Braisch et al., 2019) cohort selected using extreme SDMT scorers.

Significant differences were seen between the disease-causing CAG lengths in the early and late groups (42.6 vs 43.6 early/late), which makes CAG length an important covariate for downstream analyses when examining the *HTT* gene. Future cohort selection may find balancing CAG length useful for downstream study when considering instability, although this is further complicated by potential structural differences in *HTT* which affect the size of the pure CAG (see chapter 5), as *HTT* alleles are not routinely sequenced currently. Significant differences for irritability, VAB and depressive symptoms were seen between the early and late arms after covarying for sex and CAG. A comparable effect was seen in the generalised linear models where earlier sxrater-derived onset was negatively and significantly associated with cognitive, depression, irritability and VAB symptoms, and is also similar to the observation earlier onsets are more likely to be classed as psychiatric, cognitive or mixed onset types.

This chapter saw the derivation of ages at symptom onset in what we believe is the largest study of phenotype and CAG length in HD to date. Eight separate symptoms were analysed, primarily drawing from CCQ data. CCQ data was concordant with the rater’s estimate of onset in >90% of cases for motor symptoms and >75% for cognitive impairment. Depression was identified as having a significantly higher prevalence in women compared to men in disease course when using the CCQ, and irritability and VAB symptoms were significantly more prevalent in men. We also find age is a risk factor for several psychiatric symptoms in manifest HD individuals. The variation explained by CAG length was then calculated across a range of symptoms. Depression and psychosis were the least associated with the length of CAG and may reflect mechanistic differences in the pathology of these symptoms. Finally, an extreme onset cohort was chosen based on residual age at motor onset. These individuals will be sequenced in the proceeding chapters.

Chapter 4: Exome sequencing of an extreme motor onset cohort of HD patients

4.1 Introduction

The HD expanded CAG repeat tract is found in exon 1 the of *Huntingtin* gene (*HTT*), and its length is the is the primary determinant for age at disease onset (Wexler et al., 2004; Lee et al., 2012c) accounting for between ~50-70% of the variance observed (see chapter 3 & Tables 3.7-3.8). However, the remaining ~30-50% variation is known to have a significant and independent genetic component (Wexler et al., 2004). Recent GWA studies and other human genetics have highlighted several candidate genes associated with altered HD onset including *FAN1*, *MSH3*, *MLH1* and *HTT* allele structure (GeM-HD Consortium, 2015, 2019; Lee et al., 2017; Ciosi et al., 2019; Wright et al., 2019). Many of the implicated modifying genes from GWA study are involved in DNA repair pathways, and these likely act *via* somatic instability, wherein the extended CAG repeat undergoes progressive expansion in somatic cells (see 1.5/1.7).

To date, GWA studies have identified >70,000 trait-associated loci (Buniello et al., 2019). GWA studies are effective in identifying common variation associated with disease (Altshuler et al., 2008), however one of their main limitations is in narrowing down candidate genes from implicated loci, especially given many of the signals identified are non-coding (Maurano et al., 2012). Additionally, variants of large effect size such as loss-of-function (LoF) or other damaging non-synonymous (NS) variation are typically rare and not captured by standard array-based imputation (Cirulli and Goldstein, 2010). In contrast, whole-exome sequencing (WES) is a next-generation sequencing (NGS) modality that sequences the entire coding portion of the genome (Hodges et al., 2007; Gnirke et al., 2009), constituting ~1% of the total human genome. Unlike genotyping arrays, WES can capture the entire coding variation present in an individual regardless of rarity and can identify variation that is difficult or impossible to detect by GWAS (Visscher et al., 2017). By implicitly examining coding variation, exome sequencing has the power to detect rare variants of potentially large effect size that translate to changes to protein structure and function.

Since its outset, numerous studies have used WES to identify causative mutations in rare disease (Choi et al., 2009; Ng et al., 2009, 2010a, 2010b; Roach et al., 2010; Muona et al., 2015), and ~25% of rare diseases, predominantly those of known loci, can be clinically diagnosed using WES (Yang et al., 2013). WES has also identified *de novo* variation to be both frequent and a major contributary factor towards disorders including schizophrenia (Xu et al., 2011; Fromer et al., 2014; Rees et al., 2019), autism spectrum disorder (ASD) (Neale et al., 2012; De Rubeis et al., 2014) and other developmental disorders (Ku et al., 2013; Deciphering Developmental Disorders Study, 2017). Further, WES can also be used in rare-variant association studies to detect low frequency and rare coding variants contributing towards common, complex diseases such as diabetes (Flannick et al., 2019), Alzheimer's disease (Bis et al., 2018; Raghavan et al., 2018) and schizophrenia (Purcell et al., 2014; Genovese et al., 2016), thereby providing further insight into the genetic architecture of disease.

To build upon previous genetic work in HD, we wanted to investigate whether rare genetic variants of potentially large effect size could modulate HD disease onset. We stratified individuals from the large Registry-HD cohort (N~6000) by their age at motor onset (see chapter 3) and sequenced the 250 earliest and 250 latest individuals compared to their expected onset based on CAG length alone (age at motor onset residual) using WES. Individuals with extreme early/late onset are more likely to be enriched for genetic modifiers, and so our strategy maximised the study's power given we were unable to sequence the entire Registry cohort with our resources. This chapter details the generation and quality control (QC) of these exome data using in-house sample QC and annotation pipelines. These data are then analysed; first, *HTT* allele structure is assessed using WES. Then, candidate genes previously implicated by GWA study and other genetic work are investigated such as *FAN1*, *TCERG1* and *MSH3*. We then performed whole-exome rare-variant analysis using dichotomous and continuous phenotypes with burden regression and sequence kernel association tests (SKAT). Finally, gene set enrichment analyses of rare variation are considered. The individuals from this chapter are then used in chapter 5 where *HTT* allele structure is more rigorously assessed using a targeted NGS method.

4.2 Exome sequencing of HD patient DNA

Whole-exome libraries for the 500 selected individuals from 3.7 were prepared using TruSeq® Rapid Exome library kits as described (methods 2.6.1). Six plates were prepared as 96-plex libraries and one plate as a 12-plex library. Where possible, care was taken to distribute early and late onset samples equally during library preparation. Completed libraries were then given to the MRC core team where libraries were clustered using a cBot system and sequenced on an in-house HiSeq 4000 (methods 2.6.2).

De-multiplexed FASTQ files were aligned to a GRCh37 reference to generate variant-ready binary alignment map (BAM) files, and a compiled variant calling file (VCF) was created with a local GATK-best practices pipeline (methods 2.7.1.1). Of the six 96-plex libraries, four sequenced as expected, one had a single exome failure during initial fragmentation and barcoding (95 successful samples) and one plate had three pools fail during one of the two capture and hybridisation steps (each pool is 12-plex, *i.e.* 96-36=60 successful samples). A final total of 551 samples were sequenced (Table 4.1), most of these derived from DNA extracted from HD patient lymphoblastoid cell lines (LBCs). In addition, four samples from patient blood and nine HD induced pluripotent stem cells (IPSC) lines (HD iPSC Consortium, 2012) were also sequenced. While the analysis of the IPSC lines is outside the scope of this project, these samples were included for exome QC purposes.

Sample source	N
Early HD onset (LBC)	250 (16)
Early HD onset (Blood)	2
Late HD onset (LBC)	250 (15)
Late HD onset (Blood)	2
Normal HD onset (Blood)	7
IPSC-derived	9

Table 4.1: An overview of the samples exome sequenced. Note the numbers in brackets refer to the number of samples for which sequencing had to be repeated (usually due to some sort of QC failure). Final N=551. LBC: lymphoblastoid; IPSC: induced pluripotent stem cell.

4.3 Exome quality control pipeline

4.3.1 Establishing exome quality using Picard and Hail

For an overview of the exome quality control pipeline, see the provided flowchart (methods 2.7.1.2, Fig. 2.5). Picard, an NGS tool, was used to produce target coverage and mean sample depth statistics for each BAM file (https://github.com/broadinstitute/picard). Both target coverage and mean depth correlate strongly (*r*=0.915) and are presented in Fig. 4.1 for each plate. We opted to use target coverage as an initial QC check, and re-sequenced samples with <70% of the exome covered. Although sequencing libraries were prepared with equal numbers of early residual and late residual onset samples, there was still a small but significant difference between early and late sample QC metrics. Mean exome coverage at ≥10X in early onset exomes was 82.3% and 80.8% for late exomes (*p*=3.38E-04 Welch two sample t-test). The mean target depth in earlies was 30.34 and 28.82 for lates (*p*=2.65E-03 Welch two sample t-test).

Following Picard QC on the BAMs, we then performed a second QC on the resultant VCF using Hail (https://github.com/hail-is/hail). Hail computes several QC measures including genotype quality mean, call rate and mean variant depth. As before, these metrics correlate strongly with each other, although call rate less so than depth/genotype quality (genotype quality and depth *r*=0.98; genotype quality and call rate *r*=0.74; depth and call rate *r*=0.71). The average genotype quality was 67.82 (early) and 66.64 (late); average depth was 28.80 (early) and 28.13 (late); and the average call rate was 97.80% (early) and 97.63% (late). To be systematic, we removed exomes >3 standard deviations less than the mean for any of these metrics (Fig. 4.2). 3 standard deviations was chosen as it gave a good combination of stringency and inclusivity (*i.e.* not losing too many exomes) based on the plots in Fig. 4.2. Following sample repetition, 499 of the 500 lymphoblastoid-derived exomes passed both Picard and Hail QC.

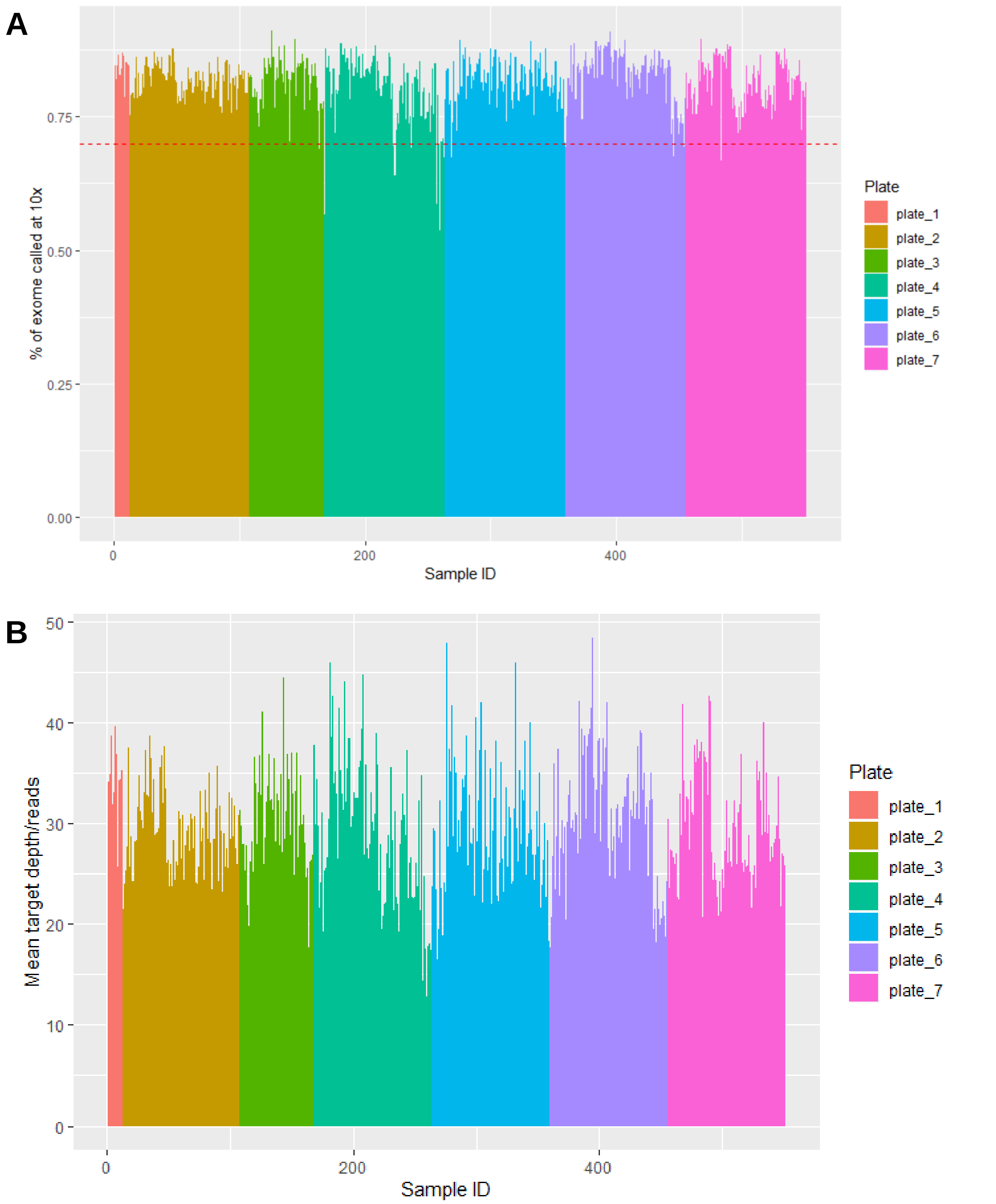


Figure 4.1: Quality control metrics for exome libraries using Picard. Picard-derived QC metrics are shown for (A) percentage of target exome covered and (B) mean target depth (average number of reads). The samples are arranged in order of pooled libraries, and the colours refer to the different plates in order of preparation. The red dashed line in (A) indicates the 70% coverage rate at 10X. N=551.

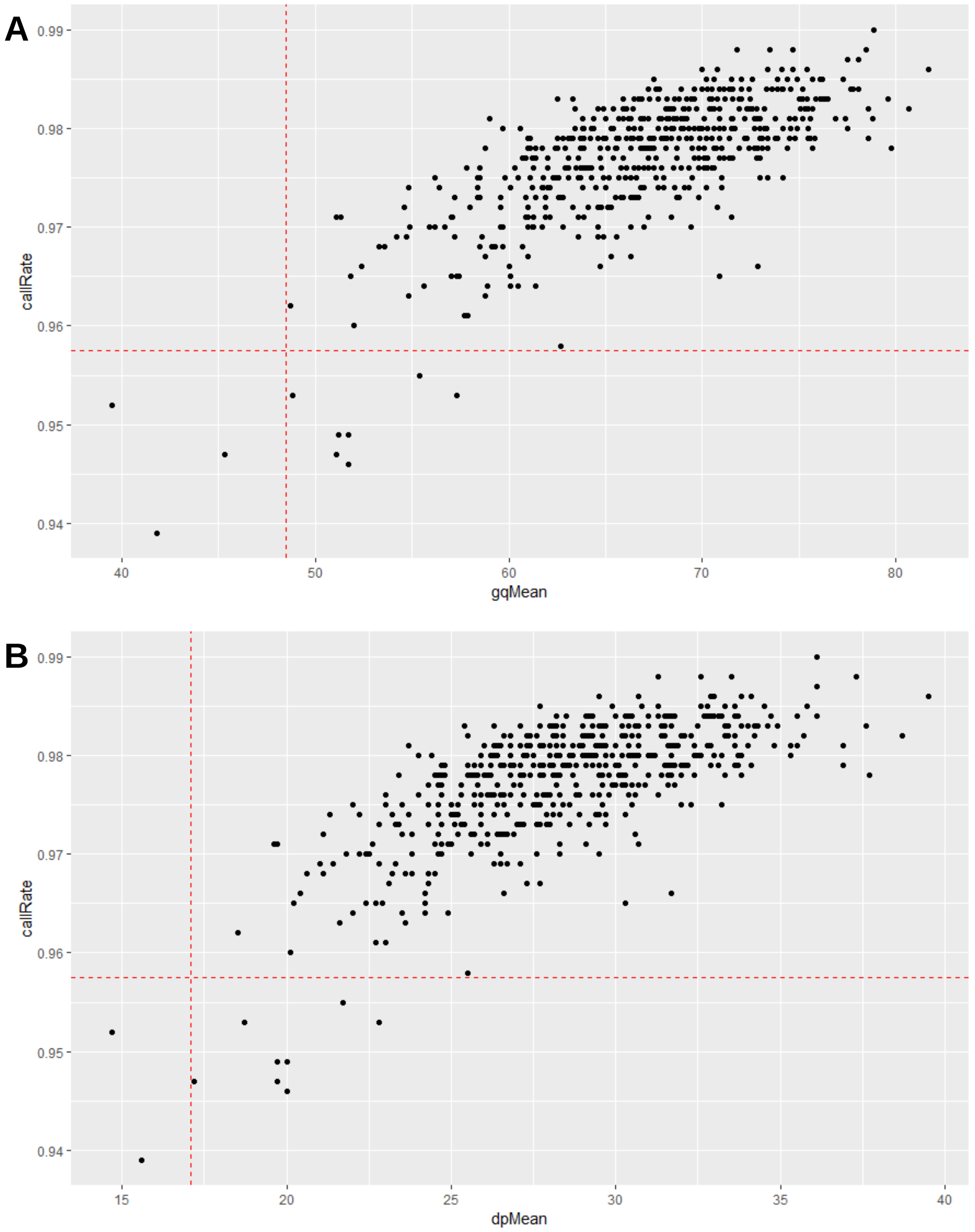


Figure 4.2: Quality control metrics for exome libraries using Hail. Call rate (callRate) and genotype quality mean (gqMean) are shown in (A); call rate and mean variant depth (dpMean) are shown in (B). The dashed red line indicates the 3 standard deviation cut-off for sample inclusion. QC metrics derived from Hail. N=551 for both (A) and (B).

4.3.2 Detecting exome contamination

Contamination of samples can arise due to handling error during library preparation or original sample contamination and must be addressed to prevent biases in the data. VerifyBamID (Jun et al., 2012) was used to generate a sequence-only contamination ratio (‘Freemix’) in our BAM files passing QC. In addition, we also calculated the heterozygote/homozygote (Het/Hom) ratio in our VCF calculated by Hail, as this is roughly equivalent (*r*=0.94). Both approaches detect sites where there are >2 alleles, as such sites are indicative of contamination, and results for these analyses are shown in Fig. 4.3. We used the same implementation as the ExAC study where samples with a contamination ratio > 0.075 were removed (Lek et al., 2016). 5 samples were predicted to have high levels of contamination and were repeated. After repetition of these samples, no further contamination was detected demonstrating the contamination likely occurred during DNA preparation. All lymphoblastoid-derived exomes up to this point, 499 of 500, passed this level of QC.

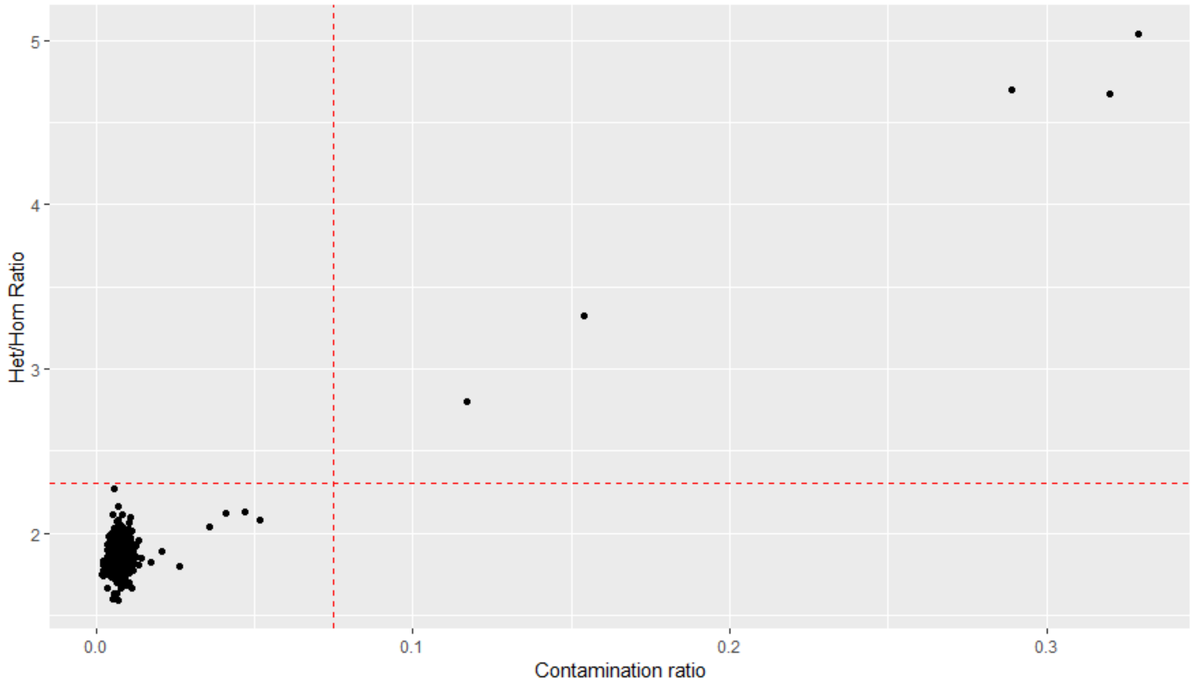


Figure 4.3: Detection of exome contamination using VerifyBamID. VerifyBamID contamination ratio (‘Freemix’) is plotted against the heterozygote/homozygote (Het/Hom) ratio from Hail. N=550 (1 sample repeat with very low quality failed to run in VerifyBamID).

4.3.3 Comparing imputed sex to our patient database

At this point, non-lymphoblastoid exomes were removed, and any replicates were similarly filtered. In cases where two lymphoblastoid-derived exome replicates passed QC, the exome with the highest QC metrics was kept. Sex was imputed using Peddy (Pedersen and Quinlan, 2017) in the 499 remaining exomes and compared to our Registry-HD patient database (Fig. 4.4). Four exomes were flagged as having different imputed sex compared to their recorded sex in our database. In all four cases, the imputed sex was male whilst recorded sex was female. In addition, one individual who had a missing sex in our database was imputed to be female. The four exomes with discrepancies between imputed sex and our clinical database were removed, and the exome whose sex was originally unknown was kept, leaving 495 of 500 exomes.

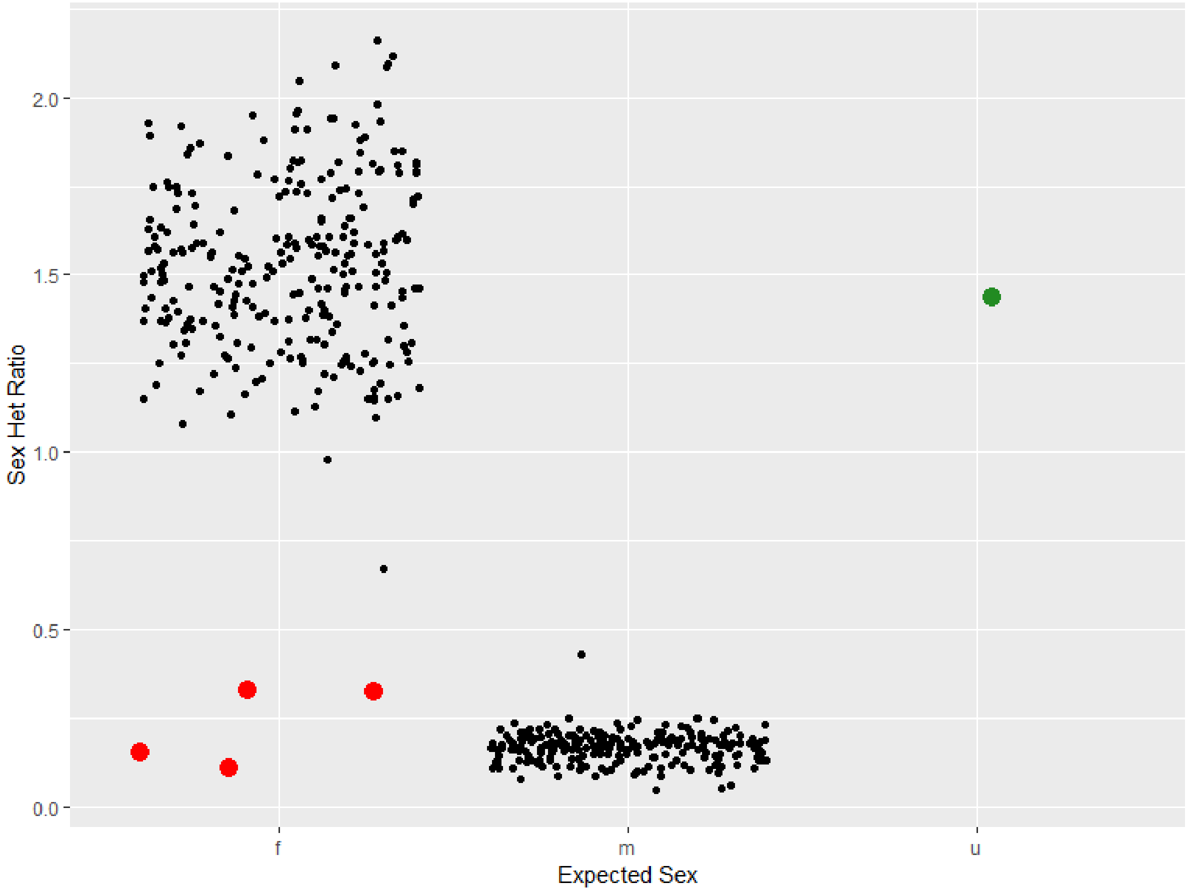


Figure 4.4: Sex checks using Peddy. The x-axis is recorded sex in Registry (f : female; m: male; u: unknown) and the y-axis is predicted sex imputed using Peddy. The four red circles are exomes which failed sex checks. The one green circle was a sample of originally unknown sex. Data are jittered on the x-axis to reduce overlap. Total N=499.

4.3.4 Principal component analysis and ancestry determination

A principal component analysis was used to determine population substructures. Population substructures (mostly through ancestry) are critical to account for as these can otherwise produce spurious gene associations (Price et al., 2010). The first 10 principal components were calculated using Hail, and the first three principal components are plotted in Fig. 4.5. Peddy was then used to estimate ancestry of samples using a principal components analysis against 2504 whole-genome samples from the thousand genomes project (1000 Genomes Project Consortium, 2015). As indicated in Fig. 4.6, 479 of the 495 exomes (~97%) came from individuals with a European ancestry as indicated by Peddy. Of those that had other ancestries, six were estimated to be ad mixed American (AMR), two were South Asian (SAS), one was African (AFR) and seven were of unknown descent (although at least three of these may be of European descent, but were not classified as such by Peddy – see Fig. 4.6).

No individuals were excluded based on ancestry, and Hail principal components were used in downstream analyses to adjust for population substructures (see 4.7-4.10). Hail principal components were preferred as Peddy only samples the VCF at ~25,000 sites whilst Hail utilises all available variant sites, at the cost of computing time. It is notable that principal components may not properly adjust for ancestry where there are only a few individuals of a given ancestry, as seen here. This is discussed in 4.11.2, however an auxiliary analysis recapitulating the main downstream burden/SKAT(-O) tests from 4.7-4.9 removing individuals with a non-European ancestry showed very similar results (Appendix 15) as having kept these individuals in the analysis.

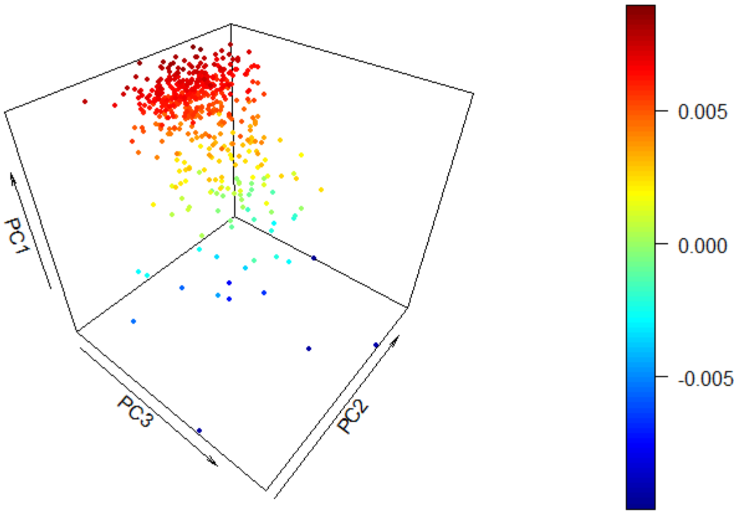


Figure 4.5: Hail-derived principal components. Principal components (PCs) are shown for the exomes, N=490 (5 individuals are excluded from this graph with very large PCs – see Fig. 4.6). Points are coloured according to the first principal component.

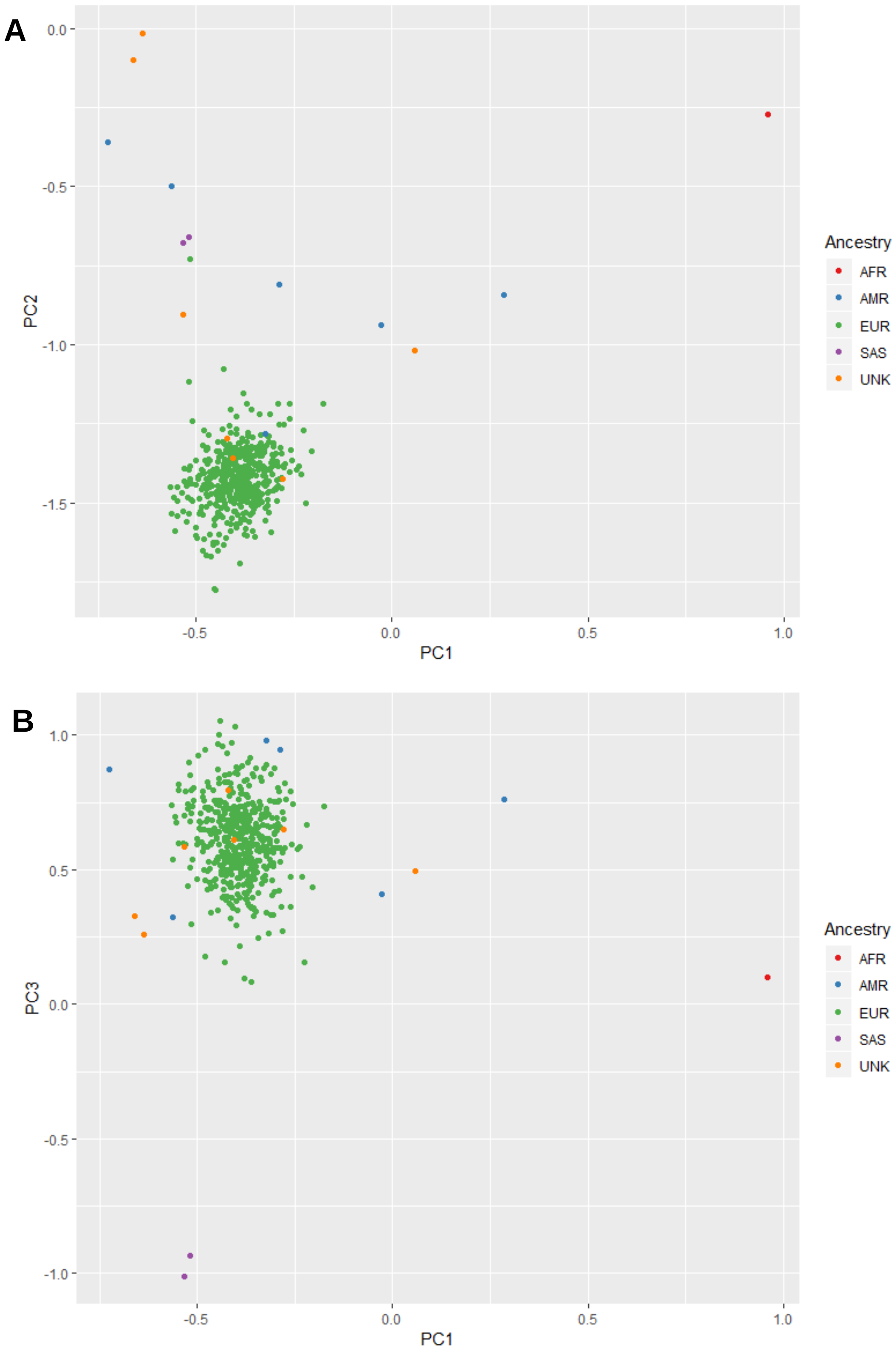


Figure 4.6: Ancestry of WES samples estimated by Peddy. Principal components (PCs) calculated using Peddy are plotted based on principal components from the 1000 genomes project (1000 Genomes Project Consortium, 2015). PC1 and PC2 are shown in (A); PC1 and PC3 are shown in (B). N=495. AFR: African; AMR: Ad mixed American; EUR: European; SAS: South Asian; UNK: Unknown.

4.3.5 Removal of related individuals

We next removed highly related individuals from our analysis using PLINK ((Purcell et al., 2007; Chang et al., 2015); www.cog-genomics.org/plink/1.9/) identity by descent (IBD) ratios. We used a PI-HAT cut-off of 0.5 to identify first-degree relatives, and identified nine pairs of highly related individuals (Fig. 4.7). For these related pairs, we kept the individual with the highest age at motor onset residual, leaving 486 exomes. Notably, in all but one of these nine cases, early onset individuals segregated with other related early onset individuals and *vice versa*. The only exception identified was a mother-daughter pair where both had the same CAG length (41 measured by MiSeq), but the daughter had a 15 years earlier onset than the mother. Using Hail's genetic relationship matrix (GRM) identified the same nine related pairs (Fig. 4.8).

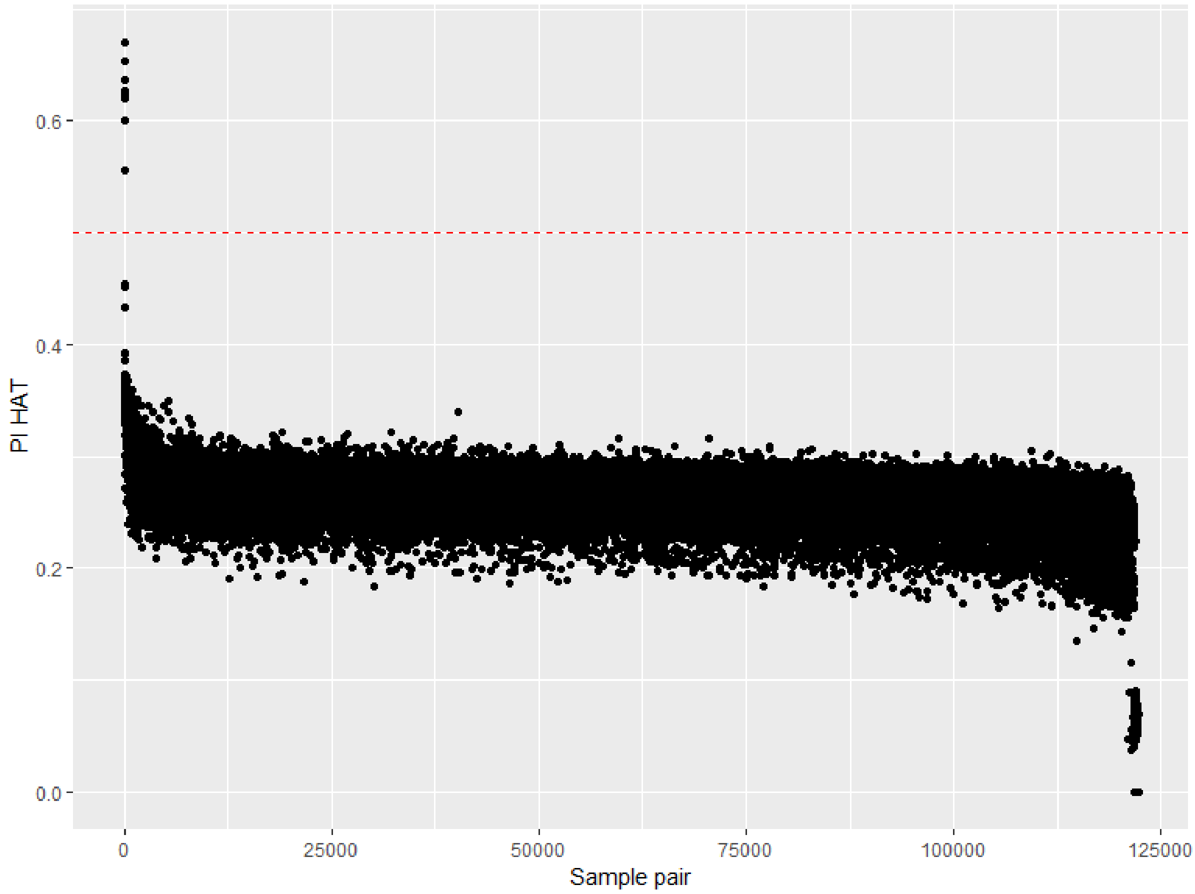


Figure 4.7: Identity by decent ratios for WES samples. VDS files were exported to PLINK, which was then used to generate identity by decent (IBD) ratios for N=495 individuals. The graph displays IBD PI-HAT results for 122,265 individual pairs in order of relatedness (x axis as sample pair). We used a cut-off of PI HAT = 0.5.

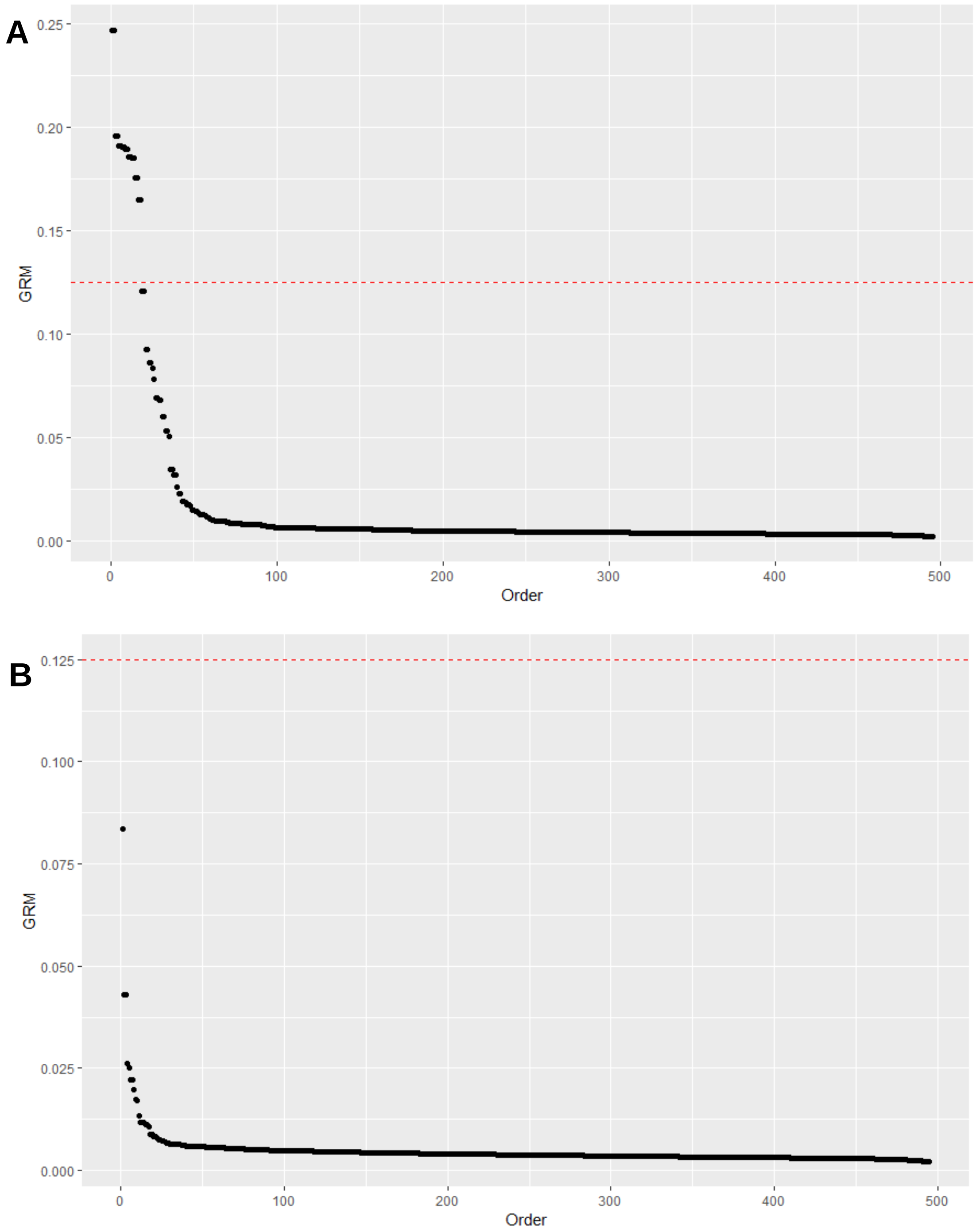


Figure 4.8: Genetic relationship matrices for WES samples. Genetic relationship matrices (GRMs) were calculated using Hail. (A) shows the most related pairs of individuals and (B) the second most related pair. Cut-off was chosen as 0.125, indicated on both plots as the dashed red line. ‘Order’ on the x-axis indicates the order in which samples are arranged (from highest to lowest GRM).

4.3.6 Redefining early and late populations post re-genotyping

Although HD individuals were originally chosen for WES based on their age at motor onset (AMO) residual, centrally measured CAG lengths were unavailable for many patients at the time of sample selection. As a result, CAGs derived from local clinical labs were instead used to calculate expected ages of onset in chapter 3. CAG lengths from local labs are generally considered less accurate as these are not (1) systematic or (2) subject to the same level of QC as centrally measured CAG lengths. To address this issue, we re-genotyped all 500 individuals chosen for sequencing using a targeted MiSeq methodology (2.8/2.9 & (Ciosi et al., 2018)) – see chapter 5 for these results in full. One individual who failed MiSeq and in-house genescan genotyping (2.11) was excluded, leaving a final 485 exomes passing QC (see Fig. 4.9 for an overview in flowchart format).

MiSeq data (see 2.8-2.9 and chapter 5) was used to calculate two AMO residuals. The uncorrected residual derived an expected age at onset using an individual's total polyglutamine length-2, to be equivalent to non-sequencing CAG sizing methods (*e.g.* genescan). Typically, non-sequencing *HTT* sizing methods assume a canonical **CAA**CAG at the end of the *HTT* CAG repeat tract (a canonical *HTT* allele is shown in Fig. 4.12, and has a single penultimate **CAA** codon at the 3' end of the repeat). The corrected residual instead used pure CAG length to estimate expected onset.

We redefined our early and late onset groups where the uncorrected AMO residual was ≥5 years earlier or later than expected using the Langbehn model, respectively (Fig. 4.10), as ~5 years is equivalent to ~±2 CAGs (Langbehn et al., 2004), and >±1 CAG error accepted for most sizing methods (Massey et al., 2018). A subset (N=44, ~9%) of the original 500 were found to have uncorrected AMO residuals between -5 to 5 years after recalculation using MiSeq polyglutamine length-2, and these individuals were reclassified as having a 'normal' or expected HD onset. With these data in mind, we created two groups for analyses. Group 1, the continuous phenotype group, contained all 485 QC-passing exomes (N=243 early, 242 late), and was used in continuous whole-exome analyses. Group 2, the dichotomous/binary group, contained 440 samples who both passed QC and had ≥5 |AMO residual| (N=225 early, 215 late) (see also Fig. 4.9). One originally late individual had an early onset upon re-genotyping and was excluded from group 2. The dichotomous group was used for determining variant counts and in whole-exome logistic analyses.

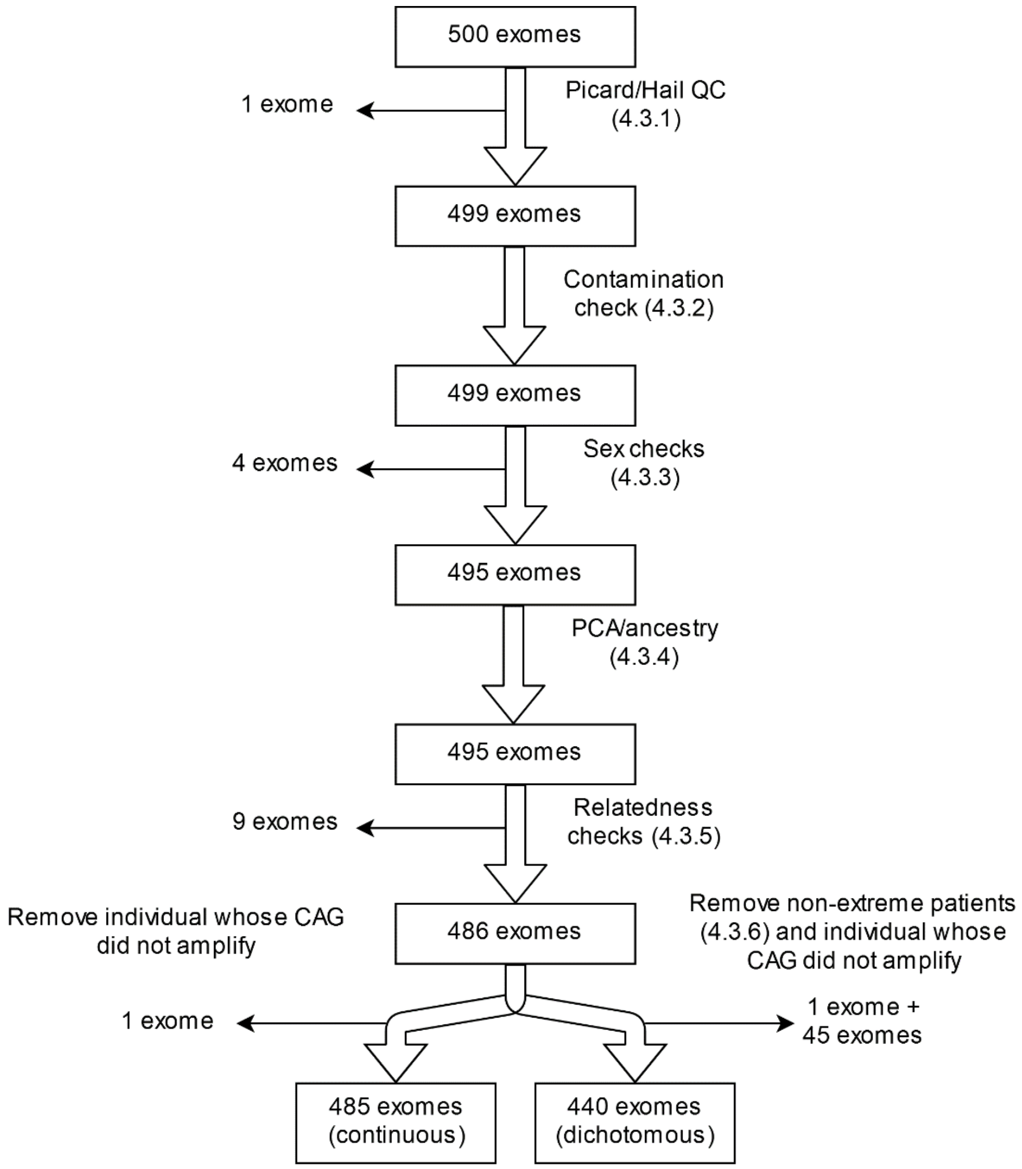


Figure 4.9: A flowchart of the exomes kept/lost in the quality control pipeline. Shown are the steps in the exome quality control pipeline (4.3); also see Fig. 2.5. Abbreviations: QC: Quality control; PCA: Principal components analysis.

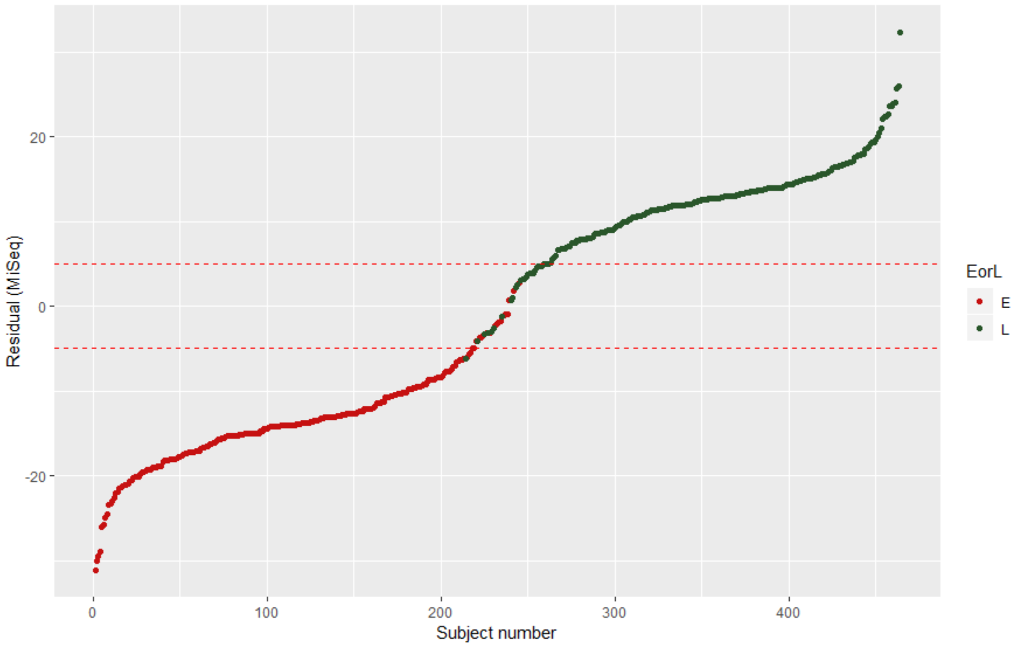


Figure 4.10: Redefining early and late populations with MiSeq genotypes. The plot uses uncorrected residual age at motor onset (polyglutamine length-2 in re-genotyped *HTT* alleles) as calculated from MiSeq CAG lengths (see methods 2.8-2.9). The dashed red lines distinguish the redefined early (≤-5 year AMO residual) and the late onset groups (≥5 year AMO residual). E: Early; L: Late.

4.4 Annotation pipeline and variant prioritisation

A flowchart for the annotation pipeline is available in Fig. 2.6 in methods 2.7.1.3. The 485 exomes passing QC from 4.2 were annotated using an in-house annotation pipeline. Multi-allelic sites were first split into separate calls, and individual variant sites were subject to QC for each sample. Variants where quality was low (<10 reads or <30 GQ) were excluded, and sites of ambiguous homo/heterozygosity were similarly filtered. Variants were then annotated using the variant effect predictor tool (VEP) (McLaren et al., 2016). Several other databases, including gnomAD (Karczewski et al., 2019) and dbNSFP (Liu et al., 2011, 2016), were additionally used to annotate variants.

An important concept in interpretation of sequencing data is that of variant prioritisation. For our purposes, we focused on three variant features. (1) Whether a variant was predicted to have a tangible functional effect (*i.e.* coding change), (2) if the variant was predicted to have a damaging effect, and (3), how rare the variant was. Taking (1) and (2) into account, we defined non-synonymous damaging (NSD) as variants that either resulted in a loss-of-function (LoF) change, such as a frameshift or loss of a splice acceptor, or a missense variant. For missense variants, we considered those with a CADD PHRED score ≥20 to be damaging. CADD score is a prediction of how damaging a particular variant is (Kircher et al., 2014), and PHRED is a logarithmic scale (Ewing and Green, 1998; Ewing et al., 1998). CADD PHRED of ≥20 represents the top 1% predicted most damaging variants in the genome.

For (3), we defined rare variants as those which were either missing from gnomAD (and may represent very rare, or private, singleton variants), or variants with a known minor allele frequency (MAF) ≤1% using gnomAD's non-Finnish European population. The non-Finnish European population was chosen as ~97% of our sequenced population were of European descent and is gnomAD's largest single sample (>50,000 exomes). This frequency cut-off for defining rare variants is similar to those implemented and suggested elsewhere (Jalali Sefid Dashti and Gamieldien, 2017; Retshabile et al., 2018; Flannick et al., 2019).

In our dichotomous early/late onset cohort that passed QC (N=440), 335,435 variants of all classes were present in at least one sample passing variant-level QC. 150,188 of these were non-synonymous (NS) changes, and slightly less than half (67,335) of these were considered potentially damaging (*i.e.* NSD) using CADD PHRED ≥20. 47,666 of these NSD variants had known gnomAD MAFs ≤1% and 34,416 had known gnomAD MAFs ≤0.1%. Of the NSD variants, 9,732 were LoF variants, and 5,661 of these had a known gnomAD MAF

≤1%. Of the 17,087 NS variants without gnomAD MAFs, 13,998 were missense variants and 2,273 were LoF variants. Variants were annotated to a total of 21,860 genes and open reading frames (ORFs), and 13,970 genes/ORFs had at least one NSD variant. Fig 4.11 shows the frequency of NSD variants across all exomes in the dichotomous population in annotated genes (21,860 genes).

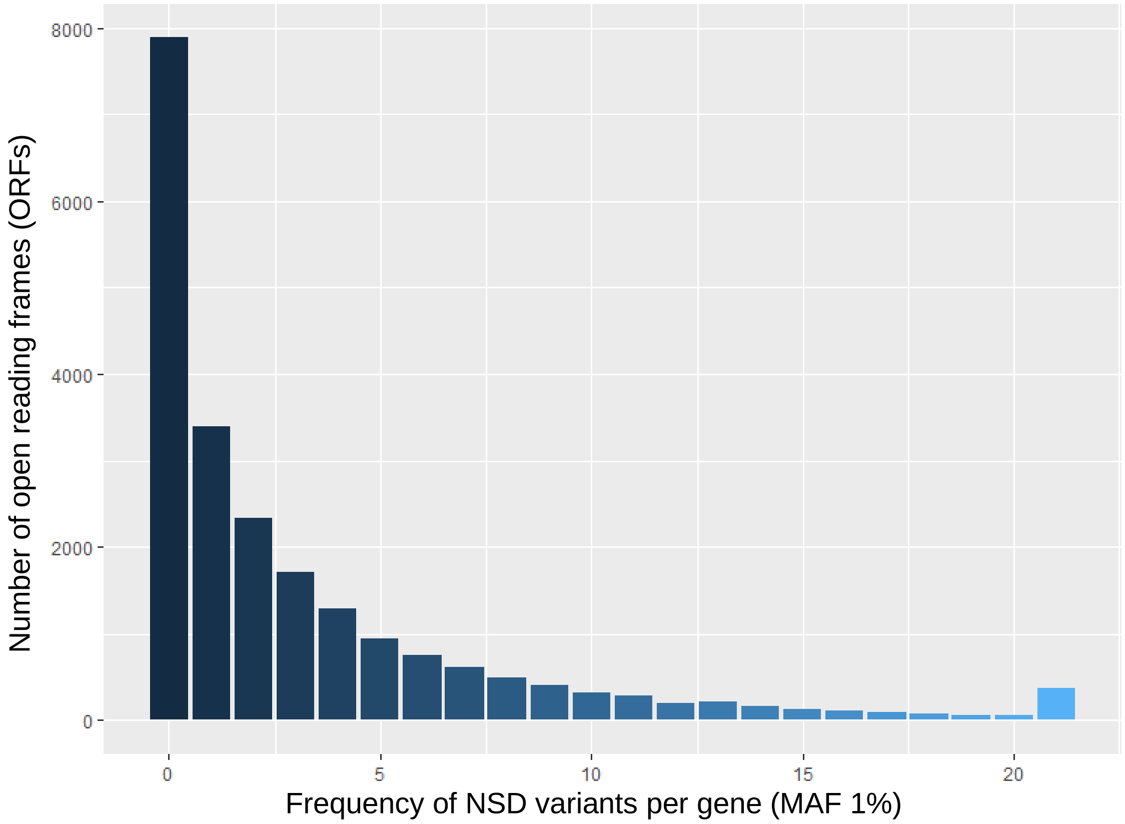


Figure 4.11: Non-synonymous damaging (NSD) variants across the exome. NSD variants (CADD ≥20 and either (1) minor allele frequency (MAF) ≤1% gnomAD or (2) not present in gnomAD) passing QC filters were collapsed across all annotated genes and ORFs (21,860 genes) and are shown between 0-20 and >20 NSD variants per gene. N=440 exomes (225 early, 215 late).

4.5 Calling *HTT* CAG structure with WES

Initially, we wanted to investigate *HTT* sequence from our exomes to examine whether cis variation in *HTT* was a modifier of HD onset. However, as read length from WES was only 75 bp, complete read-through of expanded *HTT* CAG short tandem repeat (STR) sequences (~120-160 bp) was not possible. Furthermore, the Genome Analysis Toolkit (GATK) variant calling pipeline performed poorly in identification of variants in the CAG repeat, both in expanded and wild-type alleles. To overcome this limitation, reads from the first exon of *HTT* were extracted from BAM files using SAMtools as detailed (2.7.2.2). Although our reads were not large enough to span expanded *HTT* alleles, it was possible to (1) read-through most wild-type CAGs up to ~21 CAGs (63 bp) and (2) examine the 5' and 3' structures of both wild-type and expanded alleles. Therefore, where exome read depth was high enough, we could dephase the wild-type allele structure and thus infer the 5' and 3' structure of *HTT* on each chromosome. The results from this analysis are shown in Table 4.2.

The canonical structure of the *HTT* CAG region is shown in Fig. 4.12, which contains a single penultimate CAA interruption at the 3' end of the CAG repeat (CAA also encodes glutamine). We found that the most common alternative allele is the CAG(CAG**CAA**)$_2$CAG structure, and this is found in both early and late populations. Crucially, however, this structure was only found on the expanded allele in late onset participants. A Fisher's exact test on dephased expanded alleles comparing atypical CAG(CAG**CAA**)$_2$CAG and canonical structures between early and late onset individuals shows statistical significance (*p*=8.96E-04) (183, 150; 0, 9). Furthermore, we also found two more alternative CAG structures containing additional interruptions, CAG(**CAA**)$_2$CAG and CAG(**CAA**)$_3$CAG in three late onset individuals. These were dephased in two cases, both found on the expanded allele. By contrast, we also identify a further allele with no CAA interrupting structure in seven individuals. Unlike CAG(CAG**CAA**)$_2$CAG, the pure CAG allele only appears in early onset individuals on the expanded allele where dephasing was possible. A Fisher's exact test between dephased canonical and pure CAG alleles in early and late onset individuals shows modest statistical significance (*p*=3.61E-02) (183, 150; 6, 0).

As indicated, however, allele structure could only be determined in ~80% of cases, and of those with atypical alleles, dephasing was possible only ~65% of the time. This is due to the high variability of *HTT* CAG repeat read depth, ranging from >60 reads and, in other cases, <5 reads. Furthermore, manual assessment of BAM reads was slow and potentially prone to human error. A more systematic and targeted NGS-based technique (MiSeq) and bioinformatic analysis (Scale-HD) is explored in much greater detail in chapter 5.

	Early (N=225)			Late (N=215)		
	EXP	WT	UNK	EXP	WT	UNK
(CAG)$_n$**CAA**CAGCCG	183	185	N/A	150	157	N/A
(CAG)$_n$CCG	6	0	1	0	0	0
(CAG)$_n$(**CAA**CAG)$_2$CCG	0	4	2	9	4	7
(CAG)$_n$(**CAA**)$_2$CAGCCG	0	0	0	1	0	1
(CAG)$_n$(**CAA**)$_3$CAGCCG	0	0	0	1	0	0
Unknown allele structures	33	33	N/A	46	46	N/A
Total	222	222	3 (6)	207	207	8 (16)

Table 4.2: Atypical *HTT* allele structures identified by WES. The numbers of atypical *HTT* CAG structures are shown with interruptions (CAA) emboldened. The EXP (expanded), WT (wild type) and UNK (unknown) indicate the dephasing of atypical repeat structures. UNK indicates the structure was unable to be dephased. N=440 (225 early, 215 late). Note that the 'Total' row does not add up to the exact number of alleles as would be expected in EXP/WT as the unknown alleles are capturing two alleles (shown in the brackets). Adding up the UNK bracketed number and the EXP/WT numbers gives the number of alleles as would be expected (*i.e.* 450 early, 430 late).

5'-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_7$-3'

Figure 4.12: The canonical *HTT* CAG repeat structure. Shown is the canonical structure of polyCAG in the first exon of *HTT*. The sequence in blue is the CAG repeat region (encoding polyglutamine), with the interrupting CAA triplet codon emboldened. The green sequence is CCG repeat region (encoding polyproline), with the interrupting CCA triplet emboldened. The first 16 codons of *HTT* are not shown.

4.6 Candidate gene analysis

4.6.1 Selection of candidate genes

We were next interested whether there was an excess of damaging variants observed in genes previously associated with either HD or other repeat disease. We chose a group of 13 disease-associated candidate genes whose exonic variants would be explored in more detail in our patient cohort. Nine candidate genes were genes from loci implicated by the most recent HD onset GWAS (GeM-HD Consortium, 2019): *FAN1*, *MSH3*, *LIG1*, *TCERG1*, *MLH1*, *PMS1*, *PMS2*, *SYT9* and *RRM2B*. We also included *HTT* itself as a candidate, and *MLH3*, *OGG1* and *EXO1*, all of which have been implicated in varying degrees by functional study in repeat disease systems (Pinto et al., 2013; Budworth et al., 2015; Zhao et al., 2018).

4.6.2 *FAN1*

Four independent signals at the *FAN1* locus were significant in the most recent HD GWAS (GeM-HD Consortium, 2019). We were therefore very interested in determining whether there were *FAN1* NSD variants associated with extreme onset as these may inform mechanism. A full list of all coding variants are shown in Table 4.3 (for non-coding variants, see Appendix 5). 14 distinct NSD variants were identified. Of the 35 individual instances of these variants, 26 were in early onset individuals and 9 in late individuals (note that these were found in 24 early and 8 late individuals, respectively, as two early individuals had two damaging *FAN1* variants each, and one late individual had two damaging *FAN1* variants). This enrichment is significant with a Fisher's exact test (*p*=5.54E-03) (201, 207; 24, 8). The NSD variants are plotted against a schematic of FAN1 structure in Fig. 4.13.

Two primary groups of early-associated damaging variation emerge. A cluster of variants positioned centrally encompass the Arg507His, Arg507Cys and Asp498Asn mutations in the SAF-A/B, Acinus and PIAS (SAP) domain of FAN1. The second cluster of early-onset associated variation is in the virus-type replication repair nuclease (VRR-Nuc) domain at the C-terminal end of FAN1 which includes the Val963Trp964ins, Arg982Cys and Cys1004Gly variants. The Arg969Leu, although not NSD by our classification (CADD 19.5), was also found here. A smaller sub-cluster of NSD variants, Arg377Trp and Leu395Pro, was additionally observed. Notably, both the Arg507His and Arg377Trp variants can appear in both early and late onset individuals but are highly enriched in early onset (12:4 and 6:1, respectively). We also find LoF frameshift variant, Thr187fs, in one early onset patient. The variant has a very low MAF in gnomAD (8.97E-06).

Further, we also identify a third cluster of late onset associated variants occurring in the tetratricopeptide (TPR) domain of FAN1: Arg658Trp, Asp702Glu and Lys794Arg. All three variants are rare singleton variants. Notably, however, the Pro654Leu variant, found in a single early onset individual, is also found in the same domain. Finally, although not a NSD variant, we do find a non-coding putative splice variant, 15:31212744:T:C, present only in four late onset individuals (Appendix 5). To our knowledge, the Asp498Asn and Asp702Glu NSD variants are novel and have not been described before (gnomAD, accessed July 2019).

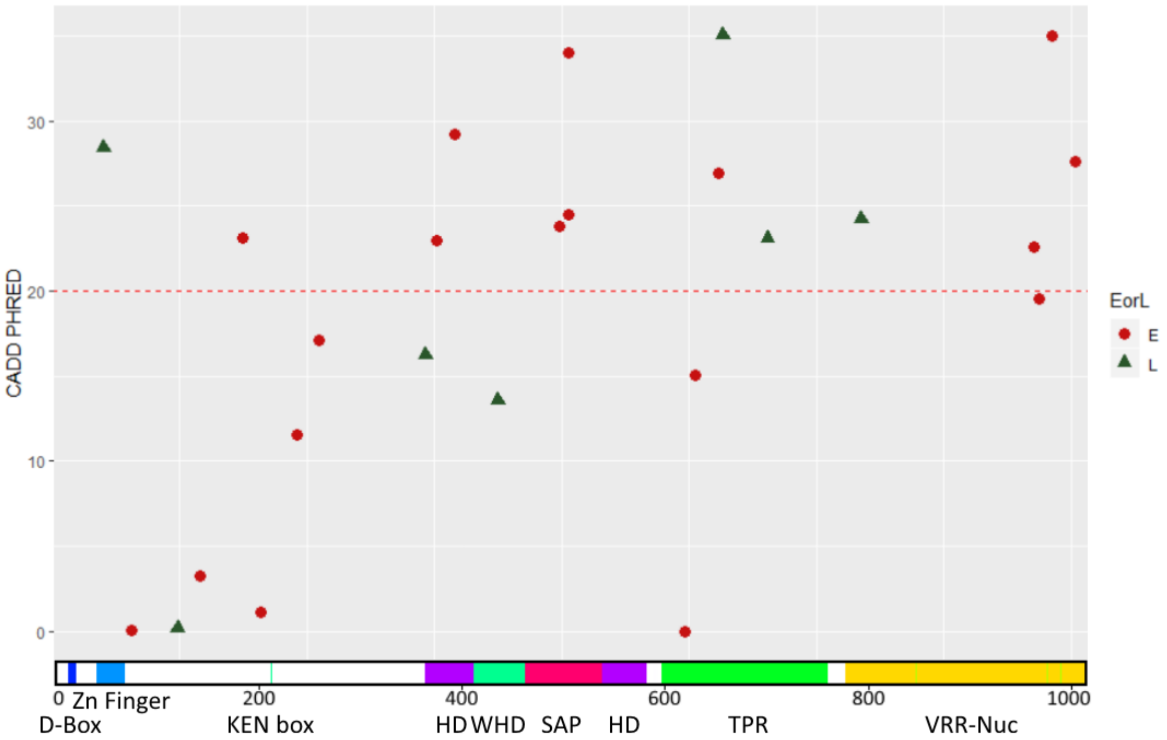


Figure 4.13: Structural overview of identified *FAN1* variants. NS variants MAF≤1% are shown plotted against a schematic of the FAN1 protein. The dashed line indicates the CADD 20 cut-off. The unmarked cyan bars indicate Mg^{++} binding sites. Numbers of each mutation are not indicated. Red circles represent variants more associated with early onset (E); green triangles are variants more associated with late onset (L). Zn finger: zinc finger; HD: Helical domain; WHD: winged-helix domain; SAP: SAF-A/B, Acinus and PIAS; TPR: tetratricopeptide repeat; VRR-Nuc: virus-type replication repair nuclease. N=225 early, 215 late HD patients. Domain boundaries are taken from UniProt (UniProt Consortium, 2019) and (Jin and Cho, 2017).

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Met50Arg	**15:31197015:T:G**	**15.72**	**2.70E-03**	**28.4**	**123**	**102**	**0**	**0**	**130**	**84**	**1**	**0**
Val77Ile	15:31197095:G:A	17.95	8.95E-06	0.1	38	186	1	0	36	179	0	0
Gln123Arg	15:31197234:A:G	32.64	0.00E+00	0.2	1	224	0	0	0	214	1	0
Arg145His	15:31197300:G:A	34.95	2.19E-03	3.3	0	223	2	0	0	215	0	0
Thr187fs [*]	**15:31197423:CCA:C**	**18.76**	**8.97E-06**	**N/A**	**30**	**194**	**1**	**0**	**38**	**177**	**0**	**0**
Gln204Arg	15:31197477:A:G	25.51	8.97E-06	1.2	5	219	1	0	3	212	0	0
Gly233Glu	15:31197564:G:A	24.08	4.28E-01	0.0	11	71	111	32	8	86	92	29
Glu240Lys	15:31197584:G:A	24.81	5.19E-03	11.5	3	220	2	0	5	209	1	0
Ala261Val	15:31197648:C:T	35.61	2.69E-05	17.1	0	224	1	0	0	215	0	0
Pro366Arg	15:31197963:C:G	35.70	NA	16.2	0	225	0	0	0	214	1	0
Arg377Trp	**15:31197995:C:T**	**31.85**	**7.21E-03**	**23.0**	**0**	**219**	**6**	**0**	**0**	**214**	**1**	**0**
Leu395Pro	**15:31198050:T:C**	**16.17**	**1.72E-04**	**29.2**	**50**	**174**	**1**	**0**	**67**	**148**	**0**	**0**
Glu437Gly	15:31200396:A:G	27.23	7.08E-04	13.6	8	216	1	0	6	207	2	0
Asp498Asn	**15:31202933:G:A**	**27.89**	**NA**	**23.8**	**2**	**222**	**1**	**0**	**0**	**215**	**0**	**0**
Arg507Cys	**15:31202960:C:T**	**25.58**	**0.00E+00**	**34.0**	**6**	**218**	**1**	**0**	**8**	**207**	**0**	**0**
Arg507His	**15:31202961:G:A**	**25.57**	**9.64E-03**	**24.5**	**7**	**206**	**12**	**0**	**8**	**203**	**4**	**0**
Asn621Ser	15:31210417:A:G	31.70	7.16E-05	0.0	0	224	1	0	1	214	0	0
Ala631Thr	15:31210446:G:A	34.33	2.69E-05	15.0	0	224	1	0	0	215	0	0
Pro654Leu	**15:31212765:C:T**	**31.38**	**3.78E-04**	**26.9**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg658Trp	**15:31212776:C:T**	**32.57**	**6.29E-05**	**35.0**	**1**	**224**	**0**	**0**	**0**	**214**	**1**	**0**
Asp702Glu	**15:31214491:C:A**	**30.96**	**NA**	**23.1**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Lys794Arg	**15:31218035:A:G**	**35.69**	**8.33E-04**	**24.2**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Pro894Ser	15:31221493:C:T	27.69	1.71E-02	0.0	8	209	8	0	14	197	4	0
Val963_Trp964insLeu	**15:31222845:G:GTGT**	**36.32**	**0.00E+00**	**22.6**†	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg969Leu	15:31222864:G:T	35.29	NA	19.5	0	224	1	0	2	213	0	0
Arg982Cys	**15:31229349:C:T**	**19.21**	**8.96E-06**	**35.0**	**47**	**177**	**1**	**0**	**50**	**165**	**0**	**0**
Cys1004Gly	**15:31229415:T:G**	**27.34**	**NA**	**27.6**	**2**	**222**	**1**	**0**	**2**	**213**	**0**	**0**

Table 4.3: Coding variation from *FAN1* from WES. Non-synonymous damaging (NSD) variants (NS CADD≥20 or LoF, and MAF≤1% in gnomAD NFE) are emboldened. Under gnomAD, 'NA' denotes variants not found in gnomAD v2.0.2, and 0.00E+00 indicates variants found in gnomAD but not in non-Finnish Europeans. Loss of function (LoF) are marked by [*]. Genomic locations are based on hg19/GRCh37 and CADD PHRED scores from dbNSFP 3.0 unless otherwise marked with a †; these were missing and were estimated using https://cadd.gs.washington.edu/score (accessed April 2019). Total N=440 (225 early; 215 late). DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

To verify the robustness of calling *FAN1* variants in WES, and to assess how well the sample quality control and annotation pipelines in sections 4.3 and 4.4 performed, Sanger sequencing was performed to confirm *FAN1* variants. 18 variants were Sanger sequenced, including all 14 NSD variants. All variants passing WES QC were confirmed by Sanger (Table 4.4; see Appendix 4 for representative Sanger sequencing traces). Notably, a variant originally called by WES, Gln717Arg, that failed subsequent exome QC was not confirmed by Sanger sequencing. In three cases, WES QC removed variants in individuals that were confirmed by Sanger sequencing. Hence, the stringency of QC may occasionally result in false negatives, but no instances of false positives were observed at our level of QC stringency in the WES in *FAN1*.

	Early			Late		
	Pre-QC	Post-QC	Sanger	Pre-QC	Post-QC	Sanger
Met50Arg	0	0	0	1	1	1
Val77Ile	1	1	1	0	0	0
Thr187fs	1	1	1	0	0	0
Pro366Arg	0	0	0	1	1	1
Arg377Trp	6	6	6	1	1	1
Leu395Pro	**2**	**1**	**2**	**0**	**0**	**0**
Asp498Asn	1	1	1	0	0	0
Arg507Cys	1	1	1	0	0	0
Arg507His	**14**	**12**	**14**	**4**	**4**	**4**
Pro654Leu	1	1	1	0	0	0
Arg658Trp	0	0	0	1	1	1
Asp702Glu	0	0	0	1	1	1
Gln717Arg	**1**	**0**	**0**	**0**	**0**	**0**
Lys794Arg	0	0	0	1	1	1
Val963_Trp964insLeu	1	1	1	0	0	0
Arg969Leu	1	1	1	0	0	0
Arg982Cys	1	1	1	0	0	0
Cys1004Gly	1	1	1	0	0	0

Table 4.4: Sanger sequencing confirmation of *FAN1* variants. Samples with *FAN1* variants identified by WES were sequenced using Sanger sequencing. The 'pre-QC' column gives variant counts in the raw VCFs called by GATK preceding per-variant quality control in the annotation pipeline (2.7.1.3 & 4.4). The 'post-QC' column refers to variant counts after per-variant QC, which removes low quality variant calls. 'Sanger' refers to variant counts determined by Sanger sequencing (Sanger sequencing was performed for all variants in the 'Pre-QC' column). Relevant variants are emboldened. Representative Sanger sequencing traces are available in Appendix 4. Variant calls which differ in at least one stage are emboldened for visibility. N=485 (continuous group).

4.6.3 *EXO1*

The endonuclease 1 gene, *EXO1*, was not originally implicated as significant *via* GWAS (GeM-HD Consortium, 2015, 2019), though it showed modest significance in a subsequent TWAS (GeM-HD Consortium, 2019). Further, it has been implicated as a modifier in a fragile X mouse model (Zhao et al., 2018). Hence, we were interested in seeing whether a similar pattern of NSD variant was also present in *EXO1*. Coding *EXO1* variants identified in WES are presented in Table 4.5 (see Appendix 5 for non-coding variants), with 13 early variants and 25 late variants (in 12 and 24 early/late patients, respectively). The 11 NSD variants found in *EXO1* are plotted in Fig. 4.14. The enrichment of late-associated damaging variation in people is approaching significance (*p*=5.69E-02, Fisher's exact test) (203, 201; 12, 24).

Strikingly, a cluster of late-associated variation appears around the N-terminal MSH3-EXO1 interaction domain (Schmutte et al., 2001). The signal here is driven by five late-associated variants, all of which very rare and predicted to be quite damaging, ranging between 25.2-34.0 CADD PHRED. The Gly274Arg variant may be of particular interest as it appears in 4 individuals. Two further variants, Gly759Glu and Lys790Arg, occur near the C-terminal end of the protein which are also associated with late onset. Interestingly, this is near the MLH1-EXO1 interaction domain of the EXO1 protein. The Gly759Glu variant here is more common (9.34E-03, gnomAD MAF) and identified in 9 late and 2 early onset individuals.

Both the Asp249Asn and Ala827Val variants, found in the MSH3-EXO1 and MLH1-EXO1 domains, respectively, occurred in equal numbers of early and late patients. Two early-associated NSD variants were observed: Ser610Gly and Arg121Trp. Arg121Trp is a singleton whereas Ser610Gly occurs in one late and three early individuals. As there are few early-associated variants, it is difficult to draw firm conclusions. Unlike the MSH3-EXO1 and (potentially) the MLH1-EXO1 interaction domains, the MSH2-EXO1 interaction domain only contains one variant (Ser610Gly). Finally, a LoF (a splice acceptor variant) was observed in equal numbers of early and late individuals (1:242048615:G:C). This is found at a modestly low gnomAD MAF (2.63E-03).

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Arg108His	**1:242016701:G:A**	**36.89**	**3.58E-05**	**33.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Arg121Trp	**1:242016739:C:T**	**37.42**	**1.25E-04**	**32.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ala137Ser	**1:242020650:G:T**	**22.19**	**1.30E-03**	**25.7**	**2**	**223**	**0**	**0**	**4**	**210**	**1**	**0**
Asp143Glu	**1:242020670:T:A**	**24.09**	**8.24E-04**	**25.2**	**5**	**220**	**0**	**0**	**2**	**212**	**1**	**0**
Gly223Val	**1:242021932:G:T**	**36.49**	**2.69E-05**	**31.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Asp249Asn	**1:242022009:G:A**	**30.89**	**4.45E-03**	**26.7**	**1**	**220**	**4**	**0**	**1**	**211**	**3**	**0**
Gly274Arg	**1:242023882:G:A**	**35.06**	**3.10E-03**	**34.0**	**1**	**224**	**0**	**0**	**0**	**211**	**4**	**0**
Asn279Ser	1:242023898:A:G	37.67	3.18E-02	26.9	0	212	11	2	0	198	15	2
His354Arg	1:242030151:A:G	45.63	5.56E-01	0.0	1	48	91	85	2	51	103	59
Thr439Met	1:242035382:C:T	22.42	8.23E-02	20.2	10	182	32	1	13	180	21	1
Val458Met	1:242035438:G:A	38.13	2.42E-01	0.0	2	134	80	9	2	126	78	9
Val460Leu	1:242035444:G:C	38.03	1.03E-02	2.3	0	223	2	0	0	211	3	1
Asn469Asp	1:242035471:A:G	38.57	NA	0.1	0	224	1	0	1	214	0	0
Glu589Lys	1:242042301:G:A	23.36	3.77E-01	0.7	11	69	102	43	12	70	101	32
Ser610Gly	**1:242042364:A:G**	**19.77**	**3.82E-03**	**21.3**	**19**	**203**	**3**	**0**	**23**	**191**	**1**	**0**
Arg634Gln	1:242042437:G:A	12.24	2.14E-03	5.0	194	30	1	0	200	15	0	0
Pro640Ser	1:242042454:C:T	12.62	6.41E-03	7.4	191	33	1	0	199	15	1	0
Glu670Gly	1:242042545:A:G	11.90	6.35E-01	0.2	174	10	16	25	177	3	20	15
Asn688Ser	1:242042599:A:G	16.29	NA	0.0	57	168	0	0	58	156	1	0
Arg723Cys	1:242045275:C:T	25.84	9.76E-01	22.6	6	0	10	209	8	0	9	198
Splice acceptor c.2212-1G>C [*]	**1:242048615:G:C**	**35.84**	**2.63E-03**	**24.2**	**0**	**223**	**2**	**0**	**0**	**213**	**2**	**0**
Pro757Leu	1:242048674:C:T	38.55	1.57E-01	29.3	2	153	61	9	0	151	59	5
Gly759Glu	**1:242048680:G:A**	**37.11**	**9.34E-03**	**22.0**	**1**	**222**	**2**	**0**	**0**	**206**	**9**	**0**
Lys790Arg	**1:242048773:A:G**	**38.14**	**NA**	**24.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Ala827Val	**1:242052841:C:T**	**36.85**	**5.65E-04**	**24.4**	**0**	**224**	**1**	**0**	**0**	**214**	**1**	**0**

Table 4.5: Coding variation from *EXO1* from WES. Non-synonymous damaging (NSD) variants (NS CADD≥20 or LoF, and MAF≤1% in gnomAD NFE) are emboldened. Under GnomAD, 'NA' denotes variants not found in gnomAD v2.0.2. Loss of function (LoF) are marked by [*]. Genomic locations are based on hg19/GRCh37 and CADD is from dbNSFP 3.0. Total N=440 (225 early; 215 late). DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

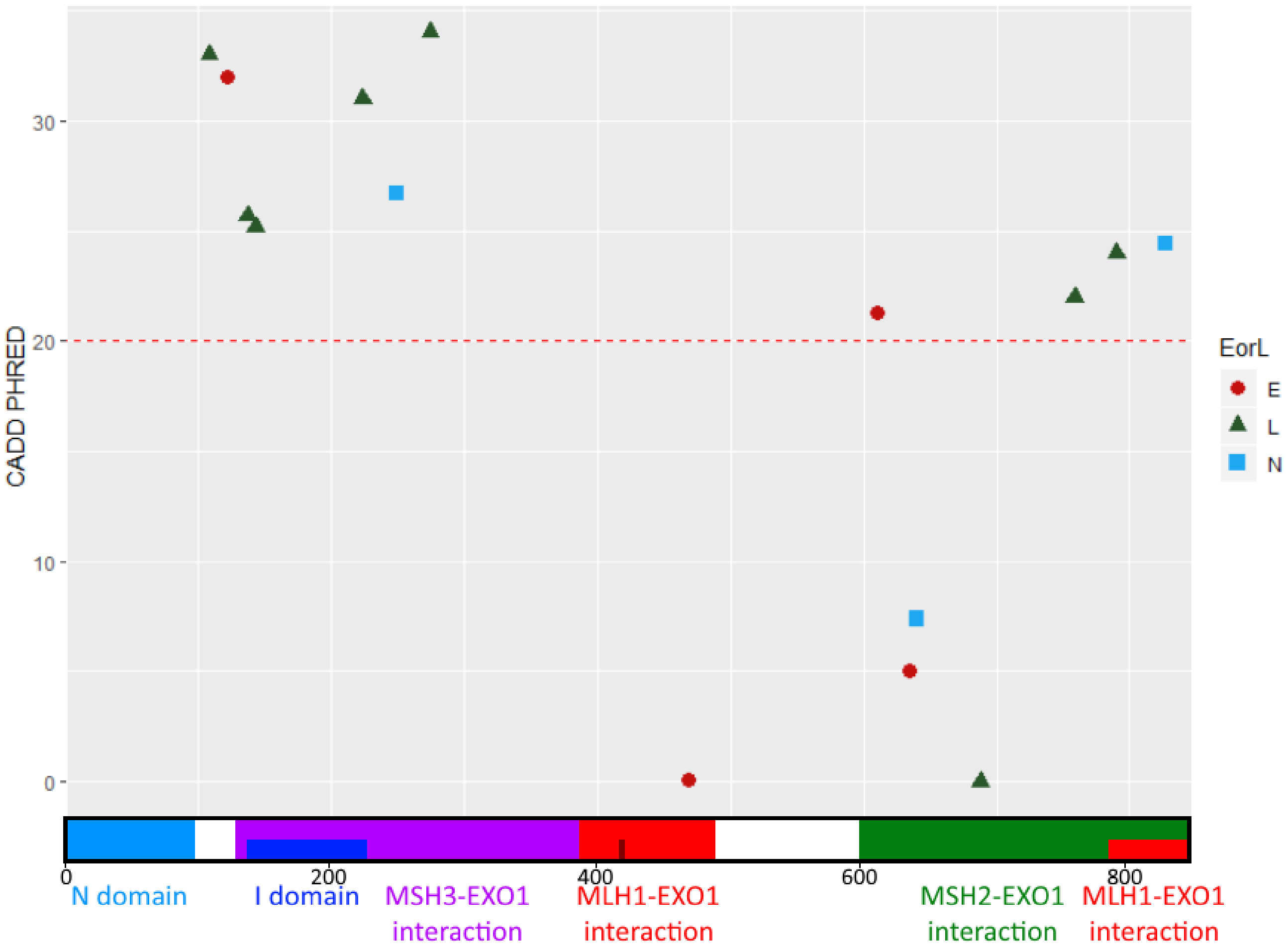


Figure 4.14: Structural overview of identified *EXO1* variants. Variants with a MAF≤1% are plotted against a schematic of the EXO1 protein. The dashed line indicates the CADD 20 cut-off. The unmarked maroon box in the first MLH1-EXO1 interaction domain refers to the nuclear localisation signal. Numbers of each mutation are not indicated. Red circles represent variants more associated with early onset (E); green triangles are variants more associated with late onset (L); blue squares are variants that are not skewed in either direction (N). N=225 early, 215 late HD patients. Domains boundaries taken from UniProt (UniProt Consortium, 2019) and (Schmutte et al., 2001).

4.6.4 MSH3

MSH3 is a mismatch repair protein, and *MSH3* has been implicated as a modifier of HD onset (GeM-HD Consortium, 2019) and progression (Hensman Moss et al., 2017), as well as in animal work in both HD and myotonic dystrophy (van den Broek et al., 2002; Tomé et al., 2013). Hence, we again wanted to examine if there was rare, damaging variation in *MSH3* that could underlie HD onset modification. Coding variants from WES are shown in Table 4.6 (see non-coding variation in Appendix 5).

Firstly, we identify three, to our knowledge, novel LoF singleton variants in MSH3 that are all in patients with late onset (note: this is only near significant using Fisher's exact due to the rarity of these events, *p*=5.69E-02 (225, 212; 0, 3)). Two are splice acceptor variants in exons 16 and 17, respectively, that both occur in the final guanine critical for splicing. The third LoF variant is a frameshift occurring in exon 9. Unlike *FAN1* or *EXO1*, numbers of other NSD variants was low (excluding LoF, N= 7), and these do not seem to segregate to functional domains associated with early/late onset in any substantive fashion. For instance, Pro681Ser and Val682Leu variants are singletons only one residue apart yet occur in a late and early residual age at motor onset HD patient, respectively. The Glu82Val and Asn118Ile singleton variants, although not NSD, lie in the EXO1-MSH3 interaction domain (Schmutte et al., 2001). Overall, non-LoF NSD variants are less clearly associated with early/late HD onset than the LoF variants in *MSH3* with 4 early onset-associated NSDs and 7 late onset-associated NSDs (in 6 late onset patients), when including LoF variants – this is not significant using Fisher's exact test (*p*=5.36E-01) (221, 209; 4, 6).

Another feature of MSH3 is its imperfect repeat encoding a Pro/Ala tract between residues 49 to 73, and this has been implicated as being associated with altered HD onset (Flower et al., 2019). However, as with the *HTT* CAG repeat, we found *MSH3* calling was poor for STRs. This is reflected in the large numbers of missing genotypes of these variants (~14% call rate) and the lower depth coverage of this region compared to the rest of the gene. Attempts to manually assess read structure using extracted reads from SAMtools was largely unsuccessful due to its variable WES coverage and repetitive structure. We used estimates for genotype length using *MSH3* STR genotype calls (see 2.7.2.3) and plotted these against corrected age at motor onset residual, however this was not significant using a linear model (*p*=0.359) (see Appendix 8). Inconsistent coverage and a depleted N (only 108 exomes with enough genotype calls) may be responsible for this finding.

					Early				Late			
Variant	Location	DP	GnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Pro49Arg	5:79950692:C:G	31.70	9.99E-06	0.0	170	54	1	0	145	70	0	0
Ala57_Ala62del	5:79950699:TGCAG CGGCTGCAGCGGCC:T	22.21	2.73E-01	NA	192	9	19	5	175	10	14	16
Ala60_Ala62dup	5:79950708:TGCAGCGGCC:T	20.45	5.76E-03	NA	180	45	0	0	163	50	2	0
Ala60Pro	5:79950724:G:C	17.80	3.48E-02	10.6	214	9	2	0	196	19	0	0
Ala61_Pro63dup	5:79950724:G:GCCGCAGCGC	17.45	3.48E-02	NA	214	10	1	0	197	16	2	0
Pro67_Pro69del	5:79950741:GCCCCCAGCT:G	23.45	3.38E-01	NA	194	13	5	13	168	17	7	23
Ile79Val	5:79950781:A:G	21.17	8.67E-01	0.0	192	0	4	29	165	2	7	41
Glu82Val	5:79952237:A:T	42.50	NA	16.3	0	225	0	0	0	214	1	0
Asn118Ile	5:79952345:A:T	42.46	5.37E-05	12.1	0	225	0	0	0	214	1	0
Tyr462fs [*]	**5:80021311:CATTT:C**	**35.91**	**NA**	**NA**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Glu523Lys	5:80024783:G:A	20.43	6.29E-05	10.2	59	166	0	0	67	147	1	0
Thr552Ile	**5:80040326:C:T**	**19.59**	**5.42E-05**	**28.2**	**11**	**214**	**0**	**0**	**19**	**195**	**1**	**0**
Arg669Trp	**5:80063860:C:T**	**36.13**	**8.96E-06**	**33.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Pro681Ser	**5:80063896:C:T**	**36.15**	**1.56E-03**	**22.8**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Val682Leu	**5:80063899:G:C**	**35.34**	**3.50E-04**	**23.7**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Splice Acceptor [*]	**5:80071512:G:C**	**27.45**	**NA**	**25.2**	**3**	**222**	**0**	**0**	**4**	**210**	**1**	**0**
Splice Acceptor [*]	**5:80074538:G:A**	**36.37**	**NA**	**26.1**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Arg779His	**5:80074556:G:A**	**36.85**	**2.33E-04**	**35.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Glu853Gln	**5:80088565:G:C**	**26.92**	**5.39E-05**	**26.9**	**3**	**221**	**1**	**0**	**4**	**211**	**0**	**0**
Asn861His	5:80088589:A:C	29.53	8.97E-06	13.8	3	222	0	0	2	212	1	0
Leu911Trp	**5:80109479:T:G**	**37.76**	**3.38E-03**	**25.0**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Gln949Arg	5:80149981:A:G	24.72	8.41E-01	4.9	9	2	67	147	13	5	65	132
Asp1000Glu	5:80150135:T:G	20.04	0.00E+00	14.9	30	193	2	0	49	166	0	0
Ala1045Thr	5:80168937:G:A	31.85	7.14E-01	7.3	3	19	102	101	5	26	81	103

Table 4.6: Coding variation in *MSH3* from WES. Non-synonymous damaging (NSD) variants (NS CADD≥20 or LoF, and MAF≤1% in gnomAD NFE) are emboldened. Under GnomAD, 'NA' denotes variants not found in gnomAD v2.0.2, and 0.00E+00 indicates variants found in gnomAD but not in NFEs. Loss of function (LoF) are marked by [*]. Genomic locations are based on hg19/GRCh37 and CADD is from dbNSFP 3.0. Total N=440 (225 early; 215 late). DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

4.6.5 *TCERG1*

The transcription elongation regulator 1 gene, or *TCERG1* (formerly *CA150*, (Suñé et al., 1997)), has been implicated as an onset modifier of HD (Holbert et al., 2001) and lies close to a significant singleton SNP in the most recent GWAS (GeM-HD Consortium, 2019). However, it is still unclear whether the other candidate gene at the locus, *GPR151*, is responsible for the significant signal seen in GWAS (GeM-HD Consortium, 2019). Upon investigation, we found there were few NSD variants in *TCERG1* in our sample (Table 4.8 and Appendix 5), with only 3 singleton NSDs observed. We find more NSD variants in *GPR151* (Table 4.7 and Appendix 5), including several LoF variants, but there is no significant segregation of variants between early and late onset patients (p=0.37, Fisher's exact test) (212, 207; 13, 8) and no segregation of variants to functional domains.

However, we identify several short tandem repeat (STR) variants centred in *TCERG1's* imperfect repeat tract associated with altered HD onset. The TCERG1 protein has a repetitive glutamine-alanine (Q-A) structure, flanked by semi-repetitive glutamine-alanine/valine tracts (Q-A/V) on both the N and C terminal regions (see Fig. 4.15). The reference sequence for the TCERG1 protein is $(Q\text{-}A/V)_2(Q\text{-}A)_{29}(Q\text{-}A/V)_7$, however, and importantly for read alignment, the repetitive coding sequence is imperfect. Both glutamine codons and all four codons each of alanine and valine are used throughout the entirety of the quasi repeat tract. Consequently, unlike *HTT* or *MSH3* repeats, our 75bp reads are capable of effectively reading through *TCERG1*'s repeat, and the average depth of these variants was very high (mean depth = 39.3-42.3) with equally high call rates (>98%).

As indicated in table 4.8, we identify five STR variants. All the variants, bar one, revolve around the deletion or insertion of a hexanucleotide repeat, $(GCCCAG)_n$, encoding Q-A. The one exception was an individual with a $(GCCCAG)_2(GCCCAA)_2$ deletion ($((Q\text{-}A)_4)$). Individuals with smaller STR tracts tended to have earlier onset, whereas insertions were associated with a later disease onset. To investigate the STR's impact on onset quantitatively, we filtered exomes with missing STRs (see 2.7.2.3), leaving N=440 exomes (E=230, L=210). We plotted a linear model where we regressed corrected AMO residual (pure CAG) on *TCERG1* genotype length. Genotype length was calculated as the glutamine-alanine repeat amino acid length compared to the reference sequence and ranged from -4 to +2 amino acids (see 2.7.2.3). Exomes with reference length *TCERG1* were given values of 0. Regressing corrected onset residual against genotype length as per Fig. 4.16 showed significance (p=2.48E-03) with B=-3.430 and β=-0.144. Hence, our model finds that for each single Q-A added this results in a ~7 year later HD onset in our dataset. Interpreting these data is difficult, however, given our extreme phenotype selection; see 6.2.3.

					Early				Late			
Variant	Location	DP	GnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Glu387Gln	**5:145894518:C:G**	**36.78**	**2.69E-05**	**26.1**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ser366Thr	5:145894580:C:G	34.52	8.95E-06	0.0	0	225	0	0	0	214	1	0
Leu304fs [*]	**5:145894764:TGA:T**	**30.51**	**1.79E-05**	**NA**	**1**	**224**	**0**	**0**	**0**	**214**	**1**	**0**
Ile288Thr	5:145894814:A:G	28.27	NA	9.8	1	223	1	0	3	212	0	0
Pro284Ser	5:145894827:G:A	28.72	NA	23.9	2	222	1	0	0	215	0	0
Leu261Val	5:145894896:G:C	44.20	2.00E-01	0.0	0	138	78	9	0	140	68	7
Phe175fs [*]	**5:145895150:CTA:C**	**25.98**	**1.70E-04**	**NA**	**2**	**223**	**0**	**0**	**6**	**208**	**1**	**0**
Ala144Val	**5:145895246:G:A**	**38.58**	**1.37E-03**	**22.8**	**0**	**225**	**0**	**0**	**0**	**213**	**2**	**0**
Arg95Ter [*]	**5:145895394:G:A**	**39.59**	**7.07E-03**	**36.0**	**0**	**216**	**9**	**0**	**0**	**211**	**4**	**0**
Pro40Leu	5:145895558:G:A	42.65	7.78E-02	16.4	0	196	28	1	0	182	33	0
Tyr27Ter [*]	**5:145895596:G:T**	**37.28**	**8.93E-04**	**36.0**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Phe23Leu	**5:145895608:A:C**	**34.43**	**8.23E-04**	**25.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**

Table 4.7: Coding variation in *GPR151* from WES. Non-synonymous damaging (NSD) variants (NS CADD≥20 or LoF, and MAF≤1% in gnomAD NFE) are emboldened. Under GnomAD, 'NA' denotes variants not found in gnomAD v2.0.2. Loss of function (LoF) are marked by [*]. Genomic locations are based on hg19/GRCh37 and CADD is from dbNSFP 3.0. Total N=440 (225 early; 215 late). DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

					Early				Late			
Variant	Location	DP	GnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Ala33Thr	**5:145834656:G:A**	**26.40**	**1.74E-03**	**27.5**	**5**	**219**	**1**	**0**	**1**	**214**	**0**	**0**
Gln222_Ala223insGlnAla	5:145838635:T:TCAGGCC	39.60	NA	NA	3	221	1	0	4	211	0	0
Gln222_Ala223insGlnAlaGlnAla	5:145838635:T:TCAGGCCCAGGCC	39.60	NA	NA	3	216	6	0	4	209	2	0
Gln222_Ala223delGlnAlaGlnAla	5:145838635:TCAGGCCCAGGCC:T	39.60	NA	NA	3	217	5	0	4	196	14	1
Gln222_Ala223delGlnAlaGlnAlaGlnAla	5:145838635:TCAGGCCCAGGCCCAGGCC:T	39.39	NA	NA	4	212	9	0	4	186	25	0
Gln222_Ala229delAlaGlnAlaGlnAlaGlnAlaGln	5:145838656:GGCCCAGGCCCAGGCCCAAGCCCAA:G	42.25	5.88E-04	NA	12	213	0	0	27	187	1	0
Thr283Ala	5:145838855:A:G	28.31	0.00E+00	14.4	0	224	1	0	3	212	0	0
Val329Ala	5:145843207:T:C	37.29	1.43E-04	19.5	0	225	0	0	0	214	1	0
Met412Val	5:145849142:A:G	22.00	6.30E-05	13.0	11	213	1	0	8	207	0	0
Asn742Ser	5:145872595:A:G	20.44	NA	15.8	16	208	1	0	18	197	0	0
Ala779Val	**5:145878203:C:T**	**21.92**	**NA**	**25.8**	**12**	**213**	**0**	**0**	**13**	**201**	**1**	**0**
Ile1008Val	**5:145888735:A:G**	**13.87**	**NA**	**22.0**	**134**	**91**	**0**	**0**	**150**	**64**	**1**	**0**

Table 4.8: Coding variation in *TCERG1* from WES. Non-synonymous damaging (NSD) variants (NS CADD≥20 or LoF, and MAF≤1% in gnomAD NFE) are emboldened. Under GnomAD, ‘NA’ denotes variants not found in gnomAD v2.0.2, and 0.00E+00 indicates variants found in gnomAD but not in NFEs. Loss of function (LoF) are marked by [*]. Genomic locations are based on hg19/GRCh37 and CADD is from dbNSFP 3.0. Total N=440 (225 early; 215 late). DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

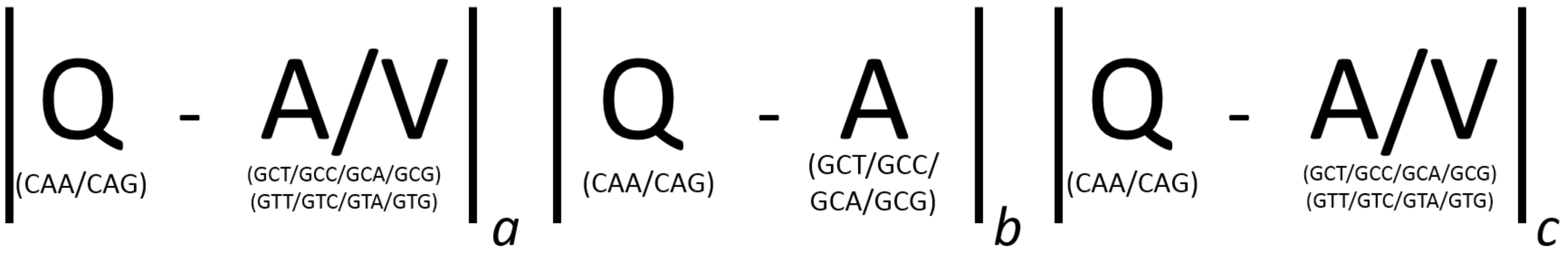


Figure 4.15: Structure of the TCERG1 repeat. Indicated is the amino acid structure of the TCERG1 protein and the semi-repetitive codons in the *TCERG1* gene. Reference sequence in GRCh37 is a=2, b=29 and c=7, *i.e.* $(Q\text{-}A/V)_2(Q\text{-}A)_{29}(Q\text{-}A)_7$. Q: Glutamine; A: alanine.

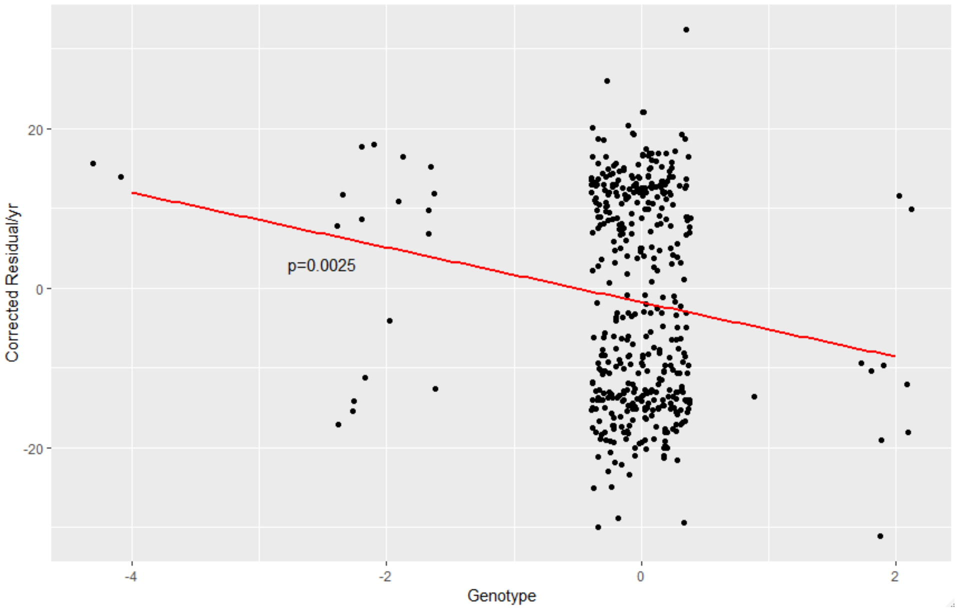


Figure 4.16: *TCERG1* genotype length plotted against corrected AMO residual. The length of the repetitive STR tract is plotted against corrected AMO residuals from MiSeq. Points have been jittered on the x-axis and exomes with missing genotypes were removed. N=240; E=230, L=210. Genotype length was defined as the length of the *TCERG1* glutamine-alanine tract (in repeating units) relative to the reference sequence (see 2.7.2.3).

4.6.6 *LIG1*

The DNA Ligase 1 gene (*LIG1*) is another DNA repair gene that participates in several DNA repair processes including mismatch repair and base-excision repair (BER), and has been implicated by GWAS as an HD onset modifier (GeM-HD Consortium, 2019). *LIG1* coding variation is shown in Table 4.9 (see Appendix 5 for non-coding variation) and plotted in Fig. 4.17. While there is no significant enrichment in either early or late onset groups (NSD early variants=11; NSD late variants=7, *p*=4.74E-01, Fisher's exact test (214, 208; 11, 7), there is some degree of domain clustering observed. The DNA ligase A N domain (DNA binding) has three early-associated variants occurring near each other: Pro395Leu, Val349Met and Leu335Phe. The segregation of the remaining variation, however, is less clear. While parts of the DNA ligase A M domain (nucleotidyltransferase) have further early-associated variants, the call rates of the Arg774Gln and Val753Met variants (in DNA ligase A C domain containing the oligonucleotide-binding fold), both of which occur on exon 24, are quite low, especially for Val753Met (call rate ~20%). The Lys845Asn variant was associated with late onset. Therefore, while functional domain clustering in *LIG1* between in HD onset may be of interest, poorer coverage of *LIG1*, especially exon 24, affects interpretation of these data.

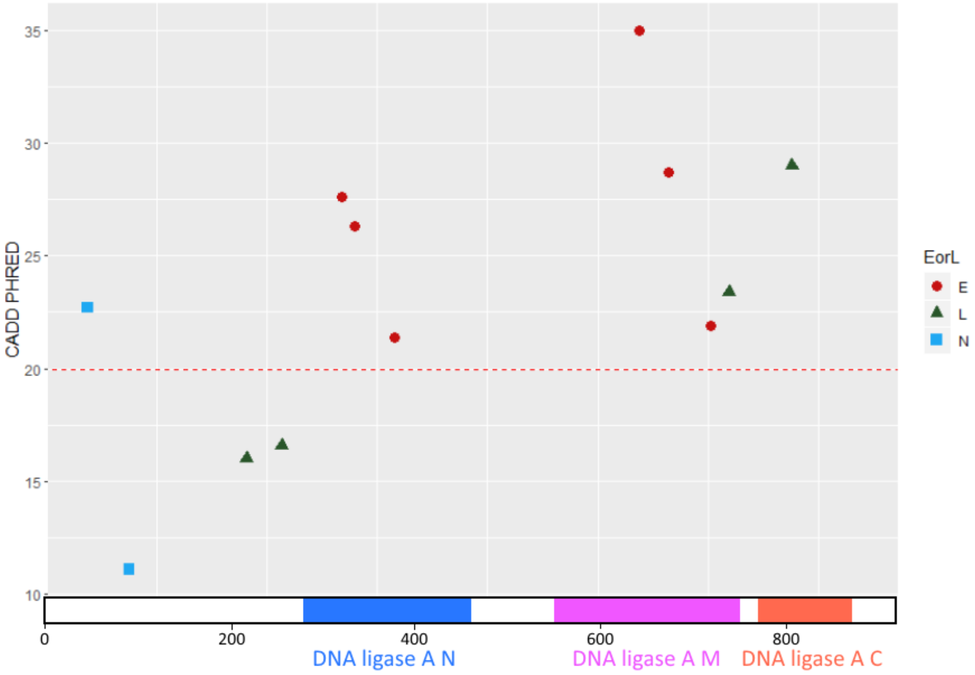


Figure 4.17: Structural overview of identified LIG1 variants. Variants with MAF≤1% are shown plotted against a cartoon of the LIG1 protein. The dashed line indicates the CADD 20 cut-off. Numbers of each mutation are not indicated. Red circles represent variants more associated with early onset (E); green triangles are variants more associated with late onset (L); blue squares are not associated strongly with either (N). Domain boundaries are taken from PhosphoSitePlus® (Hornbeck et al., 2015), (Pascal et al., 2004) and (McNally and O'Brien, 2017). N=225 early, 215 late HD patients.

					Early				Late			
Variant	Location	DP	GnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Lys845Asn	**19:48620943:C:A**	**38.36**	**1.69E-03**	**29.0**	**0**	**224**	**1**	**0**	**0**	**210**	**5**	**0**
Arg774Gln	**19:48624491:C:T**	**20.18**	**1.18E-04**	**23.4**	**27**	**198**	**0**	**0**	**28**	**185**	**2**	**0**
Val753Met	**19:48624555:C:T**	**12.61**	**9.24E-03**	**21.9**	**166**	**56**	**3**	**0**	**182**	**32**	**1**	**0**
Glu705Lys	**19:48626467:C:T**	**40.05**	**0.00E+00**	**28.7**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg672Cys	**19:48626566:G:A**	**37.04**	**3.49E-04**	**35.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg409His	19:48640807:C:T	36.30	2.06E-02	23.5	0	217	7	1	2	212	1	0
Pro395Leu	**19:48640849:G:A**	**28.25**	**5.43E-05**	**21.4**	**0**	**224**	**1**	**0**	**5**	**210**	**0**	**0**
Ala369Val$^{\Delta}$	19:48640902:G:A	16.95	1.36E-02	NA	64	159	2	0	72	137	6	0
Val349Met	**19:48643270:C:T**	**41.31**	**2.48E-03**	**26.3**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Leu335Phe	**19:48643312:G:A**	**40.67**	**2.53E-03**	**27.6**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Asn267Ser	19:48647197:T:C	40.88	7.16E-05	16.6	0	225	0	0	0	214	1	0
Thr227Met	19:48653362:G:A	25.72	8.96E-06	16.0	5	220	0	0	6	208	1	0
Arg94Cys	19:48660361:G:A	35.15	2.43E-03	11.1	1	223	1	0	0	214	1	0
Ala60Gly	19:48664693:G:C	41.55	NA	10.9	0	224	1	0	0	215	0	0
Ser47Phe	**19:48664732:G:A**	**36.69**	**9.05E-04**	**22.7**	**0**	**224**	**1**	**0**	**1**	**213**	**1**	**0**

Table 4.9: Coding variation in *LIG1* from WES. Non-synonymous damaging (NSD) variants (NS CADD≥20 or LoF, and MAF≤1% in gnomAD NFE) are emboldened. Under GnomAD, 'NA' denotes variants not found in gnomAD v2.0.2, and 0.00E+00 indicates variants found in gnomAD but not in NFEs. Loss of function (LoF) are marked by [*]. Δ denotes variants where the most damaging consequence only occurs in the non-canonical transcript. Genomic locations are based on hg19/GRCh37 and CADD is from dbNSFP 3.0. Total N=440 (225 early; 215 late). DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

4.6.7 Other candidate genes

We also investigated other genes identified as candidate modifier genes from the most recent HD GWAS (GeM-HD Consortium, 2019), and tables for these are available in Appendix 5 (non-coding) and Appendix 6 (coding). Firstly, we considered whether there were rare variants in *HTT* itself present, independent of CAG length and allele structure, associated with altered onset. However, only five NSD variants were found in *HTT*: Gly696Glu, Asp1082His, Thr1260Met, Val1551Ala and Arg2002His. There was no clear association of these variants with early or late onset.

Several NSD variants were identified in *MLH1*: Ala188Val, Tyr379Cys, Lys618Glu, Lys618Thr, Val716Met and His718Tyr, but with no clear overall segregation to early or late onset groups. Notably, though, we did identify a common NS variant (MAF=32%), Ile219Val (rs1799977), that had association with late onset (*p*=4.90E-03, Fisher exact test) (123, 84, 16; 86, 98, 28). Rs179997 in *MLH1* was identified as a candidate SNP for onset modification in GWAS and elsewhere (Lee et al., 2017; GeM-HD Consortium, 2019), and plotting genotype against our MiSeq corrected residuals in our continuous population (N=485) shows an additive, dose-dependent effect (Fig. 4.18; linear model *p*=1.63E-03, *B*=2.92 and β=0.143).

MLH1 binds PMS1, PMS2 and MLH3 forming the MutL complexes. We next investigated these as two of these three have been implicated by GWAS (*PMS1* and *PMS2*, (GeM-HD Consortium, 2019)). There were 3 NSD variants identified in the canonical transcript of *PMS1*: Thr75Ile, Gly501Arg and Arg569Gln. These were all in late onset individuals. The Arg202Lys *PMS1* variant, although not NSD (CADD = 19.2, MAF NFE gnomAD=1.38%), was found in an excess of late individuals (2:7 E:L). There was no clear early or late association of damaging coding variants in *PMS2* or *MLH3* (Appendix 6).

We also investigated *OGG1* as it has been implicated as a CAG repeat modifier elsewhere (Kovtun et al., 2009; Budworth et al., 2015). No clear onset or domain segregation of variants was seen in *OGG1*, although the Gly308Glu NSD variant was found in a slightly larger number of late onset individuals (3:6, E:L). A single early-associated (6:1, E:L) NSD variant, Leu353Val, was found in *SYT9*, although the coverage of this gene was lower; see Appendix 9. Although we attempted to investigate variation in *RRM2B*, the coverage of the gene was very low (coverage for target genes in Appendix 9), and no NSD variants were identified.

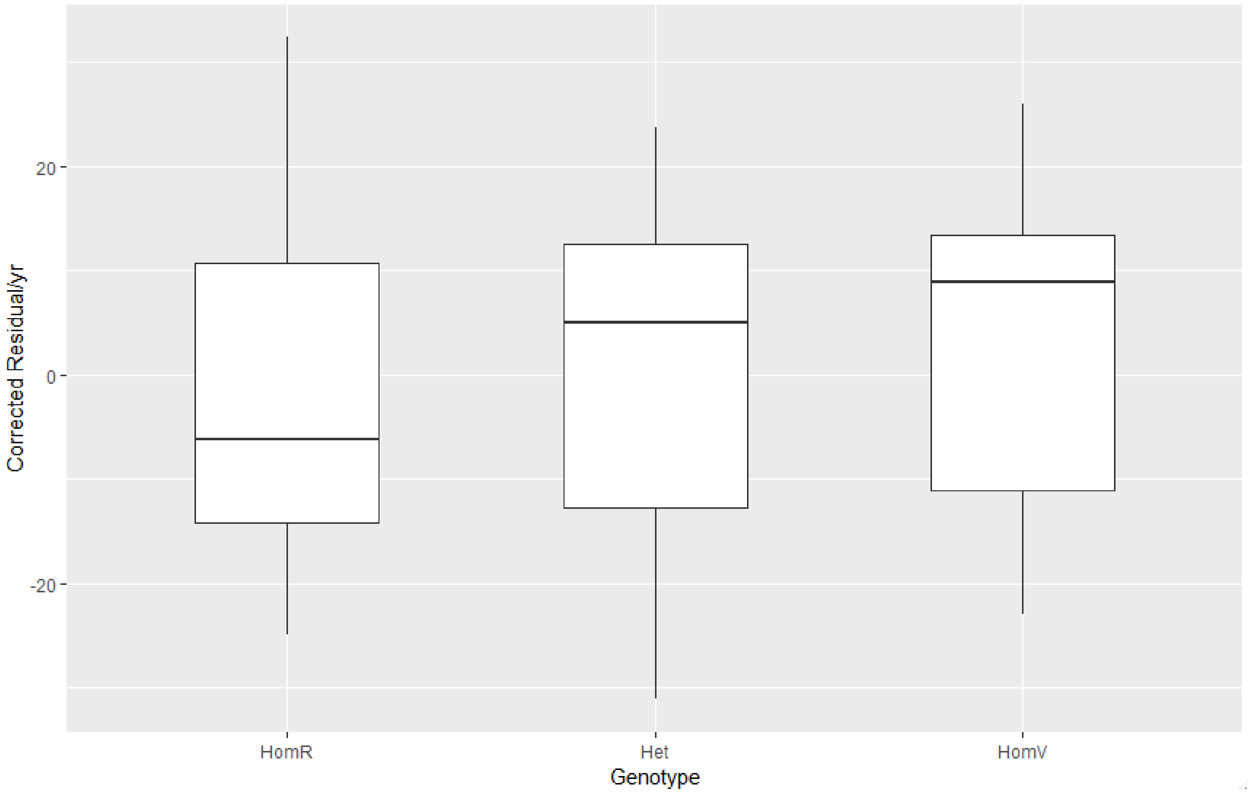


Figure 4.18: Effect of the *MLH1* Ile219Val variant (rs1799977) on residual AMO. All individuals where the *MLH1* Ile219Val (rs1799977 / 3:37053568:A:G) variant is called (N=480; 241 early, 239 late) are included. The y-axis uses corrected residual from MiSeq (pure CAG length, correcting for *HTT* allele structure). HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

4.7 Logistic burden regression analysis

Until now, our analysis has focused on genes with prior evidence for either HD onset modification in GWAS or that have been implicated through other work (*e.g. EXO1* and *OGG1*). For this reason, we wanted to implement a technique to detect genetic variation in potentially novel genes and pathways in an unbiased fashion.

As a starting point, we implemented a logistic regression using each variant identified (a variant-by-variant analysis) to see if there were single variants that were found in disproportionate numbers in our early and late populations (Appendix 10; methods 2.7.3.1). No NSD or LoF variant passed exome-wide significance (Bonferroni threshold *p*=6.68E-07 for NSD; *p*=3.55E-06 for LoF) with the covariates used (PC1-5 and mean depth). This test was likely unsuccessful as (1) Bonferroni multiple test correction results in very low *p* value thresholds and (2) rare (*e.g.* ≤1% MAF) variants are likely spread out between many different individual variant calls and are unlikely to achieve significance individually. Furthermore, we realised that we needed to further refine our input data into the burden regression by incorporating a MAF filter, as this would allow us to investigate rare modifiers only (*i.e.* ≤1% MAF).

Taking these issues into account, we devised a whole-exome burden regression test (methods 2.7.3.2) whereby individuals from our dichotomous exome group (N=440) were coded as early (1) or late (0) onset. As testing individual rare variants has limited power due to their inherent small frequency (Zuk et al., 2014), we aggregated damaging variants (≥20 CADD or LoF) across genes at different MAF cut-offs: very rare (MAF≤0.1%), rare (MAF≤1%) and uncommon (MAF≤2%). Note that variants at lower MAF cut-offs were also included at higher MAF cut-offs. We filtered variants by GATK's variant quality score recalibration (VQSR≥98.5) and call rate (call rate≥75%). Covariates chosen were principal components 1-5, mean sample variant depth and baseline variant rate (BVR). Baseline variant rate allows for differences in genetic architecture or coverage between exomes to be considered in the analysis. We initially produced these statistics for most of the previously investigated candidate genes in 4.6 (Table 4.10). Notably, *FAN1* was nominally significant at MAF 1 and 2% (*p*~1E-02) and *EXO1* is approaching nominal significance (*p*~6E-02) at the same MAFs. Possibly due to our relatively small sample set, none of the candidate genes were significant at MAF 0.1%.

We then extended this analysis to the entire exome at MAFs 0.1, 1 and 2%. The top 15 genes are available in Table 4.11. A total of 21,864 genes and ORFs were tested, with 6234 producing a *p* value. We further refined this by introducing a filter wherein only genes/ORFs with >5 damaging variants at the MAF being investigated were considered leaving 1590, 4261 and 5084 genes tested at MAF 0.1, 1 and 2%, respectively. No gene was found to pass exome-wide significance adjusting for multiple testing correction (Bonferroni threshold *p*=3.14E-05 (MAF 0.1%); *p*=1.17E-05 (MAF 1%); *p*=9.83E-06 (MAF 2%)), although *FAN1* was the 7th highest in the exome at MAF 1% (*p*=1.17E-2). Interestingly, *NOP14* was also found with a low *p* value at MAF 1%. *NOP14* is found on the short arm of chromosome 4 (4p16.3), ~133kb from *HTT*. Investigation of the *NOP14* signal revealed a single NSD variant (26.5 CADD), Arg697Cys (4:2943419:G:A), that drives most of the *NOP14* signal. Arg697Cys was highly skewed, found as a heterozygote in 18 late onset individuals and no early individuals. Following targeted MiSeq sequencing (chapter 5), we found that in 16 of the 18 variant calls, Arg697Cys was found in an individual with a (CAG**CAA**)$_2$CAG atypical *HTT* allele. In the 2 outstanding cases of this *NOP14* variant, both *HTT* alleles had canonical structures. The Arg697Cys variant, therefore, seems to be in strong linkage disequilibrium (LD) with (CAG**CAA**)$_2$CAG *HTT* structure.

We next explored potential extensions to the basic logistic burden regression approach by weighting variants. Initially, we weighted based on rarity (1/MAF), however we found that this enriched for unusual genes and the *p* values were very high (Appendix 11), and imputing missing MAFs artificially selected for variants poorly covered in gnomAD but better covered in our study. Instead, we found weighting variants based on their CADD PHRED score (a measure of deleteriousness) was viable (Appendix 12), although we did not find this greatly changed the *p* values obtained. Still, weighting may be a useful addition to future modelling.

Finally, we wanted to examine the effectiveness of logistic burden regression by plotting a quantile-quantile (Q-Q) plot of the obtained *p* values for genes with >5 variants. Fig 4.19 shows the logistic burden regression method has more deflated p values than one would expect, with a very low genomic inflation factor (λ) of 0.645. One would expect λ to be near 1 if there is no inflation or deflation of *p* values. Consequently, we wanted to explore other options available for whole-exome analyses, as detailed in the next three sections.

	MAF≤0.1%			MAF≤1%			MAF≤2%		
	B	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*
EXO1	-0.704	0.877	4.22E-01	-0.630	0.344	6.74E-02	-0.629	0.345	6.80E-02
FAN1	0.696	0.717	3.32E-01	**1.045**	**0.414**	**1.17E-02**	**1.035**	**0.414**	**1.24E-02**
HTT	-0.171	0.626	7.84E-01	0.278	0.493	5.72E-01	0.288	0.493	5.59E-01
LIG1	0.394	0.744	5.97E-01	-0.004	0.500	9.94E-01	0.004	0.500	9.94E-01
MLH1	-0.510	1.250	6.83E-01	-0.228	0.568	6.88E-01	-0.219	0.568	7.00E-01
MLH3	-0.022	0.026	3.92E-01	-0.004	0.018	8.19E-01	-0.003	0.018	8.50E-01
MSH3	-1.100	0.855	1.98E-01	-0.519	0.660	4.32E-01	-0.505	0.660	4.44E-01
OGG1	0.261	1.050	8.04E-01	0.000	0.019	9.86E-01	0.000	0.019	9.92E-01
PMS1	NA	NA	NA	NA	NA	NA	NA	NA	NA
PMS2	5.904	10.428	5.71E-01	5.879	10.226	5.65E-01	0.152	0.684	8.24E-01
RRM2B	NA	NA	NA	NA	NA	NA	NA	NA	NA
SYT9	NA	NA	NA	1.706	1.088	1.17E-01	1.681	1.088	1.23E-01
TCERG1	NA	NA	NA	-0.227	1.439	8.74E-01	-0.227	1.442	8.75E-01

Table 4.10: Logistic regression for candidate genes. Indicated are the β, *SE* and *p* values from candidate genes using logistic burden regression (Wald) in Hail (same as in Table 4.11). Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. No weighting of variants was used, and no filter based on the number of variants was in place for the targeted test. Nominally significant *p* values are emboldened. *B*: unstandardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

MAF≤0.1% (N=1590 adj)				MAF≤1% (N=4261 adj)				MAF≤2% (5084 adj)			
Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*
DENND4B	1.378	4.53E-01	2.32E-03	CUBN	1.596	4.35E-01	2.42E-04	DENND4B	1.403	4.54E-01	1.98E-03
CUBN	1.831	6.47E-01	4.62E-03	DENND4B	1.380	4.52E-01	2.28E-03	CUBN	0.685	2.31E-01	2.99E-03
MMP21	-2.444	1.06E+00	2.08E-02	ERAP2	-2.234	7.57E-01	3.16E-03	ERAP2	-2.220	7.57E-01	3.36E-03
MYO18B	1.198	5.20E-01	2.12E-02	SIPA1L2	1.828	6.31E-01	3.80E-03	PGC	2.259	7.72E-01	3.43E-03
GLDC	-1.788	7.86E-01	2.29E-02	GLI3	1.371	5.11E-01	7.34E-03	SIPA1L2	1.832	6.31E-01	3.71E-03
TENM2	-1.504	6.67E-01	2.41E-02	C9	2.011	7.76E-01	9.59E-03	ZNF462	-2.979	1.03E+00	3.99E-03
PCDH15	-1.233	5.70E-01	3.05E-02	FAN1	1.045	4.14E-01	1.17E-02	GLI3	1.371	5.10E-01	7.22E-03
FBRSL1	2.279	1.07E+00	3.27E-02	ZNF462	-2.632	1.05E+00	1.19E-02	AKR1C3	1.376	5.23E-01	8.52E-03
SYT10	-2.261	1.06E+00	3.36E-02	ATP1A4	-1.445	5.78E-01	1.24E-02	PCDH15	-0.958	3.69E-01	9.38E-03
TRPM1	1.401	6.61E-01	3.41E-02	CC2D1A	-1.613	6.53E-01	1.35E-02	C9	2.012	7.76E-01	9.52E-03
STAB1	-1.132	5.36E-01	3.47E-02	CACNA1I	-1.648	6.71E-01	1.41E-02	RNMTL1	0.776	3.00E-01	9.70E-03
MACF1	-1.145	5.46E-01	3.60E-02	ANXA11	1.875	7.64E-01	1.42E-02	TRIM66	-0.955	3.70E-01	9.90E-03
ITGB4	-1.049	5.01E-01	3.61E-02	MYO1A	1.226	5.01E-01	1.45E-02	NLRP1	-2.691	1.04E+00	1.00E-02
COL17A1	-2.190	1.07E+00	4.04E-02	NOP14	-0.922	3.78E-01	1.48E-02	DNAJA4	-1.709	6.66E-01	1.03E-02
DNAH7	-1.264	6.21E-01	4.18E-02	UNC5B	-1.572	6.47E-01	1.51E-02	ATG4B	-1.939	7.71E-01	1.19E-02

Table 4.11: Logistic regression of the entire exome (MAF 0.1, 1 and 2%), no weighting. Indicated are the β, *SE* and *p* values for the top 15 genes using logistic burden regression (Wald) in Hail. Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. No weighting of variants was used. Indicated at the top of each column are the adjusted (adj) values for each MAF cut-off, which is a count of how many genes/ORFs were tested that resulted in a *p* value and have >5 variants at the MAF tested. *B*: unstandardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

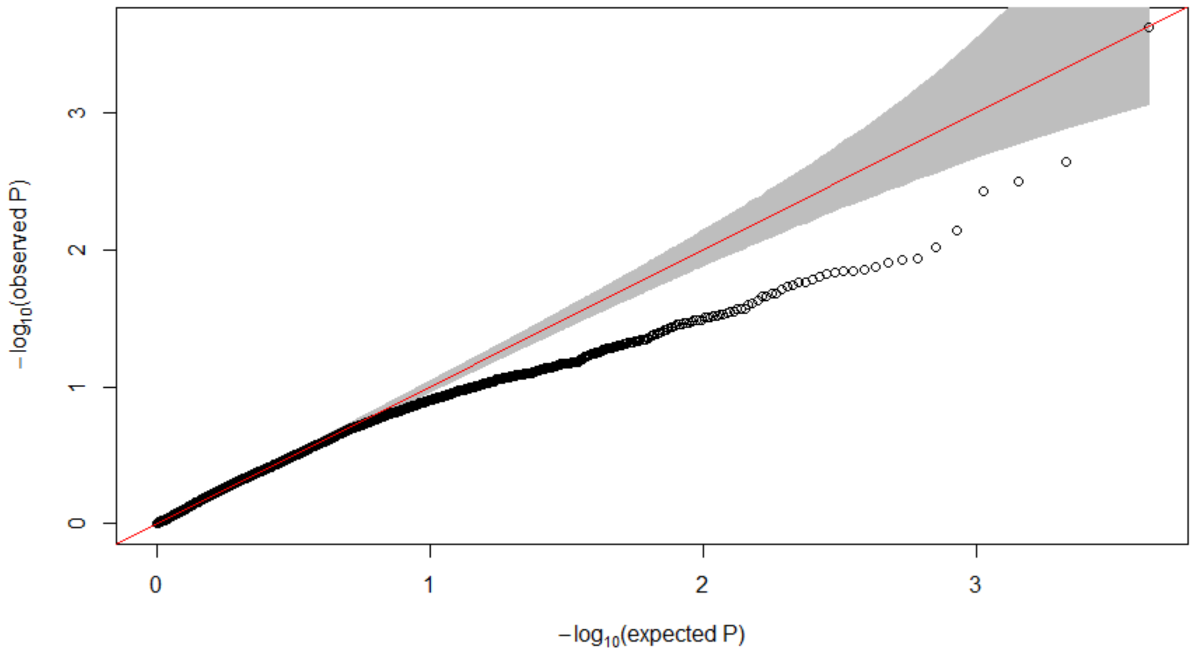


Figure 4.19: Q-Q plot of whole-exome burden regression. *p* values from unweighted whole-exome logistic regression (Wald test) are taken from those presented in Table 4.11 for MAF 1%. Genomic inflation factor (λ)=0.645. Genes plotted had >5 variants NSD variants at MAF 1%.

4.8 Linear burden regression analysis

A dichotomous whole-exome analysis indicated that burden regression can detect variation associated with altered onset in our HD exomes, including a known onset-modifying gene (*FAN1*) as well as *NOP14*, which is in LD with *HTT* CAG structure. However, logistic burden regression showed highly deflated *p* values with a very low λ. Hence, we wanted to investigate linear burden regression as an alternative method, regressing AMO residual as a continuous phenotype on the total number of NSD variants (2.7.3.3). Linear regression has the additional advantage of incorporating all exomes passing QC (N=485). We adopted a similar method as before, collapsing variants by genes using PC 1-5, BVR and mean sample depth as covariates (methods 2.7.3.3). As before, only variants with >5 damaging variants (LoF or NS CADD≥20) were considered for the whole-exome analyses.

To begin with, we regressed uncorrected AMO residuals from MiSeq (*i.e.* $(\text{polyQ})_n$-2 CAG lengths used to calculate estimated AMOs) on the NSD variant load. The top 15 genes from the burden regression are shown in Table 4.12. The uncorrected residual does not correct for CAG allele structure. Thus, it is unsurprising that *NOP14* has a low *p*-value (*NOP14*, *p*=3.42E-03), as the *NOP14* Arg697Cys variant tags the non-canonical $\text{CAG}_n(\textbf{CAA}\text{CAG})_2$ allele, as discussed previously. *CUBN* is exome-wide significant at MAF 1% (*p*=6.96E-06; Bonferroni threshold *p*=1.06E-05), although only nominally at MAF 0.1% (*p*=2.15E-04; Bonferroni threshold *p*=2.61E-05) and MAF 2% (*p*=8.86E-04; Bonferroni threshold *p*=9.05E-06). *CUBN* codes for the Cubilin protein, a cotransporter involved in the uptake of cobalamin (vitamin B12). Interestingly, the *MUT* gene, coding for methylmalonyl-CoA mutatase (MUT) also involved with vitamin B12 metabolism, has a low (but only nominally significant) *p* value (*p*=8.25E-04). The coding variants for *CUBN*, *NOP14,* and *MUT* are available in Appendix 7.

The linear burden regression was then repeated, this time using the corrected residual from MiSeq (pure CAG length) (Table 4.14). As before, *FAN1* and other candidate genes from 4.5 are shown in Table 4.13. Notably, *FAN1*, *EXO1, OGG1* and *PMS1* showed at least nominal significance at least one MAF, however none survive multiple testing correction. As expected, using the corrected AMO residual ablates most of the *NOP14* signal (*p*=3.42E-03 uncorrected to *p*=6.17E-02 corrected). *CUBN* is also exome-wide significant here at *p*=7.97E-06. As before, we were also able to weight variants based on deleteriousness (Appendix 13). Plotting a Q-Q plot of the unweighted linear burden regression data using corrected residuals (pure CAG length) (Fig 4.20) shows the *p*-values generated are much more in line with what are expected given the null hypothesis (λ=1.047). Thus, linear/continuous analyses seem to be much more appropriate compared to dichotomous/logistic techniques in our HD cohort, at least when considering burden testing.

MAF≤0.1% (N=1915)				MAF≤1% (N=4737)				MAF≤2% (N=5524)			
Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*
CUBN	-11.005	2.951	2.15E-04	**CUBN**	**-8.806**	**1.937**	**6.96E-06**	ZNF462	13.239	3.140	2.98E-05
CACNA1G	-15.685	4.414	4.18E-04	SIPA1L2	-10.962	3.009	2.99E-04	PGC	-12.064	3.235	2.15E-04
ZNF462	13.442	3.900	6.18E-04	ERAP2	11.153	3.177	4.90E-04	SIPA1L2	-10.991	3.008	2.87E-04
CGN	17.081	5.291	1.33E-03	TEKT1	18.378	5.287	5.55E-04	ERAP2	11.101	3.185	5.36E-04
DENND4B	-7.914	2.472	1.46E-03	CACNA1G	-13.718	4.056	7.79E-04	TEKT1	18.336	5.284	5.68E-04
FBRSL1	-14.091	4.418	1.52E-03	MUT	-11.678	3.470	8.25E-04	CACNA1G	-13.720	4.054	7.72E-04
GLRA4	-14.955	4.759	1.78E-03	ZNF462	12.216	3.635	8.41E-04	MUT	-11.603	3.463	8.70E-04
GLDC	11.472	3.894	3.37E-03	KIAA0319	-14.014	4.224	9.77E-04	KIAA0319	-14.161	4.226	8.71E-04
KIAA0319	-16.115	5.710	4.97E-03	ENPP7	-10.878	3.306	1.08E-03	CUBN	-4.639	1.387	8.86E-04
PRKRIR	11.239	4.034	5.55E-03	GRTP1	14.430	4.443	1.25E-03	NLRP1	10.023	3.018	9.66E-04
DMRT2	-17.302	6.265	5.97E-03	DENND4B	-7.955	2.471	1.37E-03	ENPP7	-10.908	3.305	1.04E-03
TEKT2	-14.856	5.406	6.22E-03	ANXA11	-11.195	3.526	1.59E-03	GRTP1	14.495	4.431	1.15E-03
NUP210L	-12.661	4.688	7.17E-03	GLRA4	-15.121	4.770	1.62E-03	DENND4B	-8.070	2.471	1.17E-03
SON	11.399	4.241	7.44E-03	SCYL1	-14.668	4.657	1.74E-03	RNMTL1	-5.179	1.594	1.24E-03
SYNPO2	13.189	4.951	7.98E-03	PCDH15	8.367	2.735	2.34E-03	ANXA11	-11.243	3.521	1.50E-03

Table 4.12: Linear regression (uncorrected residual) of the entire exome (MAF 0.1, 1 and 2%), no weighting. Indicated are the β, *SE* and *p* values for the top 15 genes using linear burden regression in Hail. Exome-wide significant genes which survive multiple testing correction are emboldened. The uncorrected residual from MiSeq (polyglutamine-2) was used for regression. Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were not weighted. Indicated at the top of each column are the adjusted (adj) values for each MAF cut-off, which is a count of how many genes/ORFs were tested that resulted in a *p* value and have >5 variants at the MAF tested. N=485. *B*: unstandardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

	MAF≤0.1%			MAF≤1%			MAF≤2%		
	B	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*
EXO1	3.457	5.102	4.98E-01	**4.978**	**2.067**	**1.64E-02**	**4.982**	**2.067**	**1.63E-02**
FAN1	-3.905	4.505	3.87E-01	**-6.816**	**2.25**	**2.59E-03**	**-6.768**	**2.251**	**2.78E-03**
HTT	0.507	3.945	8.98E-01	-1.147	3.075	7.09E-01	-1.102	3.072	7.20E-01
LIG1	-3.335	4.786	4.86E-01	0.429	3.222	8.94E-01	0.382	3.222	9.06E-01
MLH1	4.369	6.012	4.68E-01	1.746	3.489	6.17E-01	1.763	3.489	6.14E-01
MLH3	3.238	4.282	4.50E-01	0.538	2.648	8.39E-01	0.539	2.650	8.39E-01
MSH3	2.063	4.506	6.47E-01	1.613	3.76	6.68E-01	1.573	3.761	6.76E-01
OGG1	**-21.983**	**9.430**	**2.02E-02**	-1.527	3.767	6.85E-01	-1.570	3.76604	6.77E-01
PMS1	6.023	6.727	3.71E-01	**12.272**	**5.472**	**2.54E-02**	**12.274**	**5.472**	**2.54E-02**
PMS2	-14.521	7.718	6.05E-02	-14.672	7.718	5.79E-02	-1.673	4.276	6.96E-01
RRM2B	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
SYT9	NaN	NaN	NaN	-7.524	5.088	1.40E-01	-7.422	5.094	1.46E-01
TCERG1	15.051	13.38	2.61E-01	-0.265	9.517	9.78E-01	-0.275	9.518	9.77E-01

Table 4.13: Linear regression for candidate genes. Indicated are the β, SE and p values from candidate genes using linear burden regression in Hail with corrected residual age at motor onset (pure CAG length – the same as in Table 4.14). Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. No weighting of variants was used, and no filter based on the number of variants was in place for the targeted test. Genes and MAFs that pass nominal significance are emboldened. *B*: unstandardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

MAF≤0.1% (N=1915)				MAF≤1% (N=4737)				MAF≤2% (N=5524)			
Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*
CUBN	-10.756	2.824	1.58E-04	**CUBN**	**-8.383**	**1.856**	**7.97E-06**	ZNF462	12.293	3.012	5.26E-05
CACNA1G	-14.769	4.229	5.23E-04	SIPA1L2	-10.658	2.881	2.42E-04	PGC	-11.498	3.100	2.33E-04
DENND4B	-7.926	2.365	8.67E-04	ERAP2	10.701	3.044	4.80E-04	SIPA1L2	-10.681	2.881	2.34E-04
FBRSL1	-14.124	4.226	8.98E-04	DENND4B	-7.967	2.365	8.17E-04	CUBN	-4.827	1.326	3.01E-04
GLRA4	-15.193	4.552	9.10E-04	GLRA4	-15.245	4.564	9.03E-04	ERAP2	10.690	3.051	5.02E-04
CGN	16.870	5.064	9.31E-04	CACNA1G	-12.974	3.887	9.10E-04	RNMTL1	-5.329	1.524	5.17E-04
GLDC	12.057	3.722	1.28E-03	KIAA0319	-13.396	4.047	1.00E-03	DENND4B	-8.058	2.366	7.15E-04
ZNF462	11.916	3.742	1.54E-03	MUT	-10.886	3.326	1.14E-03	CACNA1G	-12.992	3.885	8.91E-04
DMRT2	-17.310	5.995	4.06E-03	GRTP1	13.787	4.257	1.29E-03	KIAA0319	-13.505	4.050	9.21E-04
TEKT2	-14.806	5.174	4.40E-03	ENPP7	-10.103	3.170	1.53E-03	NLRP1	9.580	2.892	9.95E-04
SON	11.350	4.059	5.37E-03	ANXA11	-10.758	3.377	1.54E-03	GLRA4	-14.901	4.543	1.12E-03
PRKRIR	10.678	3.864	5.94E-03	ZNF708	-16.047	5.089	1.72E-03	GRTP1	13.852	4.246	1.18E-03
SYNPO2	13.056	4.739	6.09E-03	ZNF462	10.940	3.488	1.81E-03	MUT	-10.783	3.320	1.25E-03
ZNF530	-15.063	5.469	6.11E-03	NCF2	12.165	3.883	1.84E-03	ANXA11	-10.806	3.373	1.45E-03
KIAA0319	-15.015	5.471	6.29E-03	SCYL1	-13.663	4.464	2.34E-03	ENPP7	-10.129	3.169	1.49E-03

Table 4.14: Linear regression (corrected residual) of the entire exome (MAF 0.1, 1 and 2%), no weighting. Indicated are the β, *SE* and *p* values for the top 15 genes using linear burden regression in Hail. Exome-wide significant genes which survive multiple testing correction are emboldened. The corrected residual from MiSeq (polyglutamine-2) was used for regression. Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were not weighted. Indicated at the top of each column are the adjusted (adj) values for each MAF cut-off, which is a count of how many genes/ORFs were tested that resulted in a *p* value and have >5 variants at the MAF tested. N=485. *B*: unstandardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

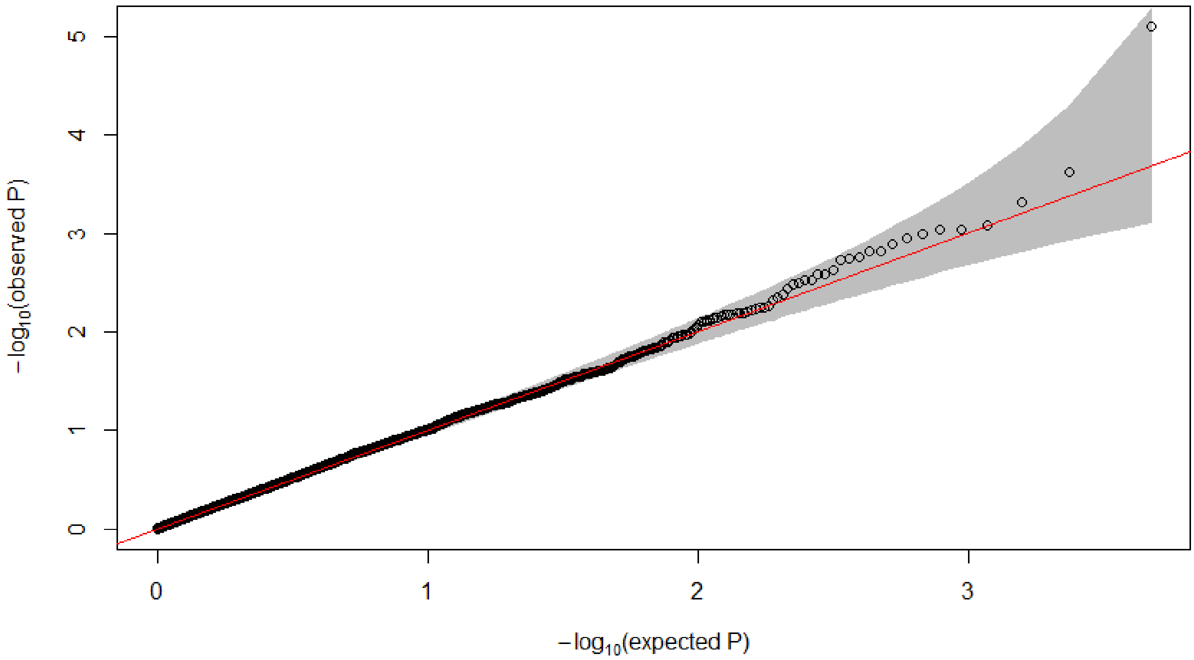


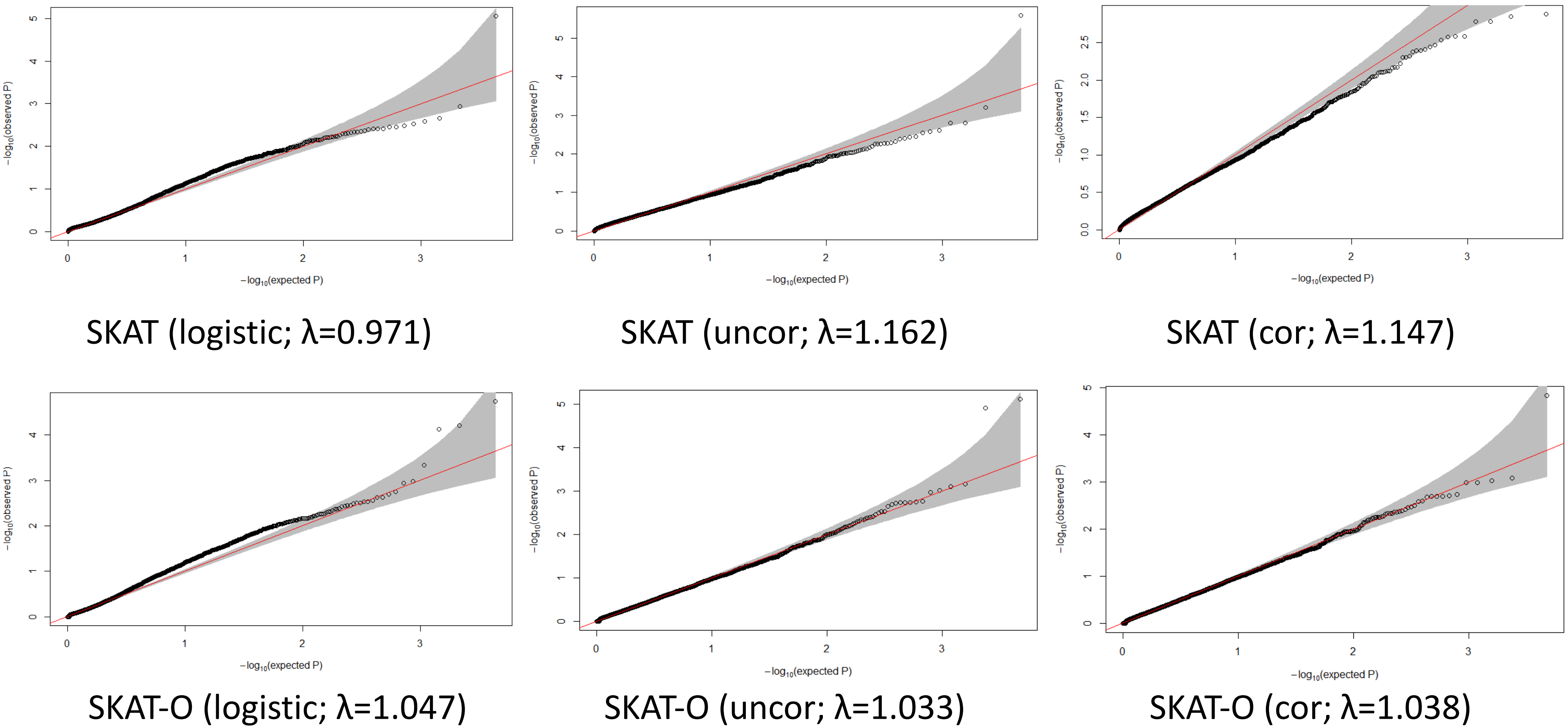
Figure 4.20: Q-Q plot of whole-exome linear regression (corrected residual). *p* values from unweighted whole-exome linear regression using corrected MiSeq residual. Data is taken from those presented in Table 4.14 for MAF 1%. Genomic inflation factor (λ)=1.047. Genes plotted had >5 variants NSD variants at MAF 1%.

4.9 SKAT and SKAT-O analyses

Burden regression is a useful method for summing variants across a gene or gene-set (pathway), however, one of its critical disadvantages is in considering genes/gene-sets with many non-associated variants or bidirectionality, *i.e.* where separate variants in a gene may be associated with opposite directions of effect. This is especially important in a gene set/pathway analysis where genes may act in different directions. To overcome this and to explore further exome-wide analytical methods, we used the SNP-set kernel association test (SKAT) (Wu et al., 2011; Lee et al., 2012e). SKAT can tolerate non-causal variants more effectively than standard burden tests (*i.e.* less affected by the presence of variants with no effect). We also use the optimised SKAT (SKAT-O) (Lee et al., 2012e), which integrates SKAT with a burden test. We ran these with the same covariates as before, PC1-5, BVR and mean sample depth, and used similar filters with missingness ≤25% and MAF=1% gnomAD (see 2.7.4). We elected to use MAF 1% as our primary cut-off as it was the lowest MAF where we had been able to detect substantial signal in our burden tests.

Q-Q plots for the six SKAT and SKAT-O tests are shown in Fig. 4.21. As indicated by the genomic inflation factors (λ), SKAT has slightly inflated *p* values in the linear models, whereas SKAT-O behaves appropriately for both logistic and linear tests, indicating SKAT-O is the best performing test.

FAN1, *PMS1*, *PMS2*, *LIG1* and *EXO1* are all nominally significant in at least one SKAT(-O) test, with *FAN1* having the lowest *p* value of these as before (Table 4.15). Both SKAT logistic and SKAT and SKAT-O continuous analyses with the polyglutamine-2 residual identified *NOP14* as exome-wide significant (*p*=2.58E-06 – 8.83E-06; Table 4.16), and the logistic SKAT-O for *NOP14* is nearly exome-wide significant (*p*=1.83E-05). As with the linear burden regression, applying SKAT/SKAT-O to the corrected CAG length residual (pure CAG) ablates most, although not all, of the *NOP14* signal (*p*=2.68E-03 SKAT; 5.22E-03 SKAT-O). No other genes pass exome-wide significance, however *CUBN* comes close in both the uncorrected and corrected residual continuous SKAT-O analyses (Bonferroni threshold *p*=1.16E-05 (4737 genes); *CUBN* *p*=1.83E-05 and *p*=2.13E-05). Notably, *CUBN* has little to no signal in both SKAT continuous tests (*p*=4.02E-02 uncorrected; *p*=3.27E-02 corrected), possibly due to having a large number of rare, singleton NSD variants driving the signal in burden regression (variants available in Appendix 7). As before, it is also possible to weight variants on deleteriousness. This time, we were able to weight classes of variants differently; NS variants missing CADD annotations were imputed as CADD PHRED = 20, whereas LoF variants were imputed as CADD PHRED = 30 (Appendix 14).



SKAT (logistic; λ=0.971)

SKAT (uncor; λ=1.162)

SKAT (cor; λ=1.147)

SKAT-O (logistic; λ=1.047)

SKAT-O (uncor; λ=1.033)

SKAT-O (cor; λ=1.038)

Figure 4.21: Q-Q plots for SKAT and SKAT-O analyses. *p* values from SKAT and SKAT-O whole-exome regressions are plotted, for genes with >5 variants. Data is taken from Tables 4.15 and 4.16. Genomic inflation factors (λ) are shown on the corresponding Q-Q plot. Filters used (for variants): MAF (1% only), NSD (LoF or CADD PHRED ≥20), missingness ≤25%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were not weighted. Genes plotted had >5 variants. Linear regression used the N=485 continuous HD exomes and logistic the N=440 dichotomous group. Uncor: linear regression on the uncorrected (polyglutamine-2) AMO residual; Cor: Linear regression on the corrected (pure CAG) AMO residual; BVR: baseline variant rate.

	Cor (SKAT-O)	Uncor (SKAT-O)	Cor (SKAT)	Uncor (SKAT)	Logistic (SKAT-O)	Logistic (SKAT)
	p	*p*	*p*	*p*	*p*	*p*
EXO1	**2.82E-02**	6.21E-02	**4.47E-02**	6.53E-02	1.08E-01	1.13E-01
FAN1	**4.39E-03**	**6.62E-03**	**6.71E-03**	**9.58E-03**	**1.52E-02**	**2.55E-02**
HTT	8.52E-01	8.34E-01	8.86E-01	9.41E-01	7.55E-01	7.63E-01
LIG1	5.60E-02	**3.97E-02**	**3.17E-02**	**2.24E-02**	9.40E-02	5.40E-02
MLH1	7.99E-01	6.40E-01	8.55E-01	8.27E-01	8.68E-01	8.67E-01
MLH3	1.00E+00	8.99E-01	8.52E-01	7.36E-01	1.00E+00	8.04E-01
MSH3	7.94E-01	8.62E-01	5.84E-01	6.66E-01	4.27E-01	2.84E-01
OGG1	1.14E-01	1.12E-01	7.25E-02	7.26E-02	3.24E-01	2.32E-01
PMS1	**3.91E-02**	**4.84E-02**	6.16E-02	7.35E-02	**1.37E-02**	7.41E-02
PMS2	**4.57E-02**	**2.76E-02**	3.86E-01	2.82E-01	2.02E-02	2.17E-01
RRM2B	N/A	N/A	N/A	N/A	N/A	N/A
SYT9	1.40E-01	1.55E-01	1.40E-01	1.55E-01	7.58E-02	7.58E-02
TCERG1	3.63E-01	4.01E-01	2.65E-01	2.96E-01	9.05E-01	7.53E-01

Table 4.15: SKAT and SKAT-O dichotomous and continuous tests. Indicated are the *p* values calculated from SKAT and SKAT-O for candidate genes examined in 4.5. Genes and MAFs which are nominally significant are emboldened (none of these pass multiple testing correction: logistic Bonferroni threshold *p*=1.16E-05 (4307 genes); linear Bonferroni threshold *p*=1.06E-05 (4737 genes). Filters used (for variants): MAF (1%), NS damaging (LoF or CADD PHRED ≥20), missingness ≤25%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were not weighted and no total variant number filter was included for target genes. Linear regression used the N=485 continuous HD exomes and logistic the N=440 dichotomous group. Uncor: linear regression on the uncorrected (polyglutamine-2) AMO residual; Cor: Linear regression on the corrected (pure CAG) AMO residual; BVR: baseline variant rate.

Cor (SKAT-O)		Cor (SKAT)		Uncor (SKAT-O)		Uncor (SKAT)		Logistic (SKAT-O)		Logistic (SKAT)	
Gene	*p*	Gene	*p*	Gene	*p*	Gene	*p*	Gene	*p*	Gene	*p*
CUBN	1.47E-05	MUT	1.32E-03	**NOP14**	**7.60E-06**	**NOP14**	**2.58E-06**	NOP14	1.83E-05	**NOP14**	**8.83E-06**
MUT	8.40E-04	CRAMP1L	1.43E-03	CUBN	1.21E-05	TEKT1	6.23E-04	NUP210L	6.18E-05	NDOR1	1.18E-03
SIPA1L2	9.49E-04	DENND4B	1.64E-03	TEKT1	6.84E-04	FBP2	1.58E-03	CUBN	7.55E-05	TEKT1	2.18E-03
ERAP2	1.03E-03	GLRA4	1.64E-03	MUT	7.93E-04	MUT	1.60E-03	ERAP2	4.61E-04	DENND4B	2.57E-03
GLRA4	1.04E-03	ZNF708	2.59E-03	SIPA1L2	9.64E-04	CRAMP1L	2.46E-03	KIAA0319	1.03E-03	TRPM4	3.02E-03
ANXA11	1.86E-03	FBP2	2.62E-03	ERAP2	1.06E-03	DENND4B	2.67E-03	ZNF462	1.14E-03	PEG3	3.27E-03
KIAA0319	1.97E-03	NOP14	2.68E-03	CACNA1G	1.72E-03	GLRA4	2.91E-03	MAMDC2	1.81E-03	BRCA2	3.53E-03
CACNA1G	2.01E-03	OR2B11	2.91E-03	ANXA11	1.80E-03	ZNF708	3.56E-03	CACNA1G	2.01E-03	SSH2	3.54E-03
GRTP1	2.03E-03	UNC5B	3.38E-03	GLRA4	1.83E-03	PRKRIR	3.74E-03	C9	2.35E-03	ETV7	3.84E-03
ZNF708	2.04E-03	CDC20B	3.64E-03	ENPP7	1.85E-03	UNC5B	3.99E-03	NDOR1	2.38E-03	UNC5B	3.91E-03
DENND4B	2.11E-03	HGFAC	3.86E-03	ZNF462	1.87E-03	OR2B11	4.21E-03	GRTP1	2.80E-03	C9	3.93E-03
ENPP7	2.51E-03	NSFL1C	4.04E-03	KIAA0319	1.88E-03	SLC38A2	4.79E-03	SIPA1L2	2.93E-03	C2CD3	4.04E-03
CRAMP1L	2.59E-03	PRKRIR	4.07E-03	GRTP1	2.02E-03	FGL1	5.11E-03	TEKT1	2.98E-03	TULP1	4.30E-03
NCF2	3.13E-03	AC012313.1	4.21E-03	FBP2	2.25E-03	FAM198A	5.33E-03	DENND4B	3.10E-03	FGL1	4.39E-03
OR2B11	3.43E-03	TEKT1	4.83E-03	ZNF708	2.92E-03	NRIP3	5.38E-03	TGM3	0.003115	ASIC4	4.58E-03

Table 4.16: Whole-exome dichotomous and continuous SKAT and SKAT-O tests. Shown are the *p* values calculated from SKAT and SKAT-O for the top 15 genes. Genes passing exome-wide significance are emboldened (logistic Bonferroni threshold *p*=1.16E-05 (4307 genes); linear Bonferroni threshold *p*=1.06E-05 (4737 genes)). Filters used (for variants): MAF (1%), NS damaging (LoF or CADD PHRED ≥20), missingness ≤25%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were not weighted, and only genes with >5 NSD variants were included. Linear regression used the N=485 continuous HD exomes and logistic the N=440 dichotomous group. Uncor: linear regression on the uncorrected (polyglutamine-2) AMO residual; Cor: Linear regression on the corrected (pure CAG) AMO residual; BVR: baseline variant rate.

4.10 Pathway analysis

After investigating whole-exome analyses where we collapsed variants based on genes, we then asked whether rare damaging variation in gene sets was enriched in early or late onset HD. To examine this, we performed a pathway analysis using gene sets taken from the Gene Ontology Consortium (The Gene Ontology Consortium, 2019). Fisher's method was used to combine *p* values across genes in each pathway. This was performed on continuous SKAT, continuous SKAT-O and unweighted linear burden regression analyses at MAF 1% on NSD variants (CADD ≥20 and MAF ≤1% gnomAD) – see methods 2.7.5. An advantage of the Fisher method is it does not make assumptions about the directionality of effects, which is useful if genes in gene sets/pathways have effects in different directions. In all three tests, we used the corrected (pure CAG length) age at motor onset residual.

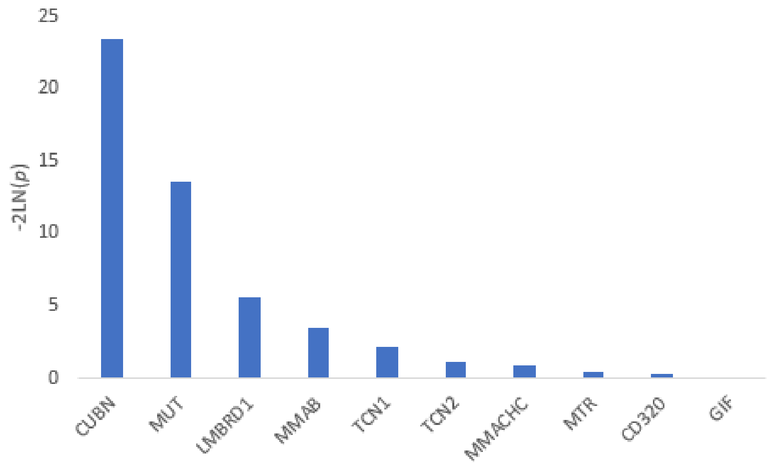
The results from the pathway analysis are presented in Tables 4.17. No pathways survive multiple testing correction (GO pathways with *p* values=7764; Bonferroni threshold *p*=6.44E-06), although in both the linear burden and SKAT-O tests cobalamin binding (GO:0031419) had the lowest *p* value. This observation is mostly driven by the contributions of *MUT* and *CUBN* variants, which constitute ~70% of the observed signal. As expected, SKAT has a less significant *p* value for the cobalamin pathway (*p*=4.12E-02) as *CUBN* has a less significant *p* value in SKAT. The endonuclease activity pathway (GO:0016803), which has *p*=4.71E-04 in linear burden regression, has *FAN1* and *EXO1* as pathway members, although these only contribute ~20% of the signal. The top two pathways from SKAT-O continuous analyses (from Table 4.17B) are shown in Fig. 4.22.

A	GO Pathway	Genes with p	Total genes in pathway	*p*	Term
	GO:0031419	10	10	1.64E-04	cobalamin binding
	GO:0016893	29	40	4.71E-04	endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 5'-phosphomonoesters
	GO:0042359	10	12	8.28E-04	vitamin D metabolic process
	GO:0045211	177	226	1.01E-03	postsynaptic membrane
	GO:0005245	36	40	1.22E-03	voltage-gated calcium channel activity
	GO:0098794	321	423	1.31E-03	postsynapse
	GO:0019203	9	10	2.15E-03	carbohydrate phosphatase activity
	GO:0050308	9	10	2.15E-03	sugar-phosphatase activity
	GO:0033017	32	39	2.20E-03	sarcoplasmic reticulum membrane
	GO:0016176	7	10	2.23E-03	superoxide-generating NADPH oxidase activator activity
	GO:0009235	19	21	2.42E-03	cobalamin metabolic process
	GO:0033013	48	57	2.51E-03	tetrapyrrole metabolic process
	GO:2001235	116	176	2.52E-03	positive regulation of apoptotic signaling pathway
	GO:0070509	28	34	2.92E-03	calcium ion import
	GO:0010803	41	59	2.92E-03	regulation of tumor necrosis factor-mediated signaling pathway

B	GO Pathway	Genes with p	Total genes in pathway	*p*	Term
	GO:0031419	10	10	2.42E-04	cobalamin binding
	GO:0042359	10	12	3.95E-04	vitamin D metabolic process
	GO:0042953	11	16	8.25E-04	lipoprotein transport
	GO:0044872	11	16	8.25E-04	lipoprotein localization
	GO:0019203	9	10	2.31E-03	carbohydrate phosphatase activity
	GO:0050308	9	10	2.31E-03	sugar-phosphatase activity
	GO:0086067	9	10	4.23E-03	AV node cell to bundle of His cell communication
	GO:2001267	18	23	5.49E-03	regulation of cysteine-type endopeptidase activity involved in apoptotic signaling pathway
	GO:0016893	32	40	5.74E-03	endonuclease activity
	GO:1902652	98	115	6.16E-03	secondary alcohol metabolic process
	GO:0006775	26	33	7.06E-03	fat-soluble vitamin metabolic process
	GO:0050711	11	14	7.61E-03	negative regulation of interleukin-1 secretion
	GO:0016176	8	10	8.21E-03	superoxide-generating NADPH oxidase activator activity
	GO:0008203	95	110	8.46E-03	cholesterol metabolic process
	GO:0050713	8	11	1.01E-02	negative regulation of interleukin-1 beta secretion

C	GO Pathway	Genes with p	Total genes in pathway	*p*	Term
	GO:2001267	18	23	2.62E-03	regulation of cysteine-type endopeptidase activity involved in apoptotic signaling pathway
	GO:0050711	11	14	6.71E-03	negative regulation of interleukin-1 secretion
	GO:0019203	9	10	7.24E-03	carbohydrate phosphatase activity
	GO:0050308	9	10	7.24E-03	sugar-phosphatase activity
	GO:0070059	25	30	8.64E-03	intrinsic apoptotic signaling pathway in response to endoplasmic reticulum stress
	GO:0032692	21	25	9.47E-03	negative regulation of interleukin-1 production
	GO:0055106	13	17	1.04E-02	ubiquitin-protein transferase regulator activity
	GO:0050713	8	11	1.25E-02	negative regulation of interleukin-1 beta secretion
	GO:0090502	54	72	1.33E-02	RNA phosphodiester bond hydrolysis
	GO:0005540	19	21	1.50E-02	hyaluronic acid binding
	GO:0016893	32	40	1.77E-02	endonuclease activity
	GO:0005031	18	24	2.02E-02	tumor necrosis factor-activated receptor activity
	GO:0005035	18	24	2.02E-02	death receptor activity
	GO:0002664	9	10	2.19E-02	regulation of T cell tolerance induction
	GO:0004521	46	58	2.38E-02	endoribonuclease activity

Table 4.17: Pathway analysis of NSD variants. *p* values were taken from whole-exome linear burden (A) and whole-exome continuous SKAT-O (B) and SKAT (C) tests at MAF=1% examining NSD variants. *p* values were combined using Fisher's method using pathways taken from the Gene Ontology (GO) database. Pathways with *p* value N=7764.



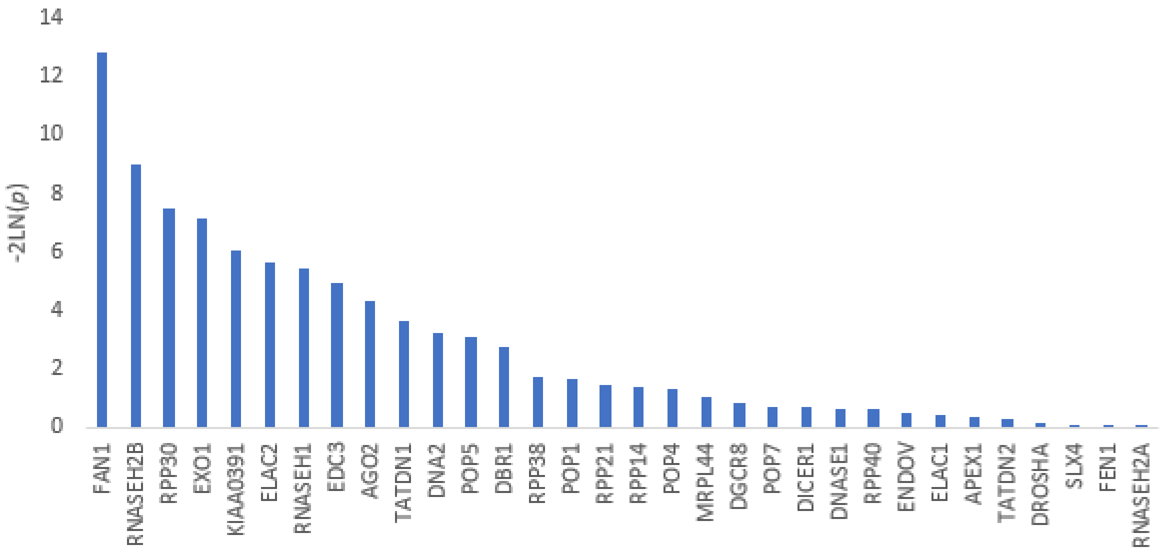


Figure 4.22: Members of the top pathways from gene set analysis. Data from continuous linear burden pathway analysis (Table 4.17A) are shown for the top two pathways (top: cobalamin binding (GO:0031419); bottom: Endonuclease activity (GO:0016893)).

4.11 Discussion

4.11.1 Overview of results

Chapter 4 detailed the development of an in-house bioinformatic pipeline for the QC and annotation of WES data from the individuals chosen for sequencing in chapter 3. Interrogation of the *HTT* CAG repeat in these samples identified atypical structures strongly associated with altered HD onset. These structures will go on to be more rigorously examined in chapter 5 using MiSeq. Further, we identified NSD variation in several genes implicated in HD onset modification including *FAN1*, *EXO1* and *MSH3*. NSD variants in other genes such as *PMS1* and *LIG1* were also identified. In addition, shorter Q-A tracts in *TCERG1* were found to be associated with a later disease onset. Finally, we used whole-exome burden regression and SKAT(-O) analyses to detect novel variation occurring in the exome associated with altered onset. These tests identified *NOP14*, a proxy for *HTT* allele structure, as exome-wide significant. *CUBN* was found as exome-wide significant using burden regression and was nominally significant using SKAT-O.

4.11.2 Development of in-house pipelines for exome analysis

We used a standard GATK best practices pipeline for initial alignment and variant discovery (McKenna et al., 2010), and then extended on this with our own custom sample QC and variant annotation pipelines for downstream analyses. We loosely based the sample QC on the ExAC/gnomAD studies (Lek et al., 2016; Karczewski et al., 2019), with some modifications, such as the use of Hail as this was unavailable during the time of the ExAC study. 495 exomes passed initial QC and 486 of these passed relatedness checks. One problem we did encounter was some of our cohort had different CAG lengths when re-genotyped with MiSeq compared with our original patient database. This was a consequence of using locally derived CAGs (explored in chapter 5) and was overcome by redefining early and late onset groups using CAG lengths from MiSeq. However, this issue did reduce the number of extreme onset individuals originally selected and slightly weakened the power of the study (see also 6.2). CAG length discrepancies are less likely in future studies as centrally measured CAG lengths are now available for a substantial portion of Registry through BioRep, and these should be more accurate and consistent than local diagnostic labs. Furthermore, the new Enroll-HD study, which superseded Registry, has collected data more rigorously, including centralised CAG length determination (Landwehrmeyer et al., 2016).

In our primary whole-exome burden analyses, we explored three main MAF cut-offs: very rare (MAF≤0.1%), rare (MAF≤1%) and uncommon (MAF≤2%) frequencies. We elected to use MAF 1% in our candidate gene analyses and whole-exome analyses (burden and SKAT(-O)) as it seemed to offer a reasonable balance between common variation, likely picked up by GWAS, and very rare variation which we were underpowered to detect at the whole-exome level. MAF 1% has also been suggested in other studies as a rarity cut-off (Li et al., 2011; Agarwala et al., 2013; Auer and Lettre, 2015; Bomba et al., 2017; Rees et al., 2019), but other studies have defined rare variation as ≤0.5% MAF (Cirulli and Goldstein, 2010; Mistry et al., 2015; Nagasaki et al., 2015; Dekker et al., 2019; Flannick et al., 2019). Our study was limited by its N (~500 total exomes), and we found that 0.1% MAF led to much higher *p* values for most genes; however, using the 1% MAF identified several variants in *FAN1* already been picked up by GWAS, namely Arg507His and Arg377Trp (GeM-HD Consortium, 2019). Future studies with a larger number of exomes may find a smaller MAF threshold (*e.g.* ≤0.5%) useful for finding entirely novel variation not at all captured by GWAS in HD.

Another element of the analytical pipeline that may need to be considered in further exome analysis, especially if using Enroll-HD, is that of ancestry. Although we did calculate ancestry and included principal components into our whole-exome analyses, candidate gene analyses should take ancestry into account if a substantial part of the sequenced population is of a non-European decent. Additionally, it is important to consider how to handle diverse ancestries. Here, we included all individuals regardless of ancestry; however, principal components alone cannot effectively account for more diverse ancestries where there are only a small number of individuals, such as we observe in the current study (see 4.2.4). Auxiliary analyses (see Appendix 15) demonstrate the removal of non-European ancestries did not substantially change the results found this study, although, *CUBN* was no longer exome-wide significant following multiple testing correction. Still, future analyses will benefit from the removal of rarer ancestries if there are an insufficient number of individuals of these ancestries for principal component analysis to be effective. Finally, other MAFs for different populations (*i.e.* not just non-Finnish Europeans) may need to be considered and integrated into future work. Linear mixed models could be used in whole-exome analyses to account for population structures, and are supported in the most recent version of Hail (v0.2).

For annotation, we used dbNSFP v3.0 (Liu et al., 2011, 2016) and added CADD (Kircher et al., 2014; Rentzsch et al., 2019), SIFT (Sim et al., 2012; Vaser et al., 2016) and Polyphen2 (Adzhubei et al., 2010) scores to our variants. We primarily focused on CADD score as it amalgamates many different scores and has been used in numerous studies, databases and tools (Liu et al., 2016; McLaren et al., 2016; Nissen et al., 2018; Cunningham et al., 2019; Sandri et al., 2019). The developers of the CADD score recommend using a scaled CADD

PHRED score cut-off of between ≥10-20; we opted to use a fixed score of ≥20, which describes the top 1% damaging variants in the genome. However, while CADD≥20 is effective for identifying possible damaging variation (Itan et al., 2016), its stringency may remove disease-relevant variants, especially given the variants may only alter the function of the resultant protein, and are not necessarily LoF. Thus, future study may find different CADD cut-offs useful (*e.g.* CADD PHRED 15), or, equally, un-scaled CADD scores could be used. Additionally, using other scores or 'meta-scores' (*e.g.* DANN, (Quang et al., 2015)) for variants may also be worth pursuing. Finally, annotation of functional domains using InterPro (Mitchell et al., 2019), or similar, would allow for variants in specific domain functional families to be aggregated across the exome, and this may inform downstream outcomes.

4.11.3 Short tandem repeats in *HTT*, *MSH3* and *TCERG1*

Short tandem repeats (STRs) are repetitive regions of the genome and are intrinsically difficult to read through with current second-generation NGS technologies, such as have been used in this study (Illumina 75bp paired end), due to STRs length and repetitive nature (discussed later generally in 6.2.3 and 6.4.1). Thus, variants from *MSH3* and *HTT* STRs tended to be inaccurate when called in our GATK-based pipeline. To overcome this, we extracted reads containing the *HTT* CAG repeat and were able to identify structural differences strongly associated with altered HD onset. These non-canonical *HTT* alleles have very recently been identified in three independent studies (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019) and constitute a major modifier of onset in a subset of patients. The atypical alleles we identify will be discussed in much greater detail in chapter 5.

Manual assessment of reads in *MSH3* was not as successful, mainly as the entirety of the repeat structure had to be assessed as opposed to just the 5'- and 3'- ends as in *HTT*. The polymorphic Pro/Ala *MSH3* repeat in *MSH3* ranges from ~3aa (27bp) to ~9aa (81bp), not including further repetitive flanking sequence (Coleman et al., 1994; Nakajima et al., 1995). We were able estimate the genotype lengths, although there was no significant association between *MSH3* genotype length and residual age at onset (*p*=0.359). Recent study has identified the *MSH3* repeat structure as being associated with altered onset in both HD and DM1 (Flower et al., 2019). Notably our (non-significant) direction of effect matches that seen in the Flower study where the shorter *MSH3* allele was associated with later disease onset in HD and attenuated somatic expansion. How the *MSH3* variant is affecting onset is unclear; it may be STRs are simply tagging non-coding modifiers (*e.g.* eQTLs) which are then driving onset modification through expression changes in *MSH3* (Flower et al., 2019). Additionally, the EXO1 interaction domain in MSH3 lies close to the repetitive tract (Schmutte et al., 2001), and this interaction could be in some way modified by the number of repeats in MSH3. There is evidence for *EXO1* as a modifier of HD in this study (discussed in 4.11.5) and as a modifier of somatic instability in a fragile X mouse (Zhao et al., 2018).

Unlike *HTT* or *MSH3*, the imperfect repetitive Q-A tract in *TCERG1* was read-through effectively with WES due to alternating codons throughout its structure. The *TCERG1* locus was significant in the most recent HD GWAS (GeM-HD Consortium, 2019), and the repeat itself in *TCERG1* has been implicated previously as an HD modifier (Holbert et al., 2001; Arango et al., 2006). Linear regression analysis in our data showed a significant negative effect of the STR on onset (*p*=2.5E-03), *i.e.* later onset is associated with shorter Q-A tracts in *TCERG1*. More work is needed to uncover the specific mechanism of action of *TCERG1*; however, our study lends further support that *TCERG1* is likely the gene responsible for the signal at this locus, and not *GPR151* where no clear excess of variation between early and late onset individuals was observed. A possible mechanism (altering the splicing and/or transcription of *HTT*) by which TCERG1 may modify onset is discussed later in 6.3.4.

4.11.4 NSD variation in FAN1

FAN1 has 5’ flap endonuclease and 5’-3’ exonuclease activity, and is known to be involved in interstrand crosslink (ICL) repair (Kratz et al., 2010; Liu et al., 2010b; MacKay et al., 2010; Smogorzewska et al., 2010). FAN1 interacts with mismatch repair (MMR) proteins including MLH1, MLH3 and PMS2 (Smogorzewska et al., 2010), and biallelic LoF or very damaging missense mutations in *FAN1* cause the kidney disease karyomegalic interstitial nephritis (KIN) in humans (Zhou et al., 2012) and in mice (Airik et al., 2016; Lachaud et al., 2016; Thongthip et al., 2016). KIN is marked by karyomegaly in kidney and progressive renal decline (Isnard et al., 2016; Hard, 2018), although there is evidence for karyomegaly in other tissues including the liver and brain (Spoendlin et al., 1995). In HD, *FAN1* has now implicated by genetic study as a modifier of disease onset (GeM-HD Consortium, 2015, 2019), and functional work has suggested a protective role of the protein in somatic expansion in fragile X and HD models (Zhao and Usdin, 2018; Goold et al., 2019). *FAN1* may also have a modifying role in other repeat diseases (Bettencourt et al., 2016).

Using WES, we identified an overall enrichment of NSD variation associated with an early HD onset (26:9 E:L; 29:9 from later Sanger sequencing confirmation). In addition, we also identified a very rare (MAF=6.17E-05) LoF frameshift variant in an early onset HD patient. Early associated variants seemed to mostly fall into two primary clusters: a cluster centred near Arg507His in the SAF-A/B, Acinus and PIAS (SAP) domain, involved in DNA binding and recruitment of FAN1 to sites of damage (Gwon et al., 2014; Zhao et al., 2014; Thongthip et al., 2016), and variants on the C-terminal end of the protein in the virus-type replication-repair nuclease (VRR-Nuc). The Arg377Trp variant and Leu395Pro variants are also nearby the former cluster. Crucially, both early variant clusters appear in areas of the protein that contact DNA (Wang et al., 2014; Jin and Cho, 2017). We also find an additional smaller cluster in the tetratricopeptide repeat (TPR) domain associated with late onset. Notably both the Arg507His (MAF=0.96%) and Arg377Trp (MAF=0.72%) have been identified now by GWAS as modifiers (GeM-HD Consortium, 2019), however ~1/3 of the variants (11 singletons) we identify here are very rare (≤0.1% MAF) and have not been previously been associated with HD. In addition, we identify at least two variants which, to our knowledge, are novel (Asp498Asn and Asp702Glu).

Variants in the VRR-nuclease domain are of particular interest as missense variants in this domain can cause KIN (Zhou et al., 2012). Indeed, the Arg982Cys variant may be especially damaging as it has a very high CADD score and is in a DNA binding domain near the active site at a position highly conserved between human and bacteria (Yan et al., 2015). Asp981-Arg982 mutated FAN1 has substantially reduced endonuclease activity (MacKay et al., 2010). Goold and colleagues (Goold et al., 2019) recently found inactivation of FAN1's nuclease activity through the Asp960Ala missense mutation did not affect the protein's ability to stabilise the *HTT* CAG repeat in an HD cell model (Goold et al., 2019). However, our study clearly suggests functionality of the nuclease domain is still important, which is perhaps unsurprising given the nuclease's importance in other *FAN1* biology including ICL repair (MacKay et al., 2010; Yoshikiyo et al., 2010; Thongthip et al., 2016; Jin et al., 2018), replication fork restart (Shereda et al., 2010; Chaudhury et al., 2014; Porro et al., 2017) and KIN (Zhou et al., 2012). Hence damaging mutations in the VRR-Nuc domain in HD patients may compromise the protective effect of *FAN1* in HD and lead to an accelerated disease onset.

TPR domains are found in a large number of proteins and often have roles in protein-protein binding (Perez-Riba and Itzhaki, 2019). FAN1's TPR domain is involved in the dimerisation of the protein (Zhao et al., 2014). The identification of a cluster of late-associated variants in this domain, then, is interesting, especially as these variants are all very rare or novel (MAF≤0.1%). It is possible these variants could affect FAN1 dimerisation or other protein-protein interaction (Jin and Cho, 2017). Potentially, as suggested by Pennell and colleagues (Pennell et al., 2014), this dimerisation could affect the substrate specificity of FAN1. Changes in the capability of FAN1 to dimerise could affect FAN1's ability to resolve *HTT* CAG repeat secondary structures. It is equally possible these late-associated variants could affect DNA binding in some way as these are still close to predicted DNA contact points (Wang et al., 2014; Zhao et al., 2014), and the TPR domain is known to bind post-nick 5'-flap DNA.

None of the variants we identify have been formally associated with KIN (ClinVar, accessed July 2019), however this is likely as KIN is a rare autosomal recessive disease and both copies of *FAN1* would have to be lost for KIN to occur. Indeed, the LoF variant we find (Thr187fs) would likely be associated with KIN in biallelic loss of *FAN1*. Interestingly, the Arg507His variant (MAF=0.97%) was associated with KIN in a recent phenotype risk score study (Bastarache et al., 2018), further suggesting that the variant is likely deleterious to FAN1 function. The variant has been implicated as modifying protein-protein interaction effects rather than a direct effect on DNA contact (Jin and Cho, 2017), and the SAP domain, where Arg507His occurs, is involved in dimerisation of FAN1 (Zhao et al., 2014).

4.11.5 NSD variation in *EXO1*

EXO1 encodes exonuclease 1 (EXO1), a 5'-3' exonuclease involved in many facets of DNA maintenance and repair, including MMR (Genschel et al., 2002; Tran et al., 2004) where it binds MLH1, MSH2 and MSH3 proteins (Tishkoff et al., 1997; Schmutte et al., 2001; Dherin et al., 2009). EXO1 also possesses substrate-specific flap endonuclease (Lee and Wilson, 1999) and RNase H activity (Qiu et al., 1999; Liu et al., 2017). Although not genome-wide significant in the most recent GWAS (GeM-HD Consortium, 2019), functional work in a murine model of fragile X identified the gene as a modifier of repeat instability (Zhao et al., 2018). In our exomes, we find an excess of late-associated NSD variation in the MSH3-EXO1 interaction domain, such as the Gly274Arg mutation present in 4 late onset individuals with a high CADD score (34). We also find two further late-associated variants in the MLH1-EXO1 interaction domain, including the more common variant Gly759Glu (MAF=0.93%). No clustering of early or late onset variants, barring one (Ser610Gly), was observed in the MSH2 interaction domain.

Somewhat surprisingly, Zhao et al. find *EXO1* to be protective in their fragile X mouse model (Zhao et al., 2018), whereas our study suggests *EXO1* is normally deleterious. A possible explanation for this is the class of mutation; in the Zhao study, they use knockout *Exo*$^{-/-}$ mice. Comparatively, we only identify clustering associated with NSD variants in EXO1. Hence this might suggest differing effects for *EXO1* depending on whether its functionality is entirely ablated or only modified (through damaging mutations). Interestingly, it has been suggested that FAN1 may act as a compensatory nuclease in the absence or knockdown of EXO1

(Desai and Gerson, 2014), suggesting at least some degree of functional overlap between the two proteins.

Why *EXO1* was not identified through GWAS is unclear, but it may be the result of the rarity of the variants observed. ~70% of the NSD variants we see are very rare (≤0.5% MAF), and array-based imputation methods struggle to effectively impute these rarer alleles (Cirulli and Goldstein, 2010). The Gly759Glu variant is of a similar MAF to *FAN1* Arg507His, however its effect size may be substantially less. Larger GWAS may identify *EXO1* in future work.

4.11.6 LoF variants in MSH3

MSH3 forms a heterodimer with MSH2 (MutSβ), and binds to small and large loop-outs in DNA, up to 14 bp in size *in vitro* (Habraken et al., 1996; Palombo et al., 1996), where it coordinates with other MMR machinery such as MLH1-PMS2 (MutLα), EXO1 and LIG1 – mismatch repair reviewed generally by (Fishel and Lee, 2016; Hsieh and Zhang, 2017). MSH3 has been shown to bind repeat trinucleotide repeat structures (Owen et al., 2005; Burdova et al., 2015; Guo et al., 2016; Lai et al., 2016), and is a modifier of disease onset in HD (Flower et al., 2019; GeM-HD Consortium, 2019) and DM1 (Morales et al., 2016; Flower et al., 2019), and a modifier of HD progression (Hensman Moss et al., 2017).

In our exomes, there seemed to be a slight excess of late-associated variation in MSH3's MutS III domain (lever domain) and early-associated variation in or near the MutS V domain (ATPase domain) (Obmolova et al., 2000; Boland and Goel, 2010; Kumar et al., 2013). However, the clustering of variants to early/late onset was somewhat unclear, and some variants (*e.g.* Pro681Ser and Val682Leu) occurred in late and early individuals. No NSD variants occurred near the EXO1 interaction domain. However, we identify three novel LoF variants in *MSH3* all associated with a late disease onset, highlighting a dose-dependent effect of MSH3 expression on HD onset as shown through imputed transcriptome wide association study (TWAS) in patients (GeM-HD Consortium, 2019). This effect is similar to the MSH3 dose-dependent effect seen on repeat instability in animal models (Foiry et al., 2006; Dragileva et al., 2009; Tomé et al., 2013). Demonstrating LoF mutations occur in late-onset HD patients helps validate *MSH3* as a potential therapeutic target, especially due to its (partial) redundancy in mismatch repair with MutSα (MSH2-MSH6 heterodimer).

4.11.7 WES of other GWAS candidate genes

We also examined other genes implicated as potential modifiers of HD from GWAS (GeM-HD Consortium, 2019) and other functional work (Budworth et al., 2015): *HTT* (independent of its CAG length/structure), *OGG1*, *RRM2B*, *SYT9* and the MMR genes *LIG1*, *MLH1*, *MLH3*, *PMS1* and *PMS2*. The coverage of *RRM2B* was poor and hence no NSD variants were identified. *SYT9*, which was a single rare significant SNP in GWAS (GeM-HD Consortium, 2019), also had lower exome coverage, although we did identify a single NSD variant associated with early onset in this gene, Leu353Val (MAF=0.98%), and this is the same directionality as seen in the GeM GWAS. *SYT9* encodes synaptotagmin 9 involved in calcium sensing and transmission, and is highly expressed in the striatum of mice (Xu et al., 2007), the most relevant disease tissue in HD. NSD variants in *HTT* did not segregate to early or late onset.

MLH1 forms heterodimers with PMS2 (MutLα), PMS1 (MutLβ) and MLH3 (MutLγ). We investigated all four of these genes as *MLH1*, *PMS1* and *PMS2* have all been identified as HD onset modifiers in GWAS (GeM-HD Consortium, 2019). Three NSD variants in *PMS1* were observed in four individuals, all of whom had a late HD onset, suggesting a potential deleterious role for PMS1 in HD onset, and matches the findings of the 2019 GeM GWAS. Neither *PMS2*, *MLH3* or *MLH1* had obvious aggregation of NSD variants with altered onset in our cohort. However, we reproduced the common signal in *MLH1* (Ile219Val) associated with altered onset seen in other genetic studies (Lee et al., 2017; GeM-HD Consortium, 2019). As Ile219Val is common and not predicted to be extremely damaging (13.9 CADD), the variant may be in LD with other factors responsible for modifying onset. Of the four MutL monomers, it is interesting that *PMS1* may harbour the most relevant NSD variants in our study. The function of the MutLβ PMS1-MLH1 heterodimer has been somewhat enigmatic with no MMR activity described (Räschle et al., 1999), so it is possible there may be some novel role the complex plays in repeat resolution alongside its binding partner MutS.

The *LIG1* gene was significant in the HD GWAS (GeM-HD Consortium, 2019), and has broad roles in DNA repair including mismatch repair and base excision repair(Montecucco et al., 1998; Levin et al., 2000). We found a small amount of clustering of early variants in the DNA ligase A N (DNA binding) and DNA ligase A M (nucleotidyltransferase) *LIG1* domains (Pascal et al., 2004), and late-associated variation in the C-terminal DNA ligase A C domain (oligonucleotide-binding fold). LIG1 can promote trinucleotide repeat expansion both *in vitro* (Crespan et al., 2015) and *in vivo* (Subramanian et al., 2005; Tomé et al., 2011). The late-associated variants we identify may affect LIG1’s ability to associate with other repair DNA repair factors. The mechanism of the early variants here are less clear, however; potentially the variants could affect LIG1’s ability to complete repair, and this may lead to a larger amount of instability (a mechanism is discussed in 6.3.3). It is also difficult to know which pathway(s) LIG1 is acting *via*; although we did not find any clear segregation of variants in

OGG1, base excision repair may still be involved in onset determination in HD in addition to mismatch repair.

4.11.8 Whole-exome analyses

Using a combination of continuous (linear) burden, SKAT and SKAT-O tests, we identified two genes as reaching exome-wide significance in at least one test. *NOP14* was exome-wide significant at MAF≤1% in SKAT and SKAT-O tests, although not significant by burden test (*p*=3.42E-03, linear uncorrected AMO residual burden), possibly as SKAT(-O) tends to perform better than burden testing when there is a single skewed variant in a gene (Wu et al., 2011; Lee et al., 2012e). The *NOP14* significance is almost entirely driven by a single late-associated coding variant, Arg697Cys, in high LD with the (CAG**CAA**)$_2$CAG atypical *HTT* repeat structure. As we demonstrate, using the corrected (pure CAG) AMO residual ablates *NOP14*'s signal. >80% (16 of 18 instances) of the (CAG**CAA**)$_2$CAG expanded alleles in our dataset possess the Arg697Cys variant in our data, which is ~133kb away from the CAG repeat tract in *HTT*. Notably, the *HTT* haplotype is not one of the seven most common HD haplotypes (Lee et al., 2012b), and the Arg697Cys variant is instead probably tagging the so-called 4AM2 *HTT* haplotype from GWAS (GeM-HD Consortium, 2019). These data suggest the majority of European (CAG**CAA**)$_2$CAG *HTT* alleles originate from a single common allelic ancestor that has expanded into the pathogenic repeat range many times.

The second exome-wide significant signal was *CUBN. CUBN* encodes Cubilin, involved in vitamin B12 (cobalamin) uptake in the small intestine, and was an unexpected result from our study. Notably, while *CUBN* was exome-wide significant by burden testing, it was only nominally significant by SKAT-O (*p*=1.21E-05) and SKAT (*p*=2.59E-02), (Bonferroni threshold *p*=1.15E-05). *CUBN* had a very large number of NS variants (72 NS variants, see Appendix 7), and this probably reflects the large size of the CUBN protein (3623aa). 23 of these variants were NSD and enriched in early onset individuals (31:7 E:L). It is difficult to know how to interpret these data. Most of the NSD variants were very rare and only singletons, and this is probably why SKAT did not find *CUBN* to be exome-wide significant (Wu et al., 2011; Lee et al., 2012e). Further study is needed with more exomes before attempting to conclude anything from *CUBN*.

Gene set analysis did not identify any significant pathways in our data, although using the linear burden regression *p* values identified cobalamin binding (driven mostly by *CUBN* and *MUT*; GO:0031419) and endonuclease activity (including *FAN1* and *EXO1*; GO:0016893) as the pathways with the lowest *p* values.

Our identification of *NOP14* as an HD modifier (indirectly tagging *HTT* allele structure) proves the utility and tractability of whole-exome rare-variant analyses in HD. It also reflects the remarkably strong effect size of the *HTT* allele structure on HD onset, which was exome-wide significant despite our relatively small sample set (N=500; discussed later in 6.2.2 and & 6.2.4). It is likely that with more HD exomes (ideally >1,000), we will have the power to detect rare variation exome-wide in DNA repair genes, and possibly in novel genes/gene sets (see 6.4.1 for future extensions to sequencing work). The following final results chapter will examine the *HTT* allele structure identified originally through WES in greater depth using MiSeq sequencing. The ramifications of these findings will then be discussed.

Chapter 5: Using MiSeq to investigate structure and instability of the *HTT* CAG repeat

5.1 Introduction

As demonstrated in chapter 4, we found expanded *HTT* allele sequence was a major modifier of disease onset in a subset of HD patients. A variant in *NOP14*, Arg697Cys, was additionally found to be exome-wide significant, and this appears to be strongly tagging the expanded atypical *HTT* $\mathrm{CAG_n(CAG\textbf{CAA})_2}$ allele. Our observation that CAG sequence of *HTT* is an important disease modifier is similar to those reported in HD recently, where variants in the repeat structure were associated with altered HD onset (Ciosi et al., 2019; GeM-HD Consortium, 2019) and HD penetrance (Wright et al., 2019).

A growing body of evidence across repeat disease points to allele structure as an important modifier of disease onset, disease penetrance and intergenerational transmission. For instance, CGG repeat expansions in the X-linked *FMR1* gene cause Fragile X; AGG interruptions toward the 5' end of the repeat confer intergenerational repeat stability (Eichler et al., 1994; Yrigollen et al., 2012; Nolin et al., 2015), and alleles without interruptions were enriched in those with disease (Falik-Zaccai et al., 1997). Various interrupted alleles have been described in several other repeat diseases including spinal cerebellar ataxia (SCA) type 1 (Chung et al., 1993; Chong et al., 1995; Menon et al., 2013), SCA2 (Choudhry et al., 2001), SCA10 (Matsuura et al., 2006; McFarland et al., 2014), and Friedreich's ataxia (Pearson et al., 1998). More recently, complex interruption arrays in myotonic dystrophy type 1 (DM1) were found to be associated with delayed onset, attenuated disease phenotype and reduced somatic expansion rates (Musova et al., 2009; Braida et al., 2010; Cumming et al., 2018; Pešović et al., 2018).

Hence, the finding that *HTT* allele structure plays a role in HD onset is consistent with other repeat diseases. However, there were several limitations in the calling of *HTT* repeat structure in our exomes: (1) coverage was found to vary substantially, and it was not possible to call repeats in all cases, (2) phasing (*i.e.* determining which allele, expanded or wild-type, harboured alternate *HTT* alleles) was not always possible and (3) we did not have an automated repeat calling pipeline, making structural calling slow and prone to error. Therefore, we wanted to confirm and dephase the repeat structures we identified in WES using an independent and systematic method.

There are several ways to assess CAG length in HD (Massey et al., 2018), with the current gold standard being capillary electrophoresis of repeat-containing PCR products using genescan (Applied Biosystems). Genescan uses fluorescent PCR to estimate *HTT* repeat size and quantify somatic instability. Genescan also has the advantage of being medium throughput as it is plate-based, and a typical genescan run can accommodate 48-96 samples. However, crucially, genescan does not determine repeat sequence, and instead approximates pure CAG size based on a ‘typical’ *HTT* sequence with a single interrupting CAA. Interruptions can be detected by triplet-primed PCR coupled with genescan (Chen et al., 2010; Hayward et al., 2016; Hayward and Usdin, 2017), and triplet-primed PCR is often used for clinical applications. But triplet-primed PCR may lack sensitivity in the context of HD where interruptions are very closely spaced, and the tandem interruptions we find here (*e.g.* CAACAA) also may not be detected using triplet-primed PCR. First generation sequencing modalities, *i.e.* Sanger sequencing, can be used to obtain sequence level information, however are low throughput and the necessary gel extraction step or cloning for isolating a single allele is laborious (Massey et al., 2018; Wright et al., 2019).

To address these issues and confirm *HTT* allelic structures from WES, we utilised a targeted next-generation sequencing approach (Ciosi et al., 2018) on the Illumina MiSeq platform to specifically amplify the CAG and CCG repeats in exon 1 of *HTT*. MiSeq is high-throughput, supporting both 96 or 384 library formats (with other plexities possible), and can sequence through repeats up to ~210-240bp (70-80 CAGs) in addition to flanking sequence. Chapter 5 will explore the use of this protocol in our extreme early/late onset patient cohort. First, repeat sequence is called and allelic structures phased using the Scale-HD bioinformatic pipeline (https://scalehd.readthedocs.io). Following this, MiSeq sequencing is validated by comparing data from the same individuals from a locally run genescan. Differences between lymphoblastoid and whole blood DNA are also considered. Finally, we consider how disease phenotype, age at onset and disease progression, may be associated with CAG instability using generalised linear models. Limitations of this modelling approach are also discussed.

5.2 MiSeq sequencing and Scale-HD

We utilised a targeted next-generation MiSeq sequencing method (developed by (Ciosi et al., 2018), see 2.8) to amplify and sequence the repetitive stretches of *HTT* exon 1. All 500 lymphoblastoid (LBC) samples exome sequenced in chapter 4, alongside 49 blood DNA from a subset of these samples, were chosen for sequencing. Seven 96 well plates were prepared and pooled to form two 384 multiplex libraries which were MiSeq sequenced. Raw sequencing files (FASTQ) were aligned and *HTT* structures called with Scale-HD (https://scalehd.readthedocs.io/en/latest/index.html). Scale-HD employs a reference sequence (refseq) alignment using the BWA-MEM alignment algorithm. Reads are aligned to 4000 imputed 'canonical' *HTT* exon 1 alleles with varying CAG and CCG structures. Samples with low alignment scores are re-aligned to 8000 'atypical' *HTT* alleles to determine structure of both wild-type and expanded *HTT* alleles. Scale-HD also produces a variety of other metrics including QC metrics *via* FastQC (Appendix 16), SNPs using freebayes (Garrison and Marth, 2012) and measures of somatic instability and polymerase slippage (Equations 2.3 & 2.4).

645 non-control samples were sequenced (including blood DNA, sequenced twice) as indicated (Table 5.1). Of the 645 samples sequenced, 642 were sequenced successfully (>99%). All three failing samples were lymphoblastoid DNA which failed amplification, however; blood and longitudinal lymphoblastoid DNA was available for one of these samples resulting in structural determination for 498 of the 500 sequenced. In addition, one sample was successfully amplified but only a wild-type allele was seen with MiSeq and subsequent genescan – this individual was removed from downstream analyses here and in chapter 4. Of the 649 samples (including 7 positive controls) successfully sequenced, the average number of mapped reads for expanded alleles was 48,542 and 48,142 for wild-type alleles. An equal balance between expanded/wild-type alleles was achieved using solid phase reversible immobilisation (SPRI) beads for size selection during the library preparation (2.8).

DNA source	Sequenced (total)	Sequenced (successful)
Lymphoblastoid	500	497
Blood	49*	49*
Longitudinal lymphoblastoid	47	47
Positive Control	7	7
Negative Control	7	N/A

Table 5.1: Details of DNA sequenced using MiSeq. Successful sequencing refers to samples where amplification and sequencing were successful. Note, for blood DNA, all DNA was sequenced twice (total 49*2=98). 42 of the 49 blood DNA had matched lymphoblastoid DNA (blood and lymphoblastoid DNA from the same person).

5.3 Determining *HTT* allele structure

5.3.1 Interruptions in the CAG repeat tract

As indicated in WES, *HTT* structure was strongly associated with altered age at onset. Using WES alone, we were originally only able to estimate allele structure in ~80% of cases and could dephase structures in ~65% of cases (4.5). Using MiSeq, we were able to successfully obtain (1) pure CAG length, (2) pure poly-proline length and (3) interruption sequences occurring in the CAG repeat for 498 of the 500 selected individuals. As MiSeq sequencing was deep (~50k per allele), structure calling was robust. As shown (Fig. 5.1 & Table 5.2), 16 different *HTT* allele structures were identified independent of CAG length. The two most common alleles, $(CAG)_n$**CAA**CAGCCGCCA$(CCG)_7(CCT)_2$ and $(CAG)_n$**CAA**CAGCCGCCA$(CCG)_{10}(CCT)_2$, represent 70.0% and 20.2% of the total (N=497) sequenced alleles, respectively. Interestingly, the 10 CCG repeat common allele appears to be less associated with expanded *HTT* than the more common 7 CCG repeat allele.

Notably, while the Scale-HD pipeline was able to detect most allele structures (494 of 498), very rare interruption structures (*e.g.* CAG(**CAA**)$_2$CAG) were miscalled by the pipeline. Hence, all alignments were manually assessed (methods 2.9.1) in addition to automatic calling. MiSeq detected a high rate (11.4%) of atypical CAG interruption structures occurring in our extreme onset group (N=438), with most (>80%) of these occurring on the expanded *HTT* allele. We also identify a novel non-Q interruption (**CAC**) in one late onset individual previously undetected by WES on an expanded allele, as well as confirming and phasing all three CAG(**CAA**)$_2$CAG and CAG(**CAA**)$_3$CAG structures to expanded *HTT* alleles. Before correcting for pure CAG size, the non-Q interruption individual had a residual age at motor onset of -25.7 years (counting the CAC interruption as a CAG), whereas after correction the residual was -10.0 years (41 pure CAGs). Comparing the structure calls from WES and MiSeq (Table 5.3) demonstrates WES effectively captured most of the (CAG**CAA**)$_2$CAG alleles (~90% of those detected by MiSeq), although with a single false positive (~7%). A single individual with a CAG(**CAA**CAG)$_2$CAG was classified as having an early disease onset after correcting for pure CAG size (uncorrected MiSeq CAG length: -3.66 years; corrected MiSeq CAG length: -9.46 years).

In contrast, WES had a low power to detect non-interrupted CAG repeats, only identifying 33% of those found through MiSeq. This probably reflects the difficulty in manually delineating where a pure CAG repeat ends without a CAA interruption. Unlike the (CAG**CAA**)$_2$ haplotype, there are at least four haplotypes we observed without interruptions. No wild-type allele we sequenced lacked an interruption. Phasing of expanded alleles shows perfect segregation of uninterrupted alleles with early onset and multiple interrupted alleles with late onset. All interruptions observed were at the 3’ end of the repeat tract; we did not find any alleles with interruptions in the 5’ or middle of the repeat.

We also observe variation in the polyproline-encoding tract. *HTT*’s polyproline tract, formed of the CCG and CCT repetitive stretch just 3’ of the *HTT* CAG repeat, was found to vary between 8 to 15 amino acids in length. In terms of coding sequence, the tract canonically contains a single CCA interruption near the 5’ start of the CCG repeat. Expanded alleles lacking this CCA interruption were associated with an early onset, although only two of these were in normally interrupted CAG alleles (both in early onset individuals). We did not see any further interruption species in the polyCCG and polyCCT tracts.

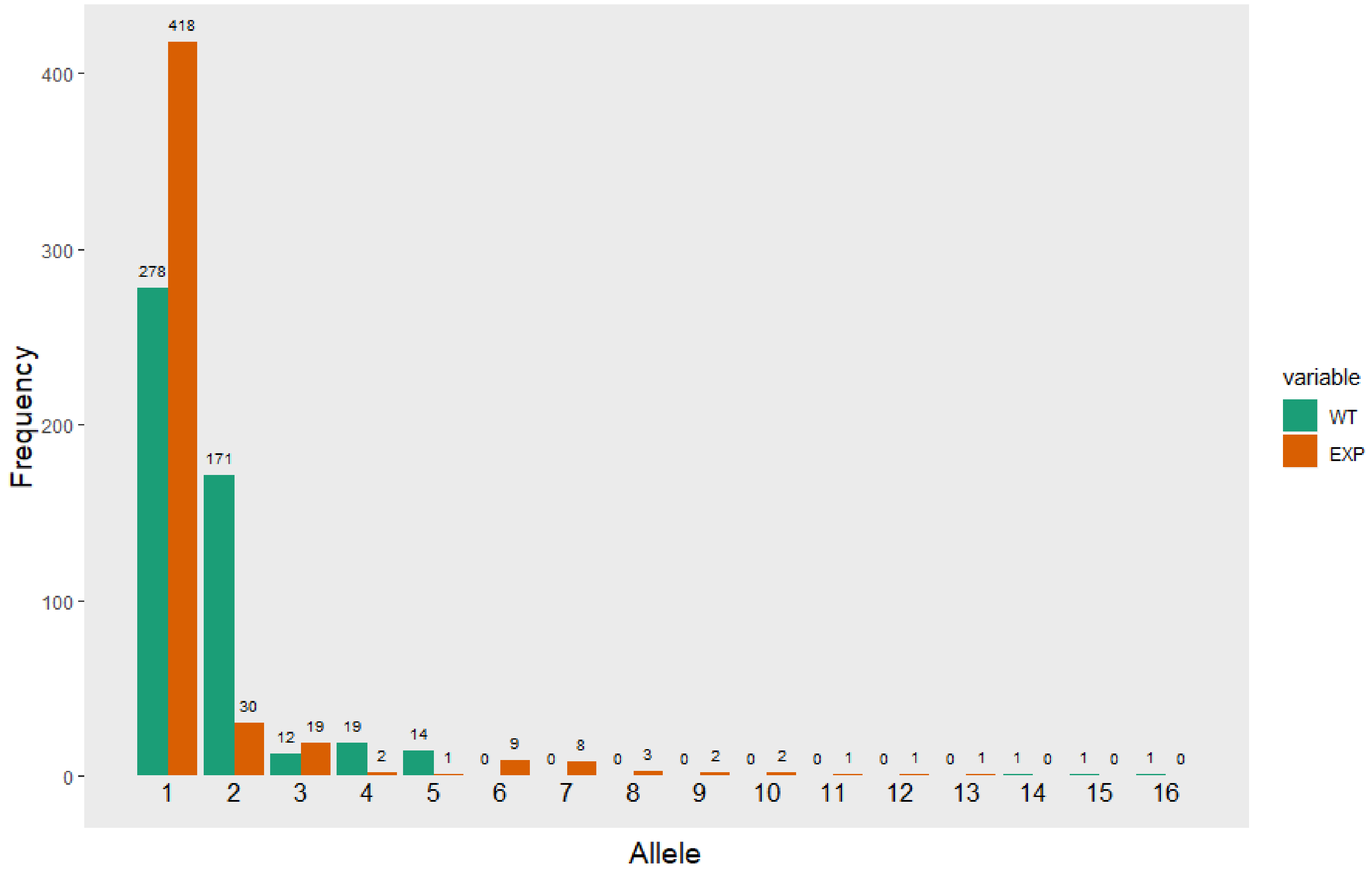


Figure 5.1: Visualisation of observed *HTT* alleles from MiSeq. The frequency of *HTT* allele structures are shown for wild-type (WT) and expanded (EXP) alleles. Note allele numbers shown refer to the numbers (No.) in Table 5.2, and these are listed in the same order. N=497.

		All individuals (N=497)				Dichotomous population (N=438)			
Allele structure	No.	E - WT	E - EXP	L - WT	L - EXP	E - WT	E - EXP	L - WT	L - EXP
5’-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_7$(CCT)$_2$-3’	1	132	209	146	209	122	188	127	180
5’-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_{10}$(CCT)$_2$-3’	2	86	13	85	17	77	12	72	13
5’-TTC(CAG)$_n$(**CAA**CAG)$_2$CCG**CCA**(CCG)$_7$(CCT)$_3$-3’	3	8	1	4	18	5	0	4	16
5’-TTC(CAG)$_n$**CAA**CAG(CCG)$_9$(CCT)$_2$-3’	4	14	2	5	0	12	2	3	0
5’-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_9$(CCT)$_2$-3’	5	6	1	8	0	6	1	7	0
5’-TTC(CAG)$_n$(CCG)$_{12}$(CCT)$_2$-3’	6	0	9	0	0	0	9	0	0
5’-TTC(CAG)$_n$CCG**CCA**(CCG)$_7$(CCT)$_2$-3’	7	0	8	0	0	0	8	0	0
5’-TTC(CAG)$_n$(CCG)$_{10}$(CCT)$_2$-3’	8	0	3	0	0	0	2	0	0
5’-TTC(CAG)$_n$(CCG)$_9$(CCT)$_2$-3’	9	0	2	0	0	0	2	0	0
5’-TTC(CAG)$_n$(**CAA**)$_2$CAGCCG**CCA**(CCG)$_7$(CCT)$_2$-3’	10	0	0	0	2	0	0	0	2
5’-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_7$(CCT)$_3$-3’	11	0	0	0	1	0	0	0	1
5’-TTC(CAG)$_n$(**CAA**)$_3$CAGCCG**CCA**(CCG)$_7$(CCT)$_2$-3’	12	0	0	0	1	0	0	0	1
5’-TTC(CAG)$_n$**CAC**(CAG)$_3$**CAA**CAGCCG**CCA**(CCG)$_7$(CCT)$_2$-3’	13	0	0	0	1	0	0	0	1
5’-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_{11}$(CCT)$_2$-3’	14	1	0	0	0	1	0	0	0
5’-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_4$(CCT)$_2$-3’	15	1	0	0	0	1	0	0	0
5’-TTC(CAG)$_n$**CAA**CAGCCG**CCA**(CCG)$_9$(CCT)$_3$-3’	16	0	0	1	0	0	0	1	0

Table 5.2: Identified *HTT* structures in an extreme onset HD cohort. Here, the ‘all individuals’ column refers to all samples which were successfully sequenced (N=497; two failed sequencing and one had no expanded allele). Equally, the dichotomous population describes the subset of samples with ≥5 |uncorrected AMO residual| sequenced (N=438; dichotomous group defined in 4.3.6). The numbers in the ‘No.’ column refer to Fig. 5.1. Interruptions in the CAG and CCG repeat tracts, CAA and CCA, respectively, are emboldened. A CAC interruption observed in one allele is similarly emboldened. Blue text refers to sequence encoding the polyglutamine repeat; green text is sequence encoding the polyproline repeat. WT: Wild-type; EXP: Expanded; E: Early; L: Late.

	Early					Late				
	Total (WES)	Total (MiSeq)	False positive	False negative	Previously unknown	Total (WES)	Total (MiSeq)	False positive	False negative	Previously unknown
(CAG)$_n$CCG	7	21	0	8	6	0	0	0	0	0
(CAG)$_n$(**CAA**CAG)$_2$CCG	6	5	1	0	0	20	20	1	0	1
(CAG)$_n$(**CAA**)$_2$CAGCCG	0	0	0	0	0	2	2	0	0	0
(CAG)$_n$(**CAA**)$_3$CAGCCG	0	0	0	0	0	1	1	0	0	0
(CAG)$_n$**CAC**(CAG)$_3$**CAA**CAGCCG	0	0	0	0	0	0	1	0	1	0

Table 5.3: Comparing WES and MiSeq *HTT* structural calls. The total counts for expanded and wild-type alleles are shown for both exome data (WES) and MiSeq data in early and late onset individuals. False positive, false negative and previously unknown counts refer to differences between WES and MiSeq calls. Previously unknown calls in WES are those with very poor coverage where an estimation of allele structure could not be made (see also Table 4.2).

5.3.2 The effect of interruption structures on age at onset

Having obtained *HTT* allele structures, we next wanted to investigate how strongly alternative allele structures were associated with modified HD onset. Generalised linear models (GLMs) were constructed regressing age at motor onset (AMO) residual on *HTT* allele structure in the continuous phenotype population previously defined (N=483; 4.3.6). Allele structure was coded by the number of interruptions present: 0, 1, 2 or 3 (see 2.9.4). We added several additional features of *HTT* structure as covariates in both expanded and wild-type alleles: number of CAG repeat interruptions, CCG interruption (CCA) and the length of the polyproline repeat. Results are shown in Table 5.4.

First, we used the uncorrected residuals calculated in 2.9.3/4.3.6, where expected age at onset was estimated using polyglutamine length-2 to be equivalent to genescan. These data showed interruptions in expanded *HTT* alleles were strongly and significantly associated with age at motor onset residual (*p*=2.83E-10). Our model predicts a ~15 years later onset for each interruption present (and thus a ~15 year earlier onset where CAA is missing), although this is probably an overestimate as we are examining an extreme onset cohort (6.2.2). Wild-type CAG repeat interruptions were not associated with onset modification (*p*=4.61E-01). Additionally, interruptions in the CCG repeat and the total polyproline length were equally not associated with onset residual in either expanded or wild-type *HTT*. However, it is notable that due to the relative infrequency of expanded alleles lacking the CCA interruption in our data (N=2 independent structures), this does not entirely preclude a role for the polyproline tract interruptions in onset modification.

Secondly, we used the corrected age at onset residual in a generalised linear model, where expected age at onset was estimated using the length of the pure CAG repeat. Here, the significance of interruptions was reduced, but still significant (*p*=4.13E-04). Therefore, a large portion of the significance is driven by mis-sizing of the pure CAG length when calculating a predicted age at onset residual for individuals with non-canonical alleles. However, as some of the signal remains, this indicates interruptions may have further effects on age at onset modification. An alternative generalised linear model (Appendix 18) coded *HTT* alleles as canonical (0, 0), containing an additional interruption of any type (1, 0) or as loss of interruption (0, 1) (see 2.9.4). This modelling approach indicates the remaining significance following correction of motor residual using pure CAG length appears to be accounted by alleles with additional interruptions (*p*=1.45E-03) and not by loss-of-interruption (LOI) alleles (*p*=1.81E-01).

	Uncorrected (N=483)				Corrected (N=483)			
	B	β	SE	*p*	*B*	β	SE	*p*
CAG tract interruption (EXP)	14.728	0.321	2.285	**2.83E-10**	8.133	0.185	2.287	**4.13E-04**
CAG tract interruption (WT)	-2.836	-0.031	3.841	4.61E-01	-2.849	-0.033	3.843	4.59E-01
CCG tract interruption (EXP)	5.249	0.065	5.034	2.98E-01	4.428	0.057	5.038	3.80E-01
CCG tract interruption (WT)	6.231	0.084	3.182	5.08E-02	6.195	0.087	3.185	5.23E-02
PolyP length (EXP)	0.120	0.009	0.777	8.78E-01	0.161	0.012	0.778	8.36E-01
PolyP length (WT)	-0.446	-0.045	0.425	2.95E-01	-0.439	-0.046	0.425	3.03E-01

Table 5.4: GLMs for *HTT* allele structure on residual HD age at motor onset. Shown are generalised linear models (GLMs) for different structural features of *HTT* in both expanded and wild-type alleles, regressing uncorrected and corrected AMO residuals on various structural features of *HTT*. Interruptions are coded between 0 to +3 depending on the number of interruptions they contain (see 2.9.4). Individuals previously quality controlled from 4.3.6 are used; N=483 as two individuals failed MiSeq sequencing. See Appendix 18 for an alternative GLM. EXP: Expanded allele; WT: wild-type allele; PolyP: polyproline. *B* = unstandardised coefficient; β = standardised coefficient; *SE* = standard error.

5.4 Validating MiSeq using genescan data

5.4.1 CAG length

Scale-HD finds the two modal peaks in its sequencing data to determine CAG size (see Fig. 5.5 later). As previously discussed in chapter 4, MiSeq was used to calculate two lengths: polyglutamine-2 (uncorrected length), which assumes a single CAA interruption at the 3' end of the repeat to be in-line with non-sequencing genotyping techniques, and pure CAG length (corrected length). Uncorrected MiSeq polyglutamine-2 lengths ranged from 39 to 53 whereas MiSeq pure CAG lengths varied between 38 to 52. To help validate and explore MiSeq CAG length data, we wanted to compare our data to that of genescan, as this is routinely used for CAG length determination (Massey et al., 2018). We generated genescan data for all individuals sequenced by MiSeq and compared CAG lengths from these to those derived by local labs in our Registry database. Graphs comparing CAG lengths from MiSeq, local labs (from Registry) and genescan (Fig. 5.2) and the resultant residual ages at motor onset (Fig. 5.3 & 5.4) illustrate several important points.

As expected, local lab derived CAG lengths show the smallest correlation with AMO (R^2=0.043), as we used these CAG lengths to derive AMO residual and select our early/late individuals in 3.7. Local CAG lengths also produced the largest age at motor onset residual standard deviation (SD) at 16.79 years. Genescan CAGs show a much higher correlation (R^2=0.239) and smaller SD (14.37 years), demonstrating local labs tended to have different (and likely less accurate) CAG measurements compared to our genescan-derived CAGs, probably due to be local technical variation. Local CAGs were the same as genescan 18.5% of the time, within ±1 CAG 61.8% of the time and within ±2 CAGs 82.5% of the time. Uncorrected MiSeq CAG lengths had comparable, but slightly larger correlations (R^2=0.284) and smaller SDs, than genescan (14.05 years), suggesting MiSeq may be more accurate in determining *HTT* allele size than genescan even when not factoring *HTT* sequence. Additionally, genescan seemed to consistently be ~1-2 CAGs shorter compared to either local lab or MiSeq lengths. Finally, as anticipated, corrected pure CAG lengths from MiSeq have both the highest R^2 (0.364) and lowest SD (13.46 years). Thus, these data show mis-sizing of the expanded pure CAG stretch due to the presence of atypical allele structures roughly accounts for ~8% of the residual age at motor onset variation in our data.

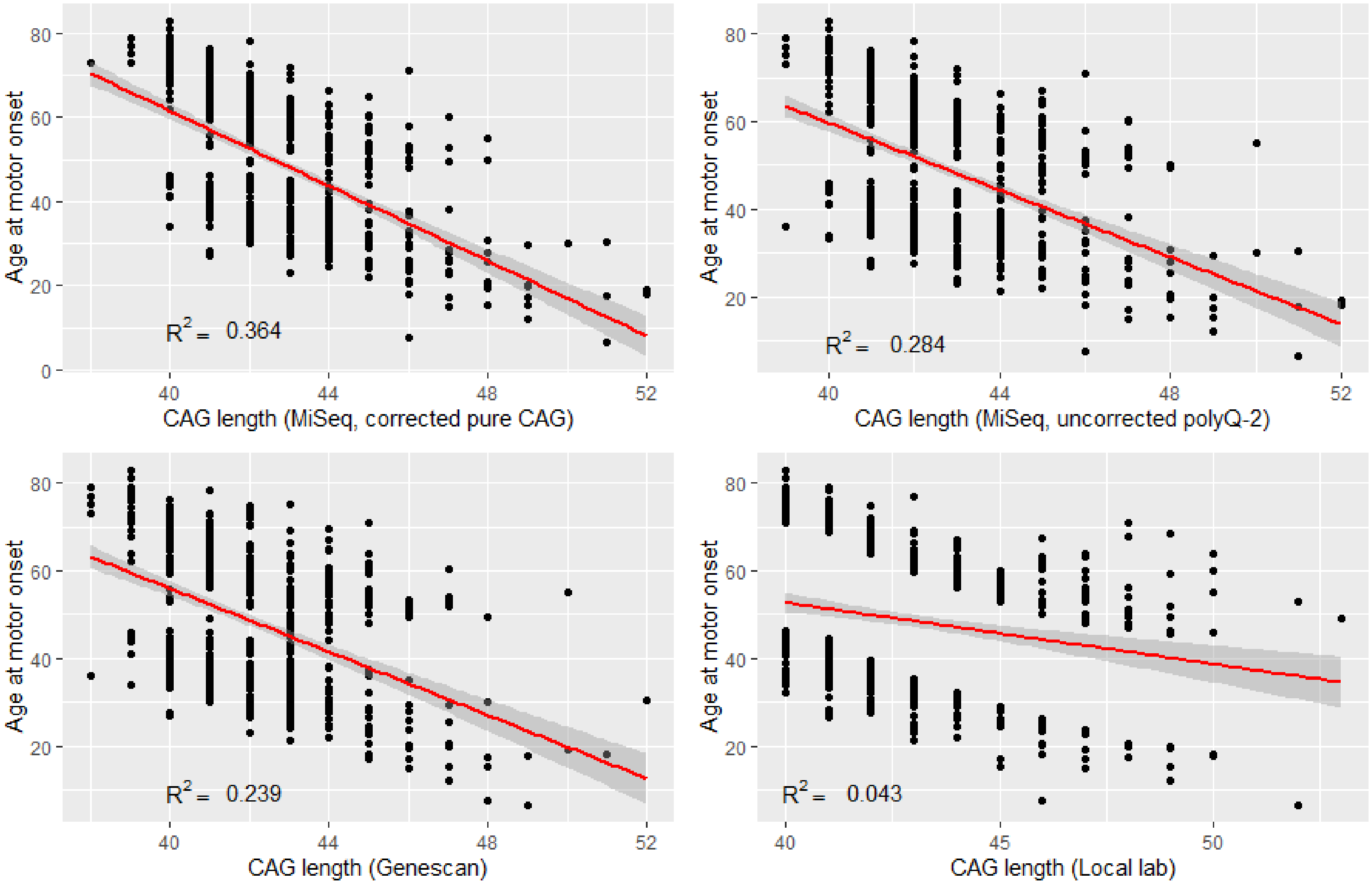


Figure 5.2: Comparing CAG lengths from MiSeq, genescan and local labs. CAG sizes from different sources are shown: MiSeq corrected (pure CAG size) and uncorrected (polyglutamine-2) CAG lengths (N=497 for both); genescan (N= 499); and local labs (N=500). The individual whose expanded allele did not amplify is not included in the genescan/MiSeq graphs.

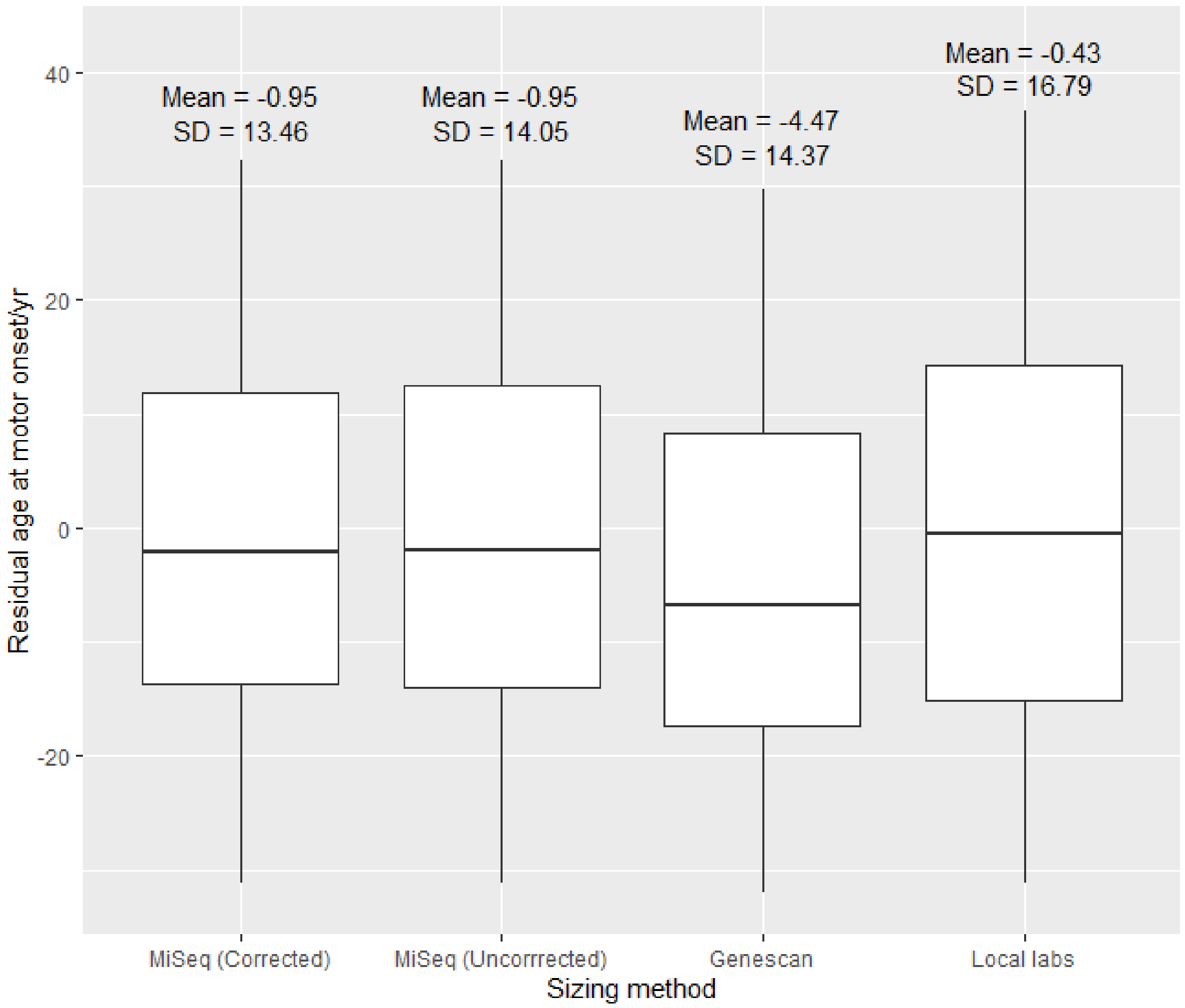


Figure 5.3: Residual ages at motor onset from MiSeq, genescan and local labs. Standard deviations (SDs) for residual age at motor onset were calculated using CAG lengths from MiSeq (corrected (N=491) and uncorrected (N=492)), genescan (N=454) and local labs (N=500) where CAG length was ≥40 CAG. Both mean and SD are given in years.

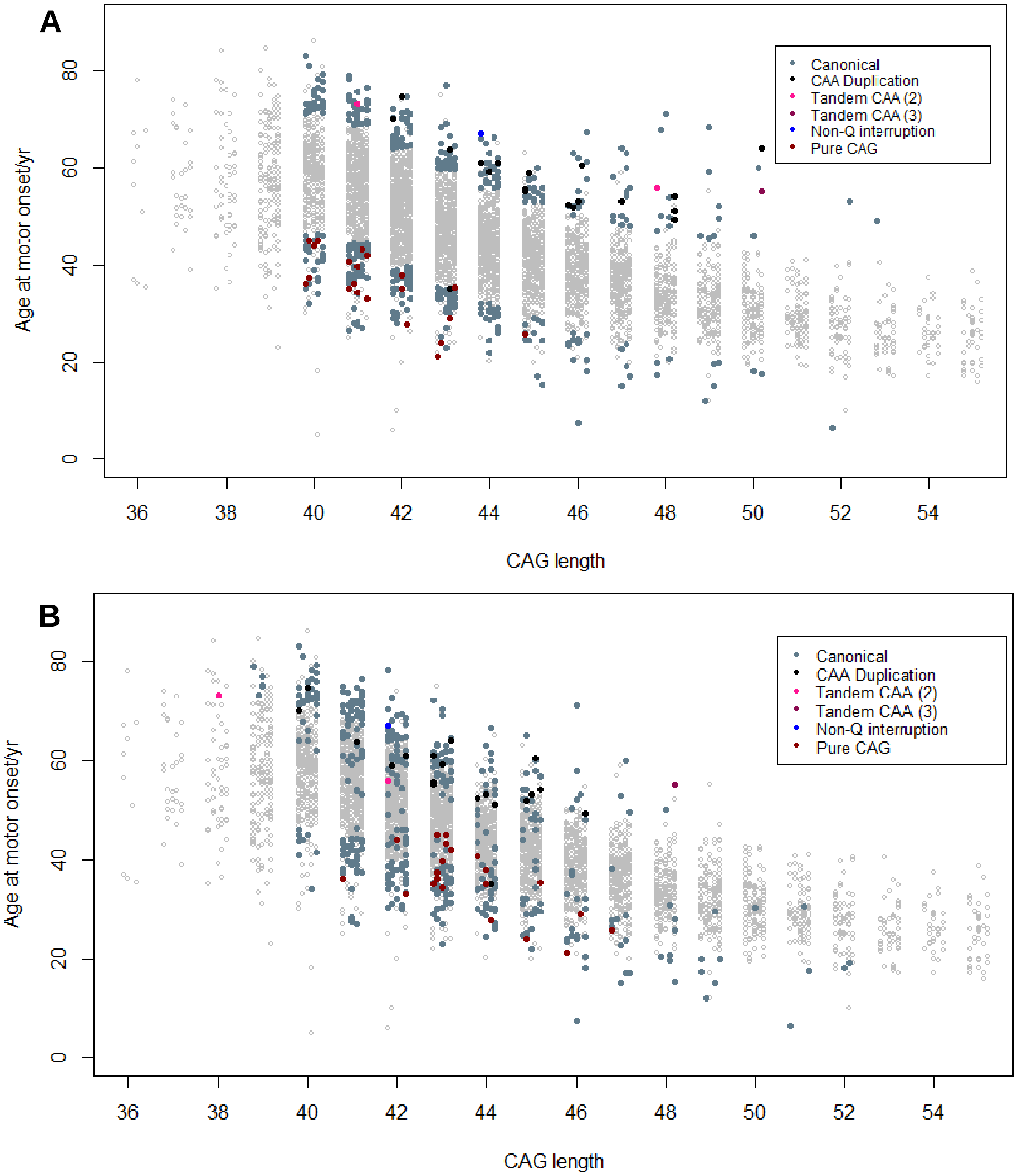


Figure 5.4: CAG lengths before and after using MiSeq against age at motor onset. CAG lengths from either local labs (A) or MiSeq (B) are plotted against age at motor onset, with individuals possessing atypical expanded *HTT* alleles highlighted in both. MiSeq CAG lengths plotted in B refer to pure CAG length. Total N of non-sequenced (grey circles) = 5851; N sequenced (various colours) = 500 (A), N=498 (B). The legend refers to interruptions occurring in the CAG repeat. Age at onset are best estimates calculated in 3.2.3. In order: canonical (5'-CAG**CAA**CAG-3'); CAA duplication (5'-(CAG**CAA**)$_2$CAG-3'); tandem CAA (2) (5'-CAG(**CAA**)$_2$CAG-3'); tandem CAA (3) (5'-CAG(**CAA**)$_3$CAG-3'); non-Q interruption (5'-CAG**CAC**(CAG)$_3$**CAA**CAG-3'); pure CAG (5'-(CAG)$_n$-3' (no interruption)).

5.4.2 Repeat instability

In addition to CAG length and *HTT* structure, MiSeq read distribution data can be used to produce metrics for both expansion and contraction of CAG repeats. These data are produced as part of the Scale-HD workflow examining peak heights (see methods 2.9.2 and Equations 2.3 & 2.4). Scale-HD calculates forward CAG repeat instability (somatic mosaicism) using the first 10 CAG peaks larger than the modal repeat, and backwards repeat instability (slippage) using the 2 CAG peaks smaller than the modal CAG. Somatic mosaicism was significantly higher in expanded *HTT* alleles compared to those in wild type alleles (0.026 vs 0.39, *p*=<2E-16 Welch two sample t-test). Similarly, slippage was also larger in expanded alleles (0.102 vs 0.405, *p*=<2E-16 Welch two sample t-test). Note, however, slippage is predominantly a technical measure of polymerase slippage, and is not correlated with MiSeq somatic mosaicism (*p*=0.71). Interestingly, we also observe a certain degree of instability of the CCG repeat, although this is much less marked than the CAG repeat. An example read frequency plot for a late onset individual for CAG and CCG sizes in both expanded and wild-type *HTT* alleles is available in Fig. 5.5.

Genescan can also produce measures of CAG instability in a similar way to MiSeq. We generated both expansion indices (which only considers forward instability) and instability indices (which consider both backwards and forwards instability) from our genescan data in line with (Lee et al., 2010). Expansion and instability indices from our genescan data were plotted against MiSeq instability measures (Fig. 5.6). For MiSeq mosaicism, both genescan expansion and instability indices were highly correlated (R^2=0.77 for expansion index; R^2=0.63 for instability index), although expansion index more so, likely as genescan expansion index is calculated in a similar way to MiSeq somatic mosaicism. Comparatively, genescan instability index considers peaks smaller than the modal CAG. MiSeq slippage was not correlated with either measure, unsurprising given it is not correlated with MiSeq mosaicism. Overall, genescan expansion index, genescan instability index and MiSeq somatic mosaicism are similar and can be used for assaying instability of the CAG repeat. Importantly, the MiSeq-derived somatic mosaicism is robust between plates with minimal batch effects observed (see Appendix 17).

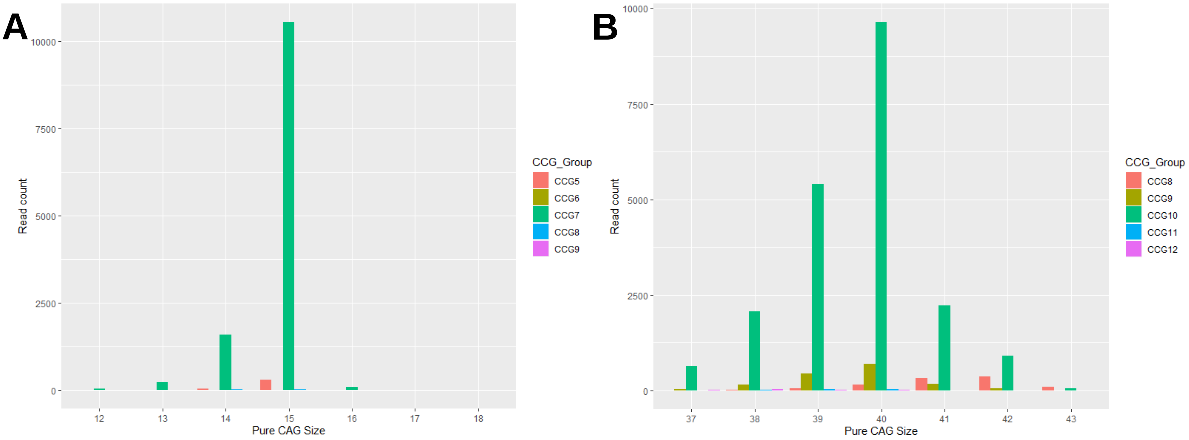


Figure 5.5: CAG repeat read frequency plots for wild type and expanded alleles. Example read distribution plots are shown for DNA taken from an individual with late HD onset (lymphoblastoid DNA). Shown are reads that map ±3 CAGs from the modal CAG repeat as well as ±2 from the modal CCG repeat at each CAG length. (A) shows the wild-type allele, (B) shows the expanded allele.

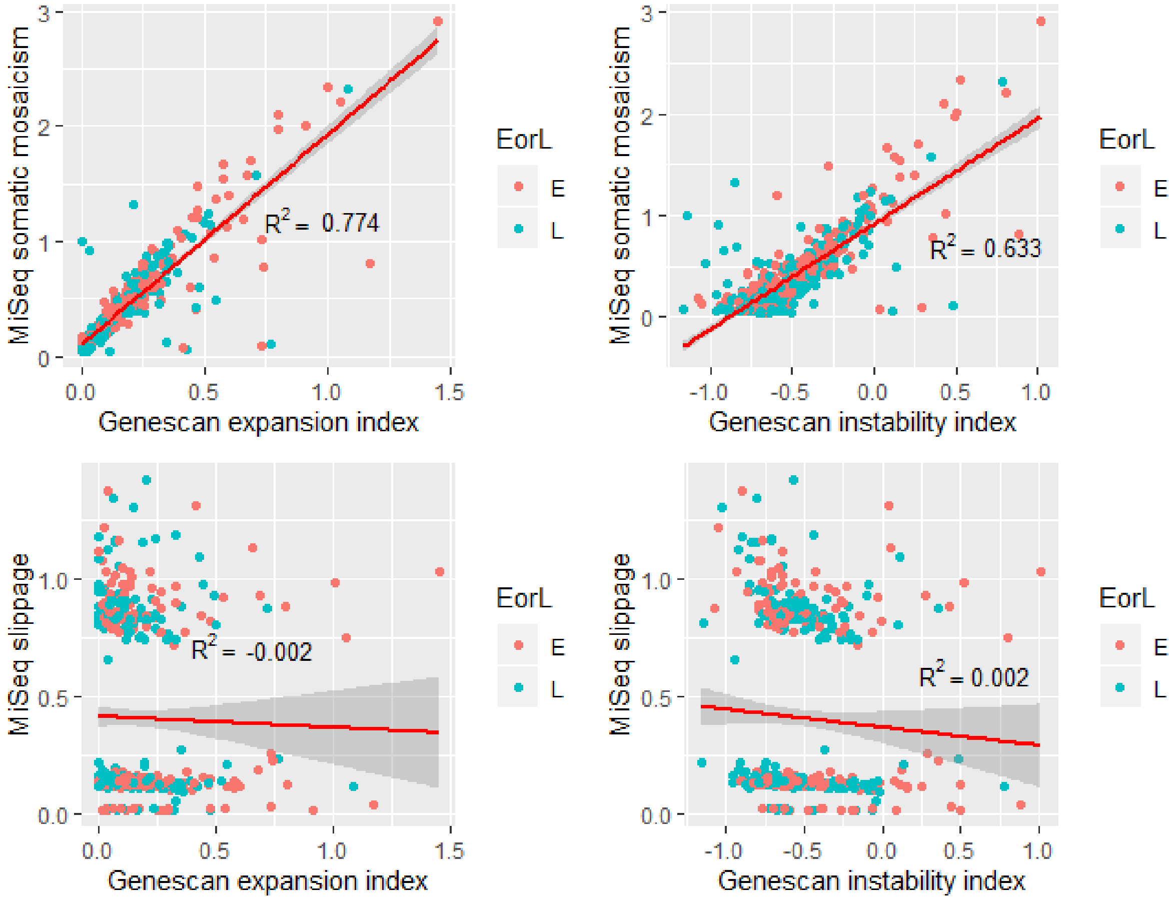


Figure 5.6: MiSeq and genescan somatic instability measures. MiSeq somatic mosaicism (top) and slippage (bottom) are shown for genescan expansion index (left) and instability index. The adjusted R^2 is shown. N=494 for all cases, and DNA are all derived from lymphoblastoid cells.

5.5 Comparing lymphoblastoid and blood DNA

So far, the analysis has focused on the lymphoblastoid-derived DNA data as these DNA were available for all subjects (N=500). However, we had both whole blood and lymphoblastoid DNA for a subset of individuals (N=42), and we sequenced both with MiSeq to investigate how CAG length and instability differed. In most cases, CAG length was identical between blood and lymphoblastoid DNA (83%; Table 5.5). In cases where differences arose, most were ±1 CAG (>95%) and no more than ±2 CAG. Somatic mosaicism measures for these samples were more varied (Fig. 5.7), and lymphoblastoid and blood-derived mosaicism measures were not significantly associated with each other in our somewhat small dataset (N=42), although the *p* value is approaching significance (*p*=0.10, *B*=0.11, R^2=0.043).

The lymphoblastoid DNA used for exome sequencing and MiSeq were early passage samples from BioRep (see 2.5.1). However, in addition to these, we also had a small number of lymphoblastoid cells which had been grown over successive months. Longitudinal samples from these lymphoblastoid cells were sequenced to examine how CAG length and repeat instability change over time in culture. As indicated in Fig 5.8, CAG length was consistent over time in most grown lymphoblastoid lines. However, somatic mosaicism in all but one line (L96) decreased or remained mostly static after being grown in culture. We were curious whether the mosaicism changes were driven by variation acquired through time in culture. A SNP genotyping array for detecting copy number variants (CNVs) was carried out to investigate (Appendix 19). Most CNVs were identical between lymphoblastoid and blood DNA. The two exceptions were CNVs in late onset lymphoblastoid samples (L21 and L31). Both samples contained the chr22:22,300,000-22,904,555 CNV (GRCh38) not found in their paired blood DNA. Given the CNV was the same in both samples, it is possible this was either picked up through the immortalisation process with Epstein Barr virus (EBV) or through time in culture.

CAG difference	Frequency
-2 CAG	2
-1 CAG	1
Same CAG	35
+1 CAG	4
+2 CAG	0

Table 5.5: Differences between blood and lymphoblastoid CAG lengths (MiSeq). Frequency of CAG size differences between lymphoblastoid (LBC) and blood are shown (LBC-blood). N=42.

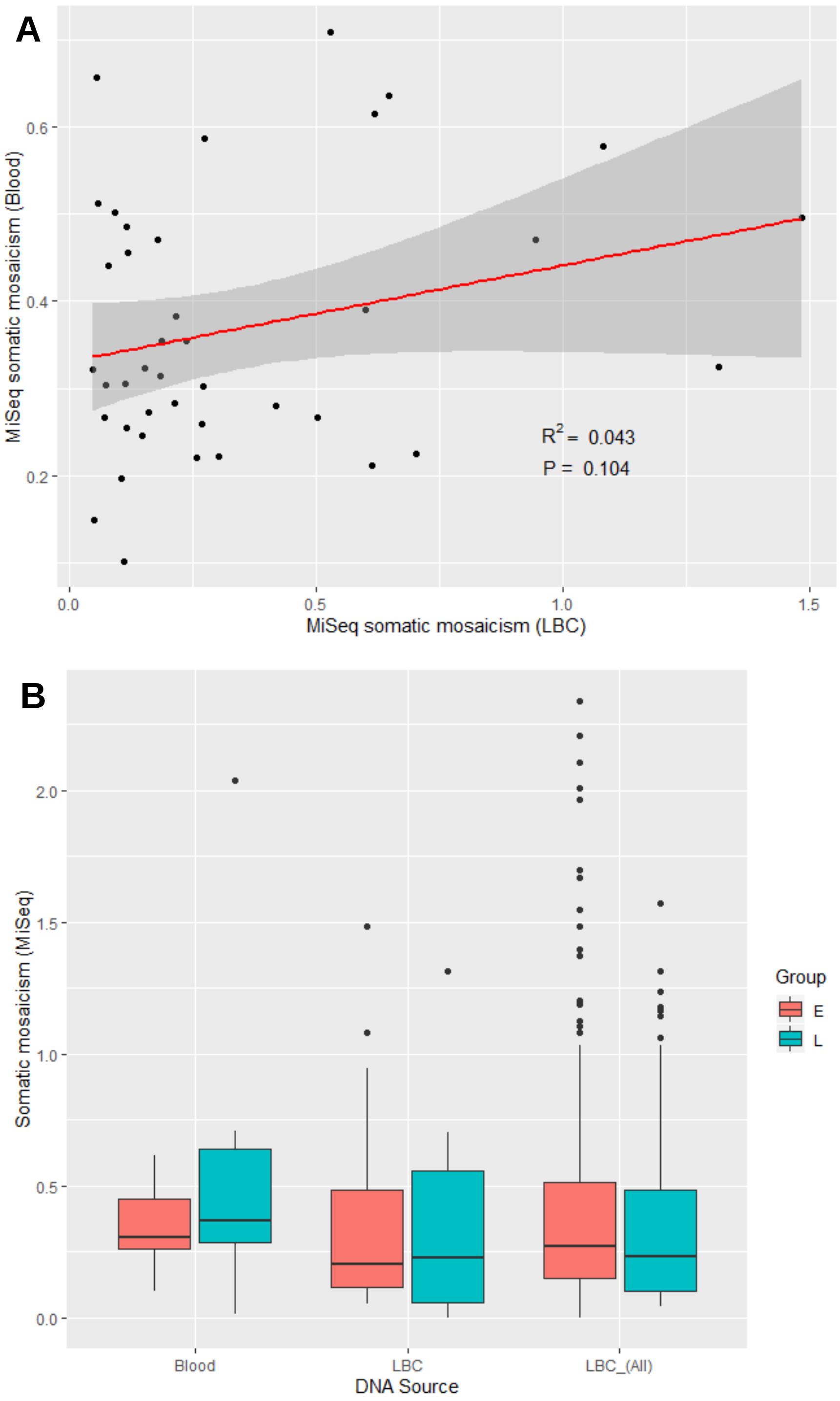


Figure 5.7: CAG instability in blood and lymphoblastoid cells. (A) Plot of lymphoblastoid (LBC) and blood DNA somatic instabilities (N=40). (B) Boxplots showing the variability of mosaicism in lymphoblastoid and blood DNA in early (E) and late (L) individuals in the same samples (N=38; 26 early, 12 late). LBC (All) covers all lymphoblastoid cell DNA (N=436; 224 early, 212 late).

Sample ID	Months in culture	CAG (earliest)	CAG (latest)	Difference
E119	1	43	43	0
E13	3	42	42	0
E14	2	46	46	0
E144	1	46	46	0
E29	1	46	46	0
E34	1	43	43	0
E40	1	41	41	0
E6	1	46	47	+1
E61	1	43	43	0
E70	2	43	43	0
E71	1	41	41	0
L118	3	50	49	-1
L21	1	42	42	0
L22	1	41	41	0
L52	1	45	45	0
L96	2	39	39	0
N25	2	50	52	+2

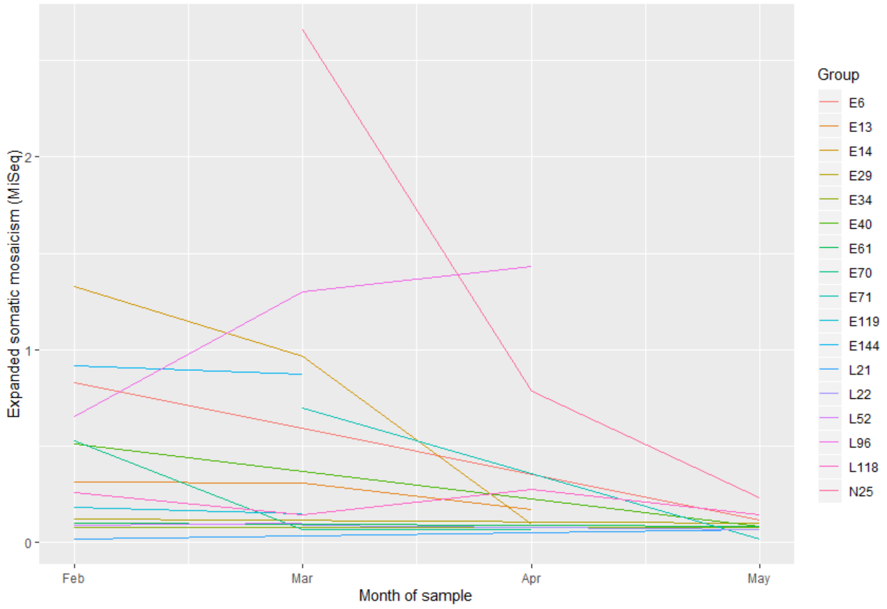


Figure 5.8: Longitudinal lymphoblastoid cell CAG length and mosaicism. Longitudinal lymphoblastoid DNA samples were sequenced. The top table indicates the earliest and latest passages for which DNA was available and the corresponding pure CAG lengths. The bottom plot shows how somatic mosaicism of the expanded CAG changes through time in culture. Month when the DNA was taken is shown on the x-axis, plotted against their MiSeq somatic mosaicism. Lines N=17, total sample N=41. The codes refer to cells from early (E), late (L) and normal/expected (N) HD onset individuals.

5.6 Repeat instability in early onset and late onset individuals

Somatic instability of the *HTT* repeat has been implicated as a driver of disease pathogenesis (Kennedy et al., 2003; Shelbourne et al., 2007; Swami et al., 2009), likely influenced by DNA repair machinery (Tomé et al., 2013; Budworth et al., 2015; GeM-HD Consortium, 2015; Hensman Moss et al., 2017; Lee et al., 2017; GeM-HD Consortium, 2019; Flower et al., 2019) (see also 1.7 & Table 1.3). Thus, we next wanted to investigate whether there were discernible differences in CAG repeat instability in early or late onset patients, and these data are plotted in Fig. 5.9. Lymphoblastoid DNA has a weak but marginally significant (*p*=1.74E-02) negative relationship between onset and mosaicism. Blood mosaicism in Fig. 5.9 appears to have the opposite relationship than expected (where later onsets are associated with larger degrees of somatic mosaicism; *p*=8.44E-03), although the sample size is very small. To understand these data further, we constructed generalised linear models accounting for (1) age of the individual and (2) CAG length, as both these can modify instability (Lee et al., 2011; Ciosi et al., 2019). We also covaried for (3) CAG interruption structure. The results are shown in Table 5.6 for both lymphoblastoid and blood-derived DNA using the corrected age at onset residual (pure CAG length).

CAG size is positively and significantly correlated with somatic mosaicism where longer CAG tracts are associated with higher degrees of CAG instability, as expected. The age of sampling, *i.e.* the age of the individual when their blood was taken (either directly for blood DNA or making lymphoblastoid cell lines), approaches significance in lymphoblastoid cells and is significant in blood DNA. The presence of a CAA interruption is significantly associated with mosaicism in blood DNA but not lymphoblastoid-derived DNA. Corrected residual age at motor onset was not significant in either model, although it is approaching so in the lymphoblastoid DNA (*p*=1.16E-01, *B*=-0.004).

The modelling approach we have utilised, however, is hampered by our study's selection criteria. By purposely enriching our sequenced population for early and late onset individuals, both CAG length (43.62 in early and 42.26 in late individuals, N=483) and sampling age (42.10 years in early and 67.75 in late individuals, N=465) are significantly different (*p*=1.13E-11 and *p*=<2.2E-16, respectively, Welch two sample t-test). This may explain the counter-intuitive findings we observed in blood mosaicism. An approach to mitigate this in an updated model is discussed later (5.8.5).

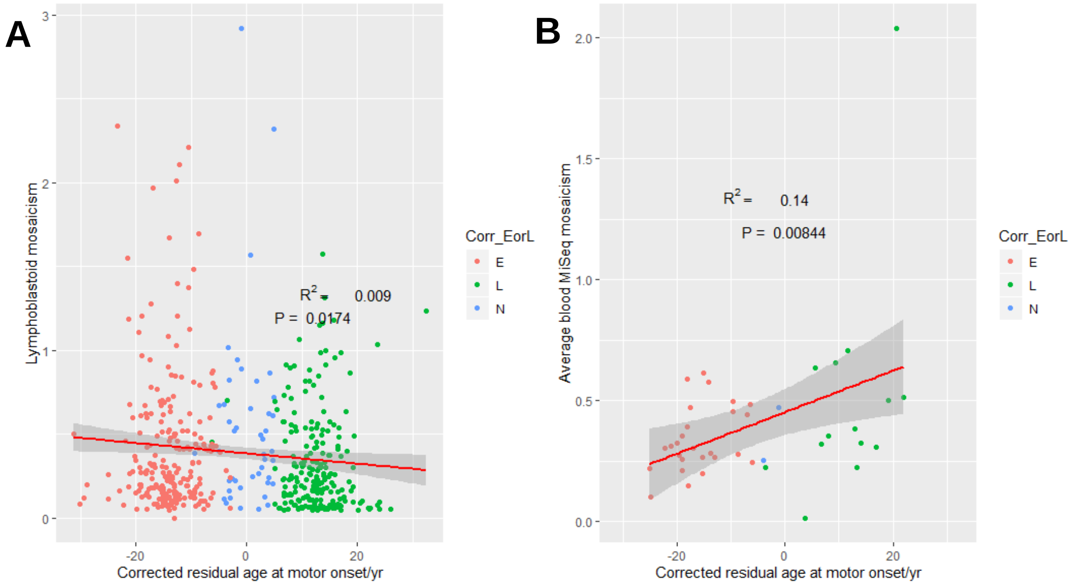


Figure 5.9: MiSeq mosaicism and age at onset (LBC and blood). Corrected residual ages at motor onset are plotted against MiSeq mosaicism from lymphoblastoid (A; N=496) and blood (B; N=42) DNA. E: early onset; L: late onset; N: normal/expected onset. Colours refer to the type of onset as defined by the corrected residual (normal onsets = <5 |age at motor onset residual|).

	Lymphoblastoid (N=461)				Blood (N=40)			
	B	β	SE	*p*	*B*	β	SE	*p*
Corrected residual	-0.004	-0.146	0.003	1.16E-01	-0.004	-0.194	0.004	3.38E-01
Interruption	-0.043	-0.032	0.056	4.51E-01	0.139	0.267	0.057	*1.96E-02*
Sample age	0.005	0.203	0.003	6.29E-02	0.011	0.543	0.004	*1.49E-02*
CAG length	0.094	0.528	0.011	**4.78E-17**	0.113	0.825	0.017	**1.71E-07**

Table 5.6: GLMs interrogating modifiers of somatic mosaicism. Generalised linear models (GLMs) regressing MiSeq mosaicism on various covariates are shown (including corrected age at motor onset). Here, sample age refers to when blood was collected (either to form lymphoblastoid cells or blood DNA directly), and interruption refers to the number of interruptions seen. For example, (CAG**CAA**)$_2$ would be coded as 2 (methods 2.9.4). LBC: lymphoblastoid cell; *B*: unstandardised coefficient; β: standardised coefficient; *SE*: standard error.

5.7 Disease progression and repeat instability

Having explored somatic mosaicism in early and late onset individuals, we finally wanted to consider if mosaicism in our HD patients was associated with an independent disease phenotype. For this, we chose a recently published HD progression measure (Hensman Moss et al., 2017). This progression measure was available for 146 individuals for whom we had lymphoblastoid cell DNA and 15 for whom we had blood DNA.

In the progression measure, larger values indicate faster disease progression. Unexpectedly, early onset HD patients tended to have slower progression (mean=-0.24) compared to late onset individuals (mean=0.23). This may be an age-related effect as regressing progression on corrected residual in a generalised linear model is significant (*p*=6.71E-03) covarying for interruption and expanded CAG length. Both age at motor onset and corrected age at motor onset residual were significantly positively associated with progression (Fig. 5.10; *p*=1.31E-02 and *p*=6.99E-4). Plotting disease progression against mosaicism calculated by lymphoblastoid DNA shows no significance (*p*=7.12E-01; Fig. 5.11), however; plotting mosaicism using whole blood approaches significance despite our low N (N=15, *p*=1.32E-01; Fig. 5.11). To investigate further, generalised linear models were built (Table 5.7). CAA interruption status was significantly associated with progression in lymphoblastoid DNA (*p*=1.86E-02, *B*=0.59), however in the opposite direction than expected. Again, this may be the result of our extreme cohort. MiSeq mosaicism nor any other variable pass the multiple testing significance threshold (*p*=2.50E-02). However, these tests are hampered by the same problems as already mentioned in section 5.6 (discussed in 5.8.5). Furthermore, our low N (N=129 and N=12) in these models further limits their usefulness.

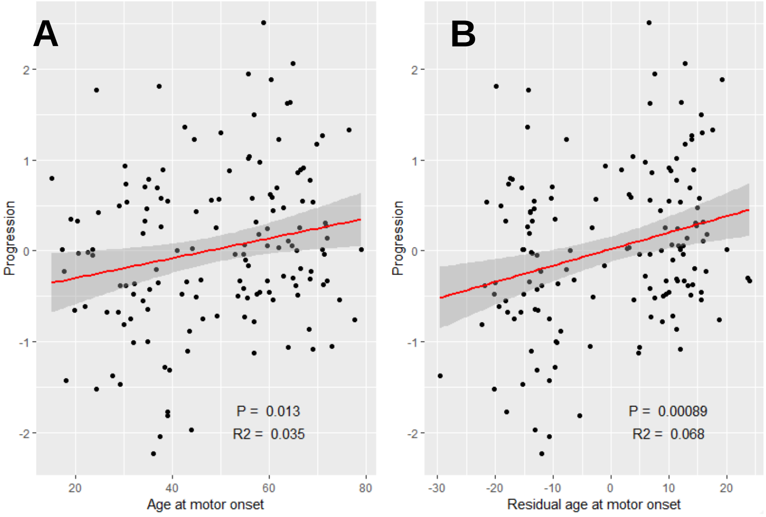


Figure 5.10: Comparing progression and age at motor onset. Indicated are plots for age at motor onset against progression (A) and residual age at motor onset against progression (B). Both N=147.

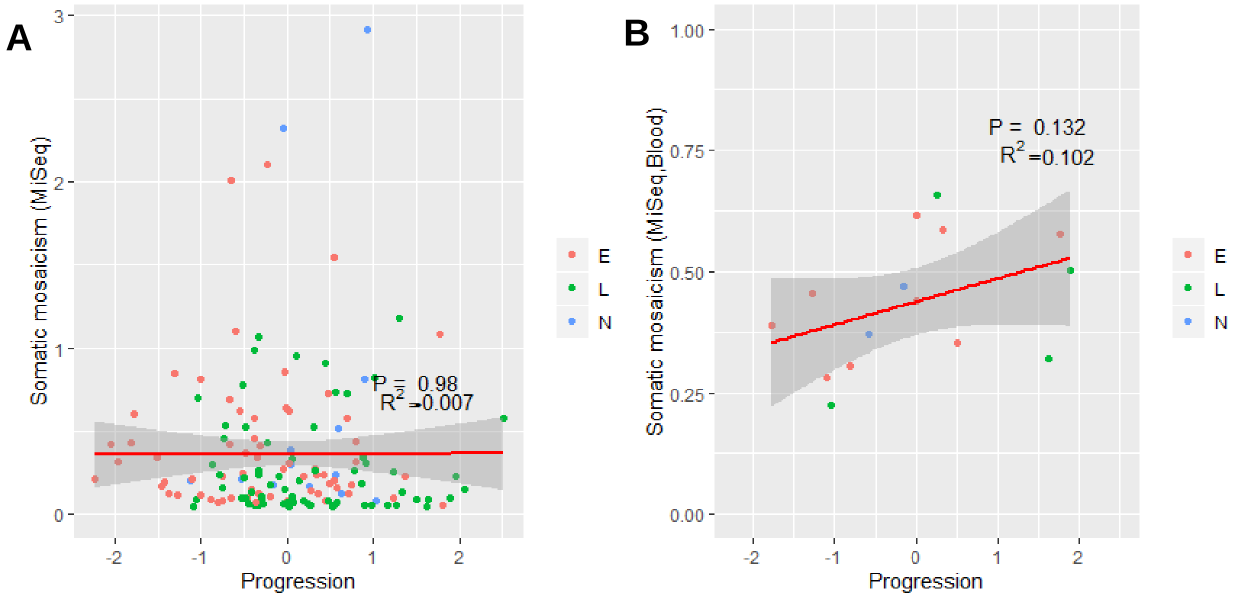


Figure 5.11: Relationship between MiSeq mosaicism and progression (LBC and blood). Plotted is progression against MiSeq CAG mosaicism for lymphoblastoid (A, N=146) and blood DNA (B, N=15). For both plots, the E (earlyonset), L (late onset) and N (normal/expected onset) colours refer to the type of onset as defined by the corrected residual (normal onsets = <5 |age at motor onset residual|).

	Lymphoblastoid (N=129)				Blood (N=12)			
	B	β	SE	p	*B*	β	SE	p
Somatic mosaicism	-0.065	-0.034	0.180	7.20E-01	0.907	0.131	2.995	7.71E-01
CAA interruption	0.588	0.212	0.247	**1.86E-02**	0.720	0.214	1.081	5.27E-01
Sample age	0.014	0.218	0.007	*4.89E-02*	0.036	0.536	0.039	3.87E-01
CAG length	0.083	0.226	0.042	*4.92E-02*	0.439	1.112	0.293	1.78E-01

Table 5.7: GLMs interrogating modifiers of HD disease progression. Indicated are generalised linear regression models (GLMs) for lymphoblastoid cell (LBC) and blood DNA. Significant values are emboldened, nominally significant values are italicised (using multiple testing correction cut-off Bonferonni *p*=2.50E-02).

5.8 Discussion

5.8.1 Overview of results

Chapter 5 has detailed the use of a targeted MiSeq method to sequence the *HTT* CAG (polyglutamine) and CCG/CCT (polyproline) repeat tracts. Using MiSeq, we identified 16 *HTT* repeat structures, and several of these are heretofore novel, including a non-glutamine interruption. Dephasing atypical alleles found those with additional interruptions were uniformly associated with late onset, and alleles lacking interruptions were wholly associated with early onset, supporting the findings in chapter 4 (4.5). We validated MiSeq by comparing it to the current gold standard of CAG length determination in HD, genescan. In our hands, MiSeq outperforms a locally run genescan with and without considering allele structure. Lymphoblastoid and blood were seen to generate equivalent CAG lengths in most cases, however CAG mosaicism was different between lymphoblastoid cells and blood DNA. Finally, we considered mosaicism in the context of two disease phenotypes: motor onset and disease progression. Although mostly inconclusive in our data, future study may investigate these phenotypes further, especially if using blood-derived mosaicism.

5.8.2 Robustness of the MiSeq protocol and analysis

The MiSeq protocol (Ciosi et al., 2018) in our hands had few failures (<1%). It was noted this was a lower rate of failure than previously observed for this protocol, possibly as we had a more rigorous DNA normalisation prior to sequencing using PicoGreen™. Few batch effects were observed. We found both CAG length and instability estimates from MiSeq were comparable to in-house genescan measurements, although MiSeq may have outperformed genescan slightly even when not considering allele structure. Non-canonical expanded *HTT* allele structure accounted for ~8% of the unexplained variation in residual age at motor onset seen in our sample.

MiSeq gives the longest continuous sequence of any second-generation sequencing platform currently available, capable of 600 sequencing cycles, and is well suited towards polyglutamine diseases where CAG repeats are often <80 CAGs in most patients (reviewed in (Lieberman et al., 2019); see also Table 1.1). Furthermore, interruption structures may still be detected even if complete read through is impossible, especially for HD where interruptions are only known to exist in the extreme 3' end of the CAG repeat. Given the importance of repeat interruptions in HD and repeat diseases in general (discussed in 3.8.4), a highly scalable method such as MiSeq has wider applications in the polyglutamine disease field. MiSeq is capable of high multiplexity (*e.g.* 384 samples) and is methodologically flexible. A modified MiSeq method was recently used to assay *MSH3* genotype in HD

(Flower et al., 2019), for instance. A similar MiSeq method could be extended to other STRs of interest such as *TCERG1* in chapter 4. One could also target multiple STR loci at once, at the cost of read depth per locus. For example, a recent study found intermediate allele sizes were higher than previously though in the general population (Gardiner et al., 2019). A high-plexity MiSeq method could sequence multiple disease-relevant loci at once, and in addition to repeat length, would provide insight into the role interruptions or other non-canonical structures play in a non-disease population.

Still, the primary limitation of second-generation short-read sequencing technologies such as MiSeq is that of read size. Illumina-based platforms are typically unable to handle large fragments (>~1,000bp) as these cluster poorly (Bronner et al., 2014), probably as longer inserts are less able to fold over during cyclical bridge amplification. Further, limitations in sequencing by synthesis chemistry leads to PHRED score drop off at the end of sequencing molecules as clusters become successively desynchronised (reviewed in (Fuller et al., 2009), also see Appendix 16). Hence while we ran the MiSeq with 400 forward cycles, the maximum resolvable sequence using this protocol is unclear. Another limitation of second-generation sequencing is the necessary PCR needed for cluster formation as these may introduce artificial PCR slippage/amplification in repeats. Thus, while for most HD patients (and probably polyglutamine diseases in general) MiSeq offers a high throughput method for assessment of both repeat length and structure, limitations in its read size restrict its usage in longer repeats found in other repeat diseases. Furthermore, MiSeq is unable to assay the repeats from HD disease models as these regularly have >120-150 CAGs. For these applications, a long-read sequencing approach will probably become the standard as protocols are established and cost comes down.

Bioinformatically, the Scale-HD pipeline (https://scalehd.readthedocs.io/) resolved most repeat structures accurately (~99%). However, several novel and very rare alleles were incorrectly called as they were missing from Scale-HD's reference sequence (refseq) library. Hence, we further had to employ manual curation of all alignments, and this was laborious and time consuming. In future analyses, we would suggest the refseq repository Scale-HD uses for alignment be expanded considering the alleles found in our study, especially the 5'-CAG(**CAA**)$_{2\text{-}3}$CAG-3' alleles. Ideally, a non-refseq approach should be used alongside the current pipeline which could flag very rare atypical repeats. Differences between the refseq and non-refseq approaches could then be examined manually.

5.8.3 Lymphoblastoid and blood DNA

Lymphoblastoid cell lines are an effective way to preserve and maintain patient DNA (Amoli et al., 2008; Sie et al., 2009; Omi et al., 2017) without expending non-renewable sources such as blood. Lymphoblastoid lines are generated using Epstein Barr virus (EBV) with few genomic changes (Neitzel, 1986; Mohyuddin et al., 2004). Most of our WES (chapter 4) and MiSeq used early passage lymphoblastoid cell samples, however it was unclear the effect immortalisation had on CAG repeat instability. Thus, we compared MiSeq sequencing data from a subset of individuals for whom we had lymphoblastoid and blood DNA. In doing so, we showed CAG length was comparable in most samples (≤1 CAG difference in ~95% cases). However, CAG mosaicism was not significantly associated between the two DNA types. Furthermore, longitudinal lymphoblastoid samples showed marked changes to CAG instability over time in culture, with mosaicism decreasing or staying the same in all but one sample. CAG length was seen to be largely static in culture in our longitudinal samples. These results are similar to Cannella et al. where the CAG repeat was stable in lymphoblastoid lines with <~60 CAGs (Cannella et al., 2009). Other studies have reported similar findings, where lymphoblastoid CAG repeats were mostly stable (Duyao et al., 1993; MacDonald et al., 1993). A genotyping array to detect copy number variants (CNVs) did not detect clear differences between blood and lymphoblastoid DNA for most samples. It is possible time in culture results in a more clonal lymphoblastoid population through successive passages, and this may explain the CAG stability observed in our data.

Our findings indicate lymphoblastoid cells are suitable for estimating the size of the CAG repeat, although with slight differences in some cases. However, for estimating an individual’s CAG mosaicism, we would strongly recommend using blood DNA, as mosaicism in lymphoblastoids was (1) different from blood and (2) changed drastically over time in culture. Blood DNA at least partially recapitulates repeat instability from disease relevant tissues (*i.e.* brain) in HD (MacDonald et al., 1993), although instability from blood tends to be lower than seen in brain (Telenius et al., 1994). Blood DNA has been used as a proxy for somatic instability in both HD (Flower et al., 2019) and myotonic dystrophy type 1 (DM1) (Morales et al., 2016; Cumming et al., 2018; Pešović et al., 2018; Flower et al., 2019). However, to our knowledge there is currently lacking a comprehensive comparison between brain and blood DNA in HD. Obtaining the necessary post-mortem brain samples, blood and phenotype data for such a study may prove difficult, but would give insight into the usefulness of blood DNA as a proxy for somatic CAG instability in patients.

5.8.4 Atypical repeat structures and HD onset

A striking finding from our study was the near perfect segregation of interrupted and non-interrupted expanded *HTT* alleles in late and early onset HD patients, respectively. Using MiSeq, we confirmed all but one of the atypical repeat structures from WES and identified several additional pure CAG repeat alleles. Several alleles were missed from WES (4.5) likely as (1) coverage of *HTT* was variable and (2) there are innate inconsistencies in manually assessing read structure, especially for pure CAG alleles. Using MiSeq, we confirmed three alleles which we believe to be novel: 5'-CAG(**CAA**)$_2$CAGCCG-3', 5'-CAG(**CAA**)$_3$CAGCCG-3' and 5'-CAG**CAC**(CAG)$_3$**CAA**CAG-3', all of which were in expanded alleles in late HD onset individuals. Of the 438 early/late HD onset individuals sequenced, 10.7% had expanded alleles with non-canonical interruptions. Neither polyproline length nor the CCA interruption in the polyCCG tract were seen to have significant associations with age at motor onset in our models.

Several interruption species have now been described across repeat disease such as CAT interruptions in SCA1 (Chung et al., 1993; Chong et al., 1995), CAA interruptions in SCA2 (Choudhry et al., 2001; Sobczak and Krzyzosiak, 2005) and SCA17 (Zühlke et al., 2001; Maltecca et al., 2003; Gao et al., 2008), GAG interruptions in Friedreich's ataxia (Montermini et al., 1997) and AGG interruptions in fragile X syndrome (Eichler et al., 1994; Yrigollen et al., 2012). Non-canonical interruption sequences in *HTT* alleles have been described before in several HD families as rare occurrences (Goldberg et al., 1995; Pêcheux et al., 1995; Yu et al., 2000). Only more recently has it become apparent these atypical alleles are more widespread, and can affect both HD penetrance (Wright et al., 2019) and age at onset (Ciosi et al., 2019; GeM-HD Consortium, 2019).

Critically, the interruptions we identify in the current study, barring one, are all CAA. As CAA and CAG both encode glutamine, the interruptions are silent coding variants and do not affect the polyglutamine length of *HTT*. Hence, this directly implicates pure CAG length as the primary determinant of age of onset in HD, not the length of the polyglutamine, as now suggested by others (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). The non-glutamine interruption we identify is the first of its kind reported in HD to our knowledge. As a result of our findings and those from the rest of the HD field (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019), sequencing approaches such as MiSeq or third generation technologies (*e.g.* Pacific Bioscience (PacBio) platforms) are likely to become the standard for assessing the *HTT* allele. Considering allele structure may be especially important in effectively stratifying HD patients and outcome measures in clinical trial design.

Previous evidence has demonstrated pure CAG alleles have enhanced repeat instability both somatically and intergenerationally (Goldberg et al., 1995). Multiple CAA interruptions confer stability in the bacterial artificial chromosome (BAC) HD mouse model (Gray et al., 2008; Pouladi et al., 2012). Finally, evidence from other repeat diseases including DM1 (Cumming et al., 2018; Pešović et al., 2018; Tomé et al., 2018), SCA1 (Chung et al., 1993; Quan et al., 1995), SCA2 (Choudhry et al., 2001), SCA3 (Almaguer-Mederos et al., 2018), SCA17 (Maltecca et al., 2003) and fragile X syndrome (Eichler et al., 1994; Yrigollen et al., 2012; Nolin et al., 2013, 2015) suggest repeat interruptions confer stability, and their loss can lead to instability both somatically and intergenerationally. In SCA2, the configuration of the interruptions can also affect disease phenotype (Charles et al., 2007; Elden et al., 2010; Yu et al., 2011).

Repeat structures readily form non-B secondary DNA structures including slipped hairpins, G-quadruplexes and R-loops (reviewed in (Neil et al., 2017; Polyzos and McMurray, 2017; Massey and Jones, 2018; McGinty and Mirkin, 2018)). Interruptions are thought to stabilise repeat tracts and reduce the frequency of secondary DNA structures, whereas loss of interruptions can lead to increased formation of non-B DNA (Pearson et al., 1998; Rolfsmeier and Lahue, 2000; Jarem et al., 2010). It is unclear the degree to which interruptions in *HTT* contribute towards (in)stability and secondary DNA structures, however, as they occur at the 3' edge of the repeat in *HTT*, and probably affect (in)stability less than interruptions in the middle of a repeat (such as SCA1). Potential mechanisms by which repeat interruptions act are discussed generally in 6.3.2.

It is also worth noting the GeM-HD study is probably under calling the frequency of atypical *HTT* alleles. In addition to very rare interruption alleles which will not be captured from GWAS, we find about ~15% of our (CAG**CAA**)$_2$ alleles are not associated with the *NOP14* variant from chapter 4 (Arg697Cys). Thus these alleles are likely on a different *HTT* haplotype than the 4AM2 *HTT* haplotype described as significant through GWAS (GeM-HD Consortium, 2019). Equally, we find at least four *HTT* haplotypes with loss of interruptions, the most common of which, 5'-(CAG)$_n$(CCG)$_{12}$(CCT)$_2$, only constitutes ~40% of all loss of interruption alleles. Hence up to 60% of loss of interruptions may not be captured by GWAS alone.

5.8.5 Repeat instability in early and late onset

Examining CAG instability in our sample was difficult as early and late onset individuals intrinsically had significantly different ages at which samples were taken (*p*<2E-16), and similarly early onset individuals had significantly larger CAGs than our late onset group (*p*=1.13E-11). Both age and CAG length had to be added as covariates as these contribute to instability (Lee et al., 2011; Ciosi et al., 2019), and the large imbalances in sample age between early/late onset groups reduces the power to detect other effects (such as instability). A possible solution in future is to use a model of CAG mosaicism from a less extreme HD population with age and CAG length as variables. We are aware of a model which may be used for this purpose once finalised ((Ciosi et al., 2019), in press at time of writing).

Furthermore, metrics from blood DNA seemed to perform much better than those from lymphoblastoid cells, especially for progression. Reasons why this may be the case have already by explored in 5.8.3, but again indicate lymphoblastoid mosaicism metrics have limited utility outside of deriving CAG length. Future work should focus on using blood DNA to derive CAG instability measurements from HD patients. This coupled with improved modelling approaches will allow for CAG instability to be explored in multiple HD phenotypic outcomes.

This chapter has explored a targeted next-generation sequencing approach (MiSeq) to examine *HTT* structure in our early/late onset cohort. We confirmed most of the alleles from WES as well as phasing and identifying several additional interruption structures. Individuals with additional interruptions had later disease onset in all but one case, whereas loss of interruption was uniquely associated with earlier onset. Lymphoblastoid DNA was roughly equivalent to blood DNA for CAG length, but blood DNA performs better for determining useful patient somatic instability. Improved modelling and more blood DNA will allow for deeper study into how instability is associated with HD disease phenotype. The final chapter will discuss the results from the thesis generally and consider future research directions.

Chapter 6: General discussion

6.1 Summary of findings

The primary objective of this study was to find modifiers of Huntington's disease (HD) onset. Data from HD patients who participated in the Registry-HD study were used to explore different HD age at onset phenotypes, using both the clinician's best estimate of disease onset (sxrater) and the clinical characteristics questionnaire (CCQ). 500 patients with extremely early or late HD motor onset given their *HTT* CAG length were then selected (250 early and 250 late), and these individuals were whole-exome sequenced to examine rare modifiers of age at onset in HD. Non-synonymous damaging (NSD) variants were found skewed between early and late onset patients in several DNA repair genes including *FAN1*, *EXO1*, *LIG1* and *PMS1*, and loss-of-function variants in *MSH3.* Length-dependent short-tandem repeats (STRs) in *TCERG1* were also found to be associated with altered HD onset. Analyses using whole-exome burden linear regression and SKAT(-O) identified a variant in *NOP14*, found to be tagging an atypical *HTT* allele 5'-(CAG**CAA**)$_2$CAG-3', as exome-wide significant. *HTT* allele sequences were then confirmed using an independent sequencing modality, MiSeq. MiSeq revealed expanded *HTT* alleles lacking CAA interruptions uniformly had early disease onset, whereas alleles with additional interruptions had late disease onset in all but one instance. These data also identified several novel expanded *HTT* alleles, two with tandem CAA interruptions, and one with a non-glutamine interruption.

6.2 Study limitations

6.2.1 Calculating residual age at onset

Although we explored several HD onset phenotypes in chapter 3 (motor, cognitive, apathy, depression, perseveration, irritability, violent/aggressive behaviour and psychosis), age at motor onset (AMO) onset was chosen as the primary outcome measure as this represents a significant milestone for most HD patients (see 3.8.7). Some of the difficulties estimating onset in HD patients have already been discussed (3.8.2), but in summary: (1) manual curation of data where the clinician and the CCQ differed was slow, and in some cases determining which data to use was difficult, (2) slight, but systematic, differences between the rater and CCQ (mainly for cognitive symptoms) and (3) lower data quality for some Registry individuals, especially those lacking rater-derived measures of onset and onset type. Any inaccuracies estimating AMO would have affected both the stratification of individuals for whole-exome sequencing (WES) and downstream analyses (such as the linear regressions undertaken in chapters 4 & 5). Considerations for additional disease phenotypes are discussed in 6.4.1.

For calculating residual AMO, we utilised the model presented by Langbehn and colleagues ((Langbehn et al., 2004); see Equations 2.1 and 2.2) to estimate expected AMO using CAG lengths from Registry, derived from local diagnostic labs. While mean Registry AMOs matched predicted onsets closely (<1 year difference for most CAGs), after re-genotyping our sequenced individuals using MiSeq, we discovered we had sequenced a less extreme cohort than originally intended. This is as CAG lengths from local diagnostic labs were of variable quality and not systematic. Consequently, we had selected some individuals with a more typical HD onset due to mis-sizing of the CAG repeat. Although we factored these findings into our quality control (QC) pipeline and re-defined our early and late cohorts for analysis (4.3.6), we probably lessened the power of our study to detect variation associated with altered onset given our small sample size (N=500).

6.2.2 Sample size and extreme phenotype sampling

A major limitation of the current study is sample size. Although we had access to ~6000 manifest HD individuals from Registry, many of whom had DNA available, we only had the resources to sequence 500 exomes, many fewer than in recent exome studies with thousands or tens of thousands of individuals (*e.g.* (Raghavan et al., 2018; Tin et al., 2018; Flannick et al., 2019; Satterstrom et al., 2019)). Small sample sizes in sequencing studies can increase both type I (false positive) and type II (false negative) error rates (Lee et al., 2014; Wang et al., 2015; Wu and Pankow, 2016). Hence, there are almost certainly modifiers of onset for which we are currently underpowered to detect. Equally, a higher rate of type I error could have resulted in spurious association. *CUBN*, an unexpected exome-wide significant gene from the sequencing in chapter 4, may be the result of such an error, especially given *CUBN* was only significant in the burden tests and not SKAT or SKAT-O. More sequencing and experimental work is needed to determine the role (if any) of *CUBN* in HD.

We partially mitigated our small sample size by utilising an extreme phenotype sampling approach. Extreme phenotype stratification has been used in a number of mostly smaller (N<1,000) sequencing studies (Emond et al., 2012; Weeke et al., 2014; Bruse et al., 2016; Gréen et al., 2016; Scott et al., 2016; Kleinstein et al., 2018; de Carvalho-Siqueira et al., 2019) as a way to increase power, as the extremes of a phenotypic distribution are more likely to harbour variants of large effect (Li et al., 2011; Barnett et al., 2013; Peloso et al., 2016). Furthermore, additional type I inflation from extreme sampling is low when population substructure is accounted for using principal components as covariates (Luo et al., 2018; Panarella and Burkett, 2019), as we have done in the current study.

Still, while detection of variants was aided by our extreme phenotype selection, one problem we encountered was effectively modelling the effect these variants had on AMO. For instance, individuals with the *MLH1* variant (Ile219Val/rs1799977) had ~3 years later onset per copy of the variant present in our data. This is three-fold higher than reported elsewhere (Lee et al., 2017; GeM-HD Consortium, 2019), and almost certainly an overestimate stemming from our extreme phenotype approach. Similarly, individuals with atypical expanded *HTT* alleles had ~14 years earlier (no interruption) or later (two interruptions) in our model, a larger effect than in other reported studies (GeM-HD Consortium, 2019; Wright et al., 2019). As explored in chapter 5, it was also difficult to model CAG instability in our sample given the large imbalance in age between our two extreme onset groups. Models examining CAG instability and progression were similarly hampered by age-associated effects. Thus overall, while our extreme sampling approach enhanced the ability of this study to detect variation associated with disease onset, it is unclear both (1) the quantitative effect (in years hastened or delayed) identified variants may have on onset and (2) the distribution and effect of these variants in a more typical, non-extreme HD patient population. Similarly, it is possible an extreme patient recruitment strategy could have been influenced by ascertainment or survival biases. For instance, late HD onset individuals may die from age-related illnesses before coming to clinic, and thus would not be included in our study.

6.2.3 Lymphoblastoid and blood DNA

Our study primarily used DNA from low passage lymphoblastoid cells from the BioRep HD biorepository. Previous literature has reported few differences between lymphoblastoid and blood DNA (Londin et al., 2011; Nickles et al., 2012; Schafer et al., 2013). The only major difference reported for some lymphoblastoid cell lines are the frequency of copy number variants (CNVs) (Jeon et al., 2007; Nickles et al., 2012; Shirley et al., 2012; Joesch-Cohen and Glusman, 2017), and we found few CNVs calls that differed between the lymphoblastoid and blood DNA in samples that were genotyped. However, as discussed in chapter 5 (see 5.8.3), *HTT* CAG repeat instability measurements from lymphoblastoid DNA were not comparable to those from blood. Therefore, in future study we would recommend the use of blood DNA for modelling repeat instability as these are more biologically relevant (see also 5.8.4); however, for CAG length determination or WES/WGS, lymphoblastoid DNA is probably faithful.

6.2.4 General challenges of NGS and WES

Next generation sequencing (NGS) has transformed the genetics field, offering highly scalable sequencing across a range of biological and clinical applications. Despite its advantages and successes, the field still faces a number of challenges (reviewed by (Bertier et al., 2016; Hoffman-Andrews, 2017; Petersen et al., 2017; Schwarze et al., 2018; Suwinski et al., 2019)). In a research setting, these include sequencing cost, computing resources and storage, a current lack of consensus in downstream analyses (especially for rare variant analyses) and the interpretation of variants of uncertain significance (VUS).

Our relatively small sample size (N=500) is the direct consequence of sequencing cost. And even using a supercomputing cluster (see 2.13), our GATK-based alignment/variant calling pipeline alone took ~26-28 computing hours per exome. While the NGS field is developing quickly, with recent large reference studies including the 1000 genomes project (1000 Genomes Project Consortium, 2015), ExAC (Lek et al., 2016), DiscovEHR (Dewey et al., 2016) and gnomAD (Karczewski et al., 2019) helping to standardise some of the analytical procedures, the nascence of many of the current bioinformatic tools capable of handling the size of sequencing data may have restricted the interpretation of our data compared to the tools/analyses available in more mature fields (*e.g.* GWAS). The introduction of new programs and analyses specifically designed for NGS will improve the interpretability and comparability of data across the field, especially for understanding rare variants – some of these are considered in 6.4.2.

It is also important to point out while variant calling is typically robust (especially when implementing per-variant read depth QC as we have here), variants may escape detection at lower quality sites (*e.g. RRM2B* in this study), and confirmatory Sanger sequencing is still considered the gold standard (Mu et al., 2016). Indeed, while we observed no false positives in *FAN1*, WES QC removed three variants in *FAN1* confirmed by subsequent Sanger sequencing. The degree to which variants were missed by variant calling was probably influenced by our coverage; we report here an average coverage of ~30 for most of our exomes and this is lower than the 100x depth suggested for clinical exome sequencing (Suwinski et al., 2019). Higher coverage sequencing in future would enable more variants to be confidently identified.

We were also limited by the variant prediction tools we have available; as discussed in 4.11.2, we primarily focused on CADD score (Kircher et al., 2014) as it is both (1) widely used and (2) considered robust, being an combinatorial score. However, while CADD greatly helped in the prioritisation of variants, especially those that were non-synonymous but not loss-of-function, the damaging scores are still only predictive. Functional work is still needed to analyse the effect (if any) identified variants have on proteins in question (see 6.4.4). It is also worth bearing in mind sequencing can only capture variation that is present – damaging variation in the ~3000 loss-of-function intolerant genes as defined by ExAC (Lek et al., 2016) may be extremely rare or not at all present, and this includes genes such as *PCNA*, important for coordinating DNA repair machinery. Not finding variants in these genes, then, does not preclude them from a role in disease mechanism.

A further limitation of the current study comes from whole-exome sequencing (WES) itself. Although cost-effective and capable of a relatively high average coverage in coding portions of the genome, WES by its very nature has poor coverage of most non-coding regions. While coding variants are easier to interpret, the role of the non-coding genome is becoming increasingly recognised (reviewed by (Takata, 2019)), particularly in regulation; considerations for whole-genome sequencing (WGS) are discussed in 6.4.2. In addition to non-coding sequence, WES also suffers from uneven coverage in regions less amenable to capture during library preparation (Wang et al., 2017), such as seen in *RRM2B* and *SYT9* in this study. In general, regions of the exome such as those with high GC content, repeats or large exons are generally less well captured by standard exonic enrichment techniques (Wang et al., 2017).

Finally, there are also limitations of second-generation NGS technologies in general, especially in the interpretation of short tandem repeats (STRs) and other repetitive regions of the genome. While we showed some imperfect repeats (*e.g.* those in *TCERG1*) can be read through with some accuracy, overall STRs in our WES are difficult to read through natively using 75 bp reads, especially in larger repetitive motifs (as in *MSH3*) or very long repeat stretches (as in *HTT*). This may have been exacerbated further due to the number of PCR cycles needed to prepare the WES libraries. Hence the interpretation of STR regions is difficult using our WES alone. MiSeq sequencing offers the longest reads of any second-generation sequencing platform, with its version 3 chemistry capable of 600 sequencing cycles. However, as previously explored in 5.8.2, MiSeq still suffers from (1) limitations of sequencing by synthesis, especially PHRED score drop off towards the 3’ end of sequence, (2) PCR amplification biases in STRs (although far fewer cycles than in WES) and (3) somewhat limited bioinformatic tools for the interpretation of STRs in targeted sequencing, especially where there may be novel interruption sequences. MiSeq is also much lower throughput compared to other Illumina platforms (*e.g.* HiSeq or NovaSeq), and hence per base WES/WGS is relatively expensive using MiSeq. Improvements to the analysis of STRs in second-generation NGS are discussed in 6.4.2; and third-generation sequencing technologies are briefly considered in 6.4.3.

6.3 Mechanisms underlying HD onset

6.3.1 An overall model for onset

Our study has further confirmed the role of the *HTT* DNA itself in onset modification of HD. Both damaging variants in several DNA repair genes and atypical *HTT* repeat sequences were associated with altered onset in HD patients. Using our data, and supported by other human genetic (Holbert et al., 2001; GeM-HD Consortium, 2015; Bettencourt et al., 2016; Hensman Moss et al., 2017; Lee et al., 2017; GeM-HD Consortium, 2019; Ciosi et al., 2019; Flower et al., 2019; Wright et al., 2019) and functional work (Holbert et al., 2001; Arango et al., 2006; Dragileva et al., 2009; Bourn et al., 2012; Tomé et al., 2013; Neueder et al., 2017; Zhao et al., 2018; Franich et al., 2019; Goold et al., 2019), we constructed a model containing mechanisms affecting HD onset, shown in Fig. 6.1.

While this model is by no means exhaustive, it does present the major modification mechanisms currently known in HD: (1) repeat sequence (a cis-modifier) , (2) members of DNA repair pathways, mostly mismatch repair components (trans-modifiers) and (3) other less-defined factors that may affect downstream pathology. Both mechanisms (1) and (2) focus on the dynamic instability of the *HTT* CAG, which we suggest are acting through somatic instability, although other interaction mechanisms could be at play (Maiuri et al., 2019). Interruptions and the presence of FAN1 help to stabilise the repeat, reducing the rate of expansion. MSH6 may also promote CAG stability by outcompeting MSH3 binding of MSH2. Conversely, aberrant handling of repeats by other DNA repair factors act to promote repeat expansion. Repeat instability (mainly expansion) of the expanded *HTT* allele then occurs over life, modified by the mechanisms in (1) and (2); longer CAG repeats probably contribute towards disease by the generation of longer (and more toxic) RNA and/or protein. The mechanisms of (1) and (2) are explored more fully in 6.3.2 and 6.3.3, respectively.

We also indicate a possible role for TCERG1 as a regulator of splicing and transcription of *HTT*. As suggested elsewhere (Sathasivam et al., 2013; Neueder et al., 2017; Franich et al., 2019), incomplete splicing of *HTT* can lead to the generation of small, highly toxic HTT exon 1 fragments that may be a contributory factor in disease pathogenesis. This pathology could additionally be mediated by mechanisms of RNA toxicity (Schilling et al., 2019). It may be, however, that TCERG1 is acting through an entirely separate mechanism (*e.g.* expression/transcription of other modifiers). We have also included SYT9 and GPR151 as additional putative downstream modifiers found through GWAS (GeM-HD Consortium, 2019), although we did not find significant enrichment of non-synonymous damaging (NSD) variants in either gene. These modifiers are briefly discussed in 6.3.4.

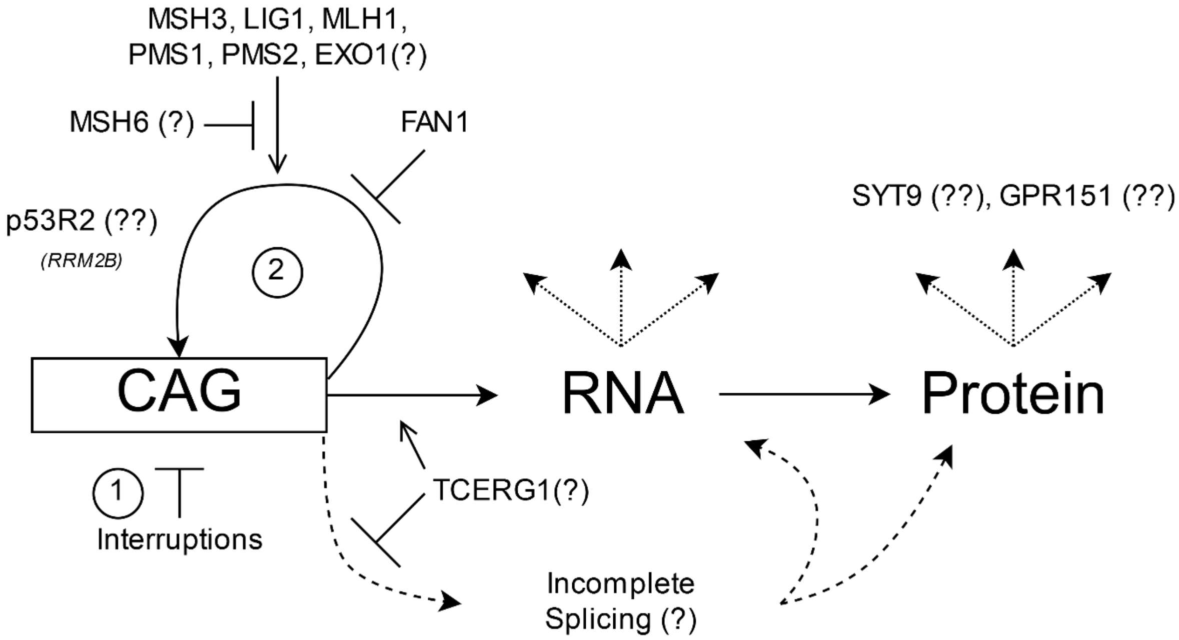


Figure 6.1: Proposed mechanisms for onset modification in HD. Here, we show mechanisms which may alter onset of Huntington's disease. In (1), interruptions in *HTT* sequence act to stabilise the CAG repeat in DNA (see 6.3.2 and Fig. 6.2). In (2), DNA repair and metabolism can act to promote or supress expansion (see 6.3.3 and Fig. 6.3). Also shown is a postulated role for TCERG1 in the splicing and/or transcription of *HTT*. SYT9 and GPR151 are shown as downstream mechanisms of modification, although these are at present poorly described. The dotted arrows in RNA and protein indicate mechanisms by which pathology is mediated downstream (see 1.5 and Fig. 1.6). Blunted arrows are mechanisms acting to slow down or inhibit a process. Note p53R2 is encoded by *RRM2B*.

6.3.2 Allelic interruptions and their contribution to HD onset

Atypically interrupted expanded *HTT* alleles comprise between ~5% (GeM-HD Consortium, 2019; Wright et al., 2019) to 10.7% (this study) of HD alleles, although the prevalence in our study is highly enriched due to examining extreme onset phenotypes. In patients with these non-canonical alleles, interruptions represent a major contributory factor towards disease onset. As previously discussed in 5.8.4, interruptions confer repeat stability across a range of repeat diseases including several spinal cerebellar ataxias (SCAs) (Chung et al., 1993; Choudhry et al., 2001; Maltecca et al., 2003; McFarland et al., 2014, 2015; Almaguer-Mederos et al., 2018), myotonic dystrophy types (DM) 1 and 2 (Bachinski et al., 2009; Cumming et al., 2018; Pešović et al., 2018), and Fragile X syndrome (Eichler et al., 1994; Yrigollen et al., 2012). Conversely, repeats lacking interruptions are instead unstable, both vertically and somatically (Goldberg et al., 1995; Wright et al., 2019). Indicated in Fig. 6.2 is a model indicating how interruptions may be contributing functionally towards disease pathology. The model shows two alleles with the same polyglutamine length but differing numbers of pure CAG. These alleles have different propensities to form non-B DNA, as secondary DNA structures are more energetically viable in uninterrupted, longer pure repeat tracts (Gacy et al., 1995; Grabczyk and Usdin, 2000; Napierala et al., 2005; Sobczak and Krzyzosiak, 2005).

The formation of repeat non-B DNA probably represents a critical step in the molecular pathology of HD, and secondary DNA structures may affect CAG repeat instability in several ways. For instance, MutSβ (MSH2-MSH3) binds to CAG hairpin loops (Owen et al., 2005; Burdova et al., 2015; Guo et al., 2016), and improper resolution by mismatch (and potentially other) repair machinery may directly act to promote expansion, as laid out in 6.3.3. It has also been suggested that non-B DNA may be more susceptible to damage; for instance, guanine in Z-DNA is more damaged by ionising and oxidative sources of DNA damage than in B-DNA (Ribeiro et al., 1992; Tartier et al., 1994). Additionally, the modification of guanine to 8-oxoguanine (8oxoG) by reactive oxygen species can affect the conformation of slipped hairpins (Volle et al., 2012; McCauley et al., 2018), and this may affect the handling of these structures by DNA repair. R-loops, RNA:DNA hybrids which transiently arise during transcription, may also contribute towards instability of CAG repeat sequences (Lin et al., 2010; Reddy et al., 2011, 2014; Su and Freudenreich, 2017; Freudenreich, 2018), although it is currently unclear whether repeat interruptions would influence the formation or stability of these structures in the same way as cruciforms or slipped hairpins. However, a role for R-loops may be compatible with the model presented in 6.3.3 (Fig. 6.3), wherein secondary DNA may form on the unbound single stranded DNA which could persist following R-loop resolution. These could then interface with DNA repair.

While probably less relevant in the neurodegeneration that occurs in HD post-mitotic neurones, secondary DNA structures also represent major challenges for replicative machinery in dividing cells (Liu et al., 2010a, 2013a). Issues during replication may lead to fork collapse and necessitate DNA repair (reviewed in (Polyzos and McMurray, 2017) and (McGinty and Mirkin, 2018)), and these mechanisms could affect intergenerational repeat mutability. It is also possible repeat interruptions may additionally guard against the formation of repetitive RNA hairpins, as these can occur in CAG repeat sequences (Sobczak et al., 2003; Kiliszek et al., 2010; Yildirim et al., 2013; Tawani and Kumar, 2015) and may affect disease pathology through various mechanisms of RNA toxicity (Li et al., 2008; de Mezer et al., 2011; Hsu et al., 2011; Bañez-Coronel et al., 2012; Rué et al., 2016; Urbanek et al., 2016; Jain and Vale, 2017).

An integral question is whether repeat interruptions, such as those we find in the current study, are associated with altered onset simply due to mis-sizing of the pure CAG repeat, or whether interruptions have an additional stabilising effect on the repeat. For instance, would repeat instability in a pure $(CAG)_{44}$ allele and a doubly interrupted $(CAG)_{44}$(**CAA**CAG)$_2$ be the same, given they both have the same pure CAG length? Our data suggests interruptions may exert an additional stabilisation effect even after correcting residuals for the pure CAG length in alleles with additional interruptions (5.3.2), but the age imbalances in our cohort make the interpretation of these data difficult. The recent GeM-HD study did not find this effect to be genome-wide significant through GWAS following pure CAG length correction (GeM-HD Consortium, 2019). More study of *HTT*-centric 3' repeat interruptions (for instance, *in vitro* modelling) will elucidate the mechanistic role of interruptions in HD disease pathology.

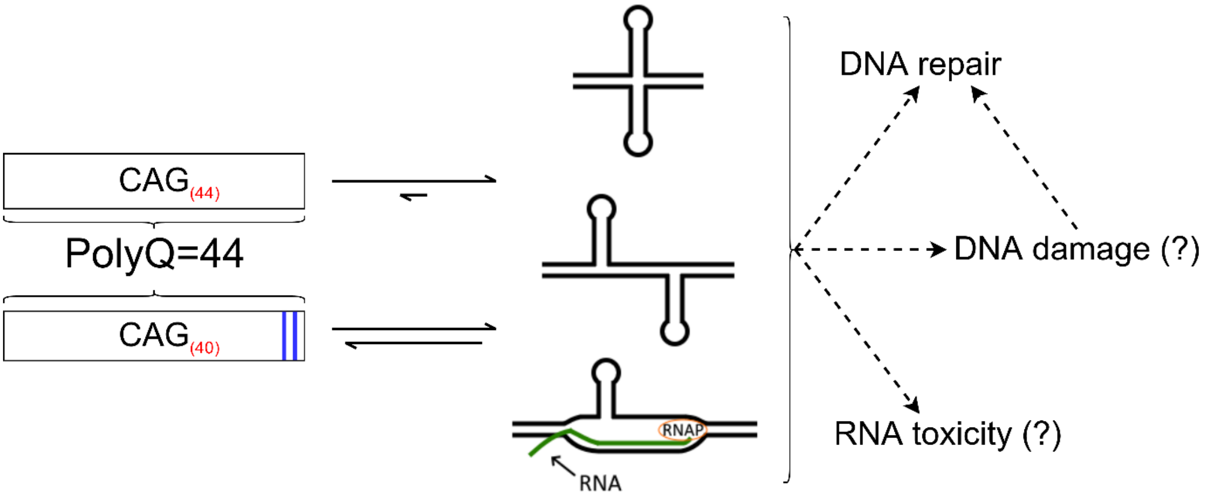


Figure 6.2: Repeat sequence and secondary DNA structure. Shown on the left are two *HTT* repeat sequences: $(CAG)_{44}$ (top) and CAG_{40}(**CAA**CAG)$_2$ (bottom). The blue bars in the bottom CAG_{40}(**CAA**CAG)$_2$ allele are CAA interruptions. Both alleles have the same polyglutamine (polyQ) length, however different pure CAG lengths. Shown on the right are secondary DNA structures; from top to bottom, cruciform DNA, slipped hairpin DNA and an R-loop (RNA:DNA hybrid) with a slipped hairpin formed on the single-stranded DNA. As indicated by the central arrows, longer pure CAG tracts (without interruptions) are more readily able to form secondary DNA structures. Secondary DNA could influence pathology by through DNA damage/repair, leading to subsequent somatic instability. Possibly, secondary RNA hairpins could exert alternative toxic effects through sequestration of RNA-binding proteins. RNAP: RNA polymerase.

6.3.3 DNA repair mechanisms may stabilise or destabilise repeats

In Fig. 6.3, I present my two-fated pathway model for repeat expansion. This model is focused on mechanisms underlying DNA repair occurring in post-replicative cells, given neurodegeneration in HD occurs in terminally differentiated neurones (Gonitel et al., 2008). For a recent review of replicative pathways implicated in repeat expansion in replicative cells, see (McGinty and Mirkin, 2018). As with most DNA repair pathways, my presented pathway has four main steps: damage recognition (1), excision and removal of DNA damage (2-4a/b), re-synthesis (5) and ligation (5). These steps are explored more fully below.

First (Fig 6.3, (1)), a slipped hairpin is formed. As already explored in 6.3.2, hairpins (and other secondary DNA) can form intrinsically in repetitive sequence (Gacy et al., 1995; Mitas et al., 1995; Grabczyk and Usdin, 2000; Pearson et al., 2002; Napierala et al., 2005; Sobczak and Krzyzosiak, 2005; Liu et al., 2010a). It is also possible hairpins could arise from single-stranded DNA in an R-loop, as suggested by Freudenreich (Freudenreich, 2018) and others, and hairpins may persist following R-loop resolution. Regardless of origin, here we indicate the mismatch repair heterodimer MutSβ (MSH2-MSH3) binding the secondary DNA hairpin structure, as MutSβ is known to bind hairpins in repetitive DNA (Owen et al., 2005; Burdova et al., 2015; Guo et al., 2016). As discussed in 4.11.6, our study identified novel loss-of-function variants in *MSH3* associated with late disease onset, which is further supported by recent studies indicating *MSH3* is a modifier of HD onset (Flower et al., 2019; GeM-HD Consortium, 2019) and HD progression (Hensman Moss et al., 2017). *Msh3* also drives repeat expansion in a dose-dependent manner in mouse models of HD (Dragileva et al., 2009; Tomé et al., 2013).

We also suggest a possible minor role for MSH6 in onset modification (see Fig 6.1), potentially through competitive binding of MSH2, although MSH6 was not significant in the current study or by GWAS as onset modifying (GeM-HD Consortium, 2019). Where MutSβ binds insertion/deletion loops up to 14 bp in size *in vitro* (Habraken et al., 1996; Palombo et al., 1996), and smaller (~3 CAG) slip-outs in repetitive DNA (Panigrahi et al., 2010; Zhang et al., 2012), MutSα (MSH2-MSH6) instead binds small 1-2 bp mismatches with high affinity (Drummond et al., 1995; Palombo et al., 1995; Acharya et al., 1996). However, whether MSH6 promotes or protects from repeat expansion is currently unclear, and this may be context dependent. For instance, *MSH6* expression seems to protect against repeat expansion in a yeast (Kantartzis et al., 2012) and human cell system (Nakatani et al., 2015). Similarly, *Msh6* expression appears to protect against somatic CTG repeat expansion in a DM1 mouse model (van den Broek et al., 2002). However, *MSH6* expression promoted repeat expansion in other repeat disease models (Bourn et al., 2012; Du et al., 2012; Zhao et al., 2016), and *Msh6* knockdown in a DM1 mouse decreased maternal intergenerational repeat expansion (Foiry et al., 2006). Thus, the role of *MSH6* we have suggested is largely speculative, and it is currently unclear as to the specific role of *MSH6* (if any) in modifying HD onset in humans.

Shown next (Fig 6.3, (2)) is the recruitment of MutL to the hairpin by MutSβ, followed by MutL-mediated endonucleotide cleavage of DNA. Note while we have suggested a 3' DNA nick in (2), it is possible additional nicks may occur 5' to the hairpin either additionally or alternatively (and possibly utilising other endonucleases), as mismatch repair can proceed 5'- or 3' (Hsieh and Zhang, 2017). As for the specific MutL complex, both MutLα (MLH1-PMS2) and MutLβ (MLH1-PMS1) complexes are probably involved, as in the current study we find rare damaging variation in *PMS1* associated with late onset, as well as a common variant associated with altered onset in *MLH1* (the same found in (Lee et al., 2017) and (GeM-HD Consortium, 2019)). GWAS has identified all three dimer components of MutLα and MutLβ (PMS1, PMS2 and MLH1) as modifiers of onset (GeM-HD Consortium, 2019). These data are interesting given MutLβ is not known to have a role in canonical mismatch repair (Räschle et al., 1999; Cannavo et al., 2005); this suggests a novel mechanism through which repair is proceeding.

Although excess variants in *MLH3* were not found to be associated with altered onset by this study or by GWAS (GeM-HD Consortium, 2019), there may be additionally be a role for MutLγ (MLH1-MLH3). Indeed, evidence from HD mice show MutLγ is a powerful mediator of somatic instability (Pinto et al., 2013), and a similar finding was recently reported in Fragile X mice (Zhao et al., 2018) and a Fredrich's ataxia disease human cell model (Halabi et al., 2018). Interestingly, another recent study in yeast found a role for MutLγ in R-loop-dependent CAG repeat instability (Su and Freudenreich, 2017); again, such a mechanism (R-loop -> hairpin -> DNA repair component binding) may be feasible.

The next step is the recruitment of further downstream effectors by MutL (Fig 6.3, (3)), and here we have indicated competitive binding of MLH1 by FAN1 and EXO1. Our data show rare, late-associated damaging variants in the MSH3-interaction domain of EXO1 (suggesting a normally deleterious role for EXO1:MSH3 interaction). Conversely, we identify rare, early-associated variants in FAN1's DNA-contacting SAP and VRR-nuclease domains (suggesting a normally protective effect for FAN1 DNA handling). Hence, we believe this to be the step that decides the pathway fate: either an error-free FAN1-driven pathway (4a-5a) or an error-prone EXO1-driven pathway (4b-5b).

In support of this concept, FAN1 is known to directly interact with most MutL heterodimer components: MLH1, MLH3, and PMS2 (Cannavo et al., 2007; Kratz et al., 2010; Smogorzewska et al., 2010), and possibly indirectly with PMS1 (Smogorzewska et al., 2010). This also lends support to the notion discussed that all MutL complexes (α, β and γ) could play a role in hairpin resolution, although only MutLα and MutLβ in humans have been implicated by this study and others (Lee et al., 2017; GeM-HD Consortium, 2019). Crucially, MLH1 is required for FAN1-mediated stability in an HD mouse model (Wheeler VC, personal communication), further indicating MutL driven recruitment of FAN1 is a critical step in the pathway. EXO1, on the other hand, is the canonical 5'->3' nuclease involved in mismatch repair (reviewed in (Fishel, 2015)). EXO1 interacts with MLH1, MSH2 and MSH3, and is similarly recruited by MutL in mismatch repair (Tishkoff et al., 1997; Schmutte et al., 1998, 2001). It is notable FAN1 and EXO1 both possess 5'->3' exonuclease activity and 5'-flap endonuclease activity, and it has been reported there is redundancy between the two nucleases, with FAN1 compensating for Exo1 loss in a mouse model (Desai and Gerson, 2014).

Following this, the pathways diverge. In the FAN1-centric error-free pathway (4a), FAN1 is effectively able to carry out the removal of the secondary DNA hairpin without expansion. Although it is not currently known how this occurs, it may proceed in several ways. For instance, FAN1 could initiate repair and then act as a scaffold to recruit additional effectors. These putative effectors, however, are not known, but could include alternative exonucleases such as FEN1 (which has been shown to modify repeat expansions before (Liu and Bambara, 2003; Liu et al., 2009)), or the structure specific endonuclease MUS81 as suggested by (MacKay et al., 2010) and (Pizzolato et al., 2015) in the context of interstrand crosslink repair. Equally, this may even include Fanconi anaemia proteins, with which FAN1 is known to interact (MacKay et al., 2010). Indeed, a blended mismatch-interstrand crosslink repair pathway is appealing given the known overlap between the two pathways. For instance, MutS can bind to cisplatin DNA interstrand crosslinks (Yamada et al., 1997), and MLH1 loss can sensitise cells to crosslinking agents (Williams et al., 2011). It has recently been demonstrated that repair, driven by MutS, MutL and EXO1, could repair interstrand crosslinks *in vitro* (Kato et al., 2017).

Alternatively, or additionally to the above concept, FAN1 could act by the mechanism suggested by Zhao and colleagues, whereby FAN1 dimerises and can locally unwind the DNA in interstrand crosslinks (Zhao et al., 2014). So, in similar fashion, FAN1 may be able to resolve secondary hairpins by local DNA unwinding of the atypical non-B DNA. Given we have found damaging variants in the exonuclease domain of FAN1 associated with early disease, we suggest the exonuclease activity of FAN1 is used for excision of hairpin-containing DNA in (4a).

In the error-prone pathway mediated by EXO1 (4b), we suggest a similar mechanism to that of canonical mismatch repair. EXO1 is recruited to the site of damage by MutL where EXO1 can excise the hairpin-containing DNA using its 5'->3' exonuclease activity. However, this pathway is unable to effectively deal with the hairpin, and this instead persists during repair. Why this may be the case is unknown, but possibly due to substrate differences between EXO1 and FAN1, EXO1 is unable to properly handle the repeat structure as FAN1 (or its effectors) does. For instance, EXO1 is not able to process interstrand crosslink-containing DNA substrates *in vitro* whereas FAN1 can (Pizzolato et al., 2015). Small DNA hairpins may also enhance EXO1 activity and excision, as shown recently (Li et al., 2019). As mentioned in 4.11.5, the recent report of EXO1 protecting against expansions in a murine Fragile X model (Zhao et al., 2018) we believe indicates an alternative, protective role for EXO1 distinct from the MSH3-driven mechanism we describe here.

Both (4a) and (4b) result in the resolution of hairpin containing DNA, either successfully or unsuccessfully, by re-synthesis by a polymerase and re-ligation. The canonical mismatch repair polymerase Polδ could be involved (Longley et al., 1997). But equally, given the non-canonical nature of the described pathway, Polβ, typically involved in base-excision repair (Matsumoto and Kim, 1995; Podlutsky et al., 2001), or Polθ, as in interstrand crosslink repair and microhomology-mediated end joining (Beagan et al., 2017; Wang et al., 2019), may instead be involved. The choice of polymerase may also depend on the pathway (either (5a) or (5b)). Interestingly, *LIG1* has been identified both in this study and by GWAS (GeM-HD Consortium, 2019) as a modifier of disease onset, thereby implicating re-ligation as an important mechanism in hairpin resolution. Possibly, as we find mostly early-associated damaging variation in the LIG1 ligase A N (DNA binding) and A M (nucleotidyltransferase) domains, slower re-ligation by LIG1 may be deleterious. Potentially, slow ligation could allow for a DNA hairpin to reform on nascent, un-ligated DNA. Aberrant gap-filling by polymerases could then incorporate the hairpin into DNA, leading to repeat expansion.

Notably, we suggest there could be a degree of overlap between mismatch and interstrand crosslink repair pathways, possibly indicating a substrate overlap between interstrand crosslinks and smaller hairpins/loop outs in repetitive DNA. However, this does not preclude other DNA repair pathways as being involved in repeat expansion, although these are not currently implicated in HD patients. For instance, a hybrid base excision-mismatch repair pathway was recently described in the context of repeat expansion (Guo et al., 2016; Lai et al., 2016), and the base excision glycosylase *OGG1* has been previously implicated in HD (Kovtun et al., 2009; Budworth et al., 2015), although we did not find clear segregation of damaging variants in *OGG1* in our data.

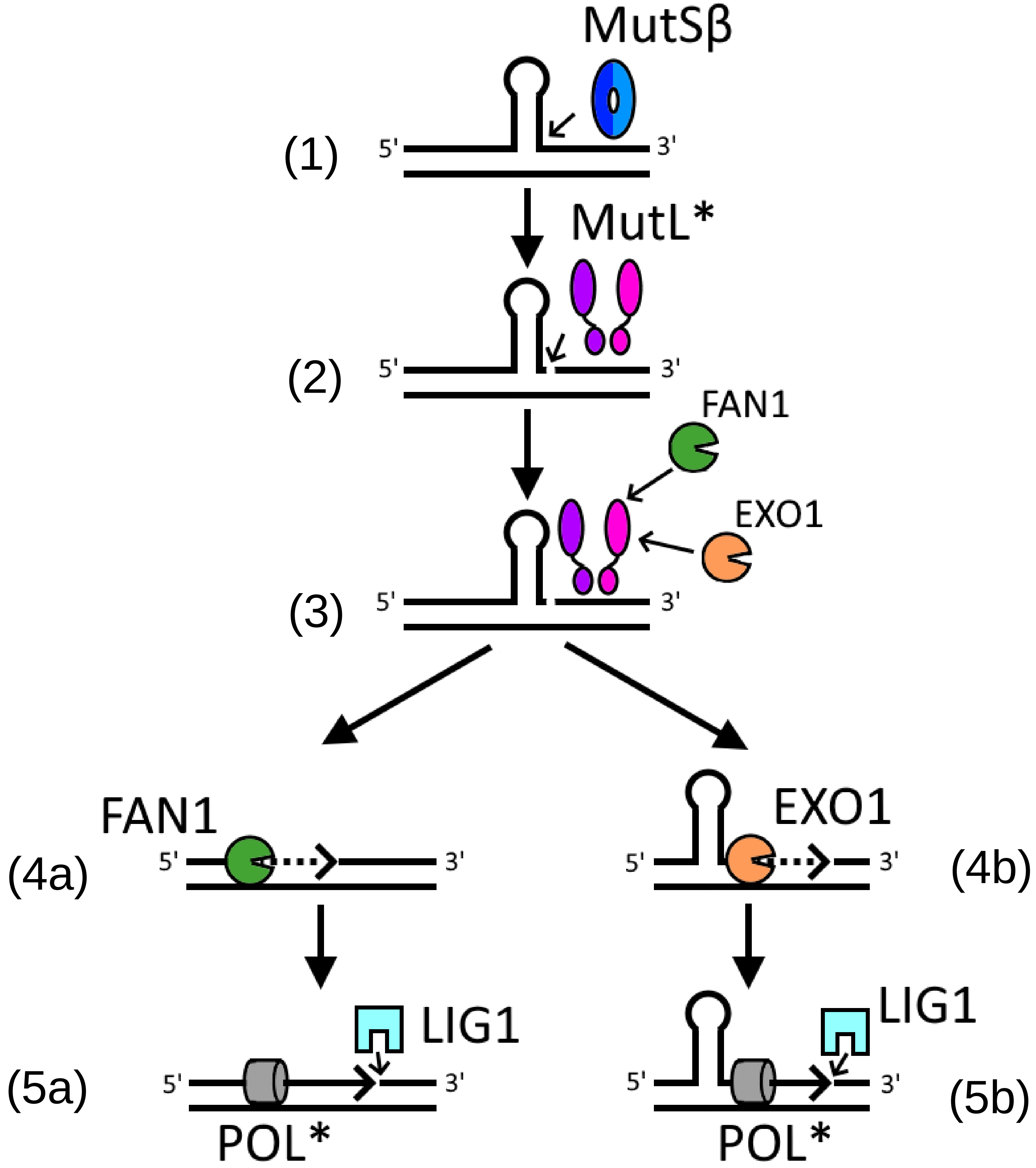

MutSβ
(1)
5'
3'
MutL*
(2)
5'
3'
FAN1
EXO1
(3)
5'
3'
FAN1
(4a)
5'
3'
EXO1
5'
3'
(4b)
LIG1
(5a)
5'
3'
POL*
LIG1
5'
3'
(5b)
POL*

Figure 6.3: A two-fated pathway model for DNA repair in HD.

(1) *Structural recognition*: A secondary DNA structure (in this case a DNA slipped hairpin) is recognised and bound by the MutSβ (MSH2-MSH3) heterodimer. The secondary structure may have arisen intrinsically, or possibly during the resolution of an R-loop.

(2) *Endonucleotide cleavage*: MutSβ recruits the MutL heterodimer. An endonuclease nicks the 3' end of the top hairpin, shown in this case by MutL. The * on MutL denotes it is unclear which MutL complex(es) participate. MutLα (MLH1-PMS2); MutLβ (MLH1-PMS1); MutLγ (MLH1-MLH3).

(3) *Competitive binding of FAN1/EXO1*: MutL recruits downstream exonuclease effectors, shown here as FAN1 and EXO1. The exonuclease recruited determines the pathway fate.

(4a) *Expansion-free resolution by FAN1*: FAN1 exonuclease activity excises nicked DNA without expansion occurring, possibly through local unwinding of secondary structure or entire excision of non-B DNA. Additionally (or alternatively), FAN1 may act as a scaffold for further effectors involved in repair.

(5a) *Re-synthesis and ligation*: DNA is successfully re-synthesised with a polymerase (possibly POLβ or POLθ) and ligated by LIG1 with no expansion occurring.

(4b) *Error-prone resolution by EXO1*: EXO1 successfully excises nicked DNA *via* its exonuclease activity but is unable to resolve the secondary DNA present. Other 5'->3' exonucleases or factors may have a role.

(5b) *Re-synthesis and ligation*: DNA is re-synthesised with a polymerase (possibly POLβ or POLθ) and ligated by LIG1, but the secondary structure escapes DNA repair and is instead incorporated in the resultant DNA, leading to a repeat expansion.

6.3.4 Contribution of downstream modifiers of HD onset

Notably, our model in Fig. 6.1 primarily focuses upstream modifiers of *HTT* DNA as these were identified both in this sequencing study and by others (GeM-HD Consortium, 2015, 2019; Bettencourt et al., 2016; Hensman Moss et al., 2017; Lee et al., 2017; Ciosi et al., 2019; Wright et al., 2019). Hence while downstream modifiers probably exist, most of these have yet to be identified. Possible downstream modifiers of HD onset include the loci containing *SOSTDC1* (sclerostin domain containing 1) (Chao et al., 2018), *SYT9* (synaptotagmin 9) and *GPR151* (G protein-coupled receptor 151) (GeM-HD Consortium, 2019). But the mechanism by which these putative modifiers could be operating is currently unknown, and further work is necessary to confirm the gene responsible for the signals observed at these loci (Chao et al., 2018; GeM-HD Consortium, 2019). As previously mentioned in 4.11.7, it is interesting to consider *SYT9* as (1) we identified a single damaging variant skewed in patients to early onset (6:1) and (2) Syt9 is highly expressed in mouse brain striatum (Xu et al., 2007).

We do, however, suggest a possible role for TCERG1 in the splicing and/or transcription of *HTT* in Fig 6.1. In our WES data, shorter glutamine-alanine tracts in TCERG1 were associated with delayed HD onset, and TCERG1 has been implicated as a modifier of HD previously (Holbert et al., 2001; Arango et al., 2006). TCERG1 is a transcription elongation factor, with high expression in the brain (GTEx Consortium, 2017), that regulates elongation and splicing. The protein localises to nuclear speckles (Sánchez-Hernández et al., 2012) and binds RNA polymerase II (Goldstrohm et al., 2001; Liu et al., 2013b). Work has shown the Q-A domain seems to be necessary for proper localisation to occur (Miller et al., 2016), so it is possible coding STR variants may modify *TCERG1*'s ability to localise to nuclear speckles, and this could affect splicing of either *HTT* or other disease-relevant genes. Aberrant splicing of *HTT* has been implicated in HD (Sathasivam et al., 2013; Neueder et al., 2017; Franich et al., 2019), and the generation of toxic terminal exon 1 HTT fragments may contribute towards disease pathogenesis (Mangiarini et al., 1996; Hazeki et al., 1999; Wang et al., 2008; Barbaro et al., 2015). Interestingly, Holbert and colleagues showed HTT interacts with the TCERG1 protein (Holbert et al., 2001); the STR variants we identify could also be acting *via* modulation of this interaction, and possibly HTT and TCERG1 proteins could cooperate in transcription.

6.4 Future directions

6.4.1 Additional disease phenotypes

While onset of motor symptoms represents a significant milestone for most HD patients, future genetic study should consider other disease phenotypes. Different phenotypes may be influenced by other disease-relevant factors; indeed, this is exemplified by the findings in chapter 3 where CAG size was seen to influence symptom age at onset to varying degrees (especially for depression and psychosis). Recently, it was shown there was a degree of overlap in genetic architecture between some of the psychiatric and cognitive symptoms experienced by HD patients and those in other neurological and psychiatric disease (Ellis et al., 2019), further indicating modifiers of different HD symptom onsets may be driven by different factors. As well as investigating onset types independently, one could model multiple onset phenotypes at once using multi-SKAT (Dutta et al., 2019) (see 6.4.2).

Phenotypes not directly related to age at onset are also highly relevant. The progression measure from Hensman Moss and colleagues (Hensman Moss et al., 2017), modelled in chapter 5 with respect to MiSeq CAG instability, is a composite measure that can provide insight into patient disease trajectory. Their study found *MSH3* is a significant modifier of disease progression through GWAS despite a small sample size (N=2,078) (Hensman Moss et al., 2017), demonstrating the utility of detailed phenotypic measures. Furthermore, it is possible there may be novel modifiers of progression (whether a composite score or otherwise) that are not at all or are poorly captured using motor onset of disease alone. Intergenerational changes to CAG repeat length would also be worth pursuing, as modifiers of vertical CAG transmission may implicate meiotic replicative machinery involved in repeat expansion. As explored in 3.8.5, anticipation (the difference in onset between the affected parent and offspring) could provide an estimate for vertical repeat expansions, although genetic differences between the parent and offspring could make interpretation of anticipation complex.

Disease penetrance may also be worth considering. A major modifier of HD penetrance has been shown to be *HTT* allele structure (Wright et al., 2019), where HD individuals in the partial penetrance 36-39 CAG range are more likely to have disease when in possession of a pure CAG *HTT* allele. However, many individuals with normally interrupted *HTT* alleles can still develop disease, and it is likely these people harbour HD modifiers. Understanding these modifiers would further our understanding the mechanisms contributing towards disease.

6.4.2 Additional sequencing and analysis

This study has shown HD genetic architecture is amenable to whole-exome rare-variant analyses, even in a relatively small cohort of HD patients (N=500). Thus, further WES with a larger sequencing cohort (>1,000) may have sufficient power to detect variation exome-wide in novel genes/gene sets. An extreme phenotype approach, as employed by this study, would further augment power, although may make modelling the effect variants have on phenotype difficult (see 6.2.2). With enough numbers, it would be possible to forgo extreme phenotype sampling, and instead sequence a more representative HD population. Centrally measured CAG sizes, now available in both Registry-HD and Enroll-HD, will improve patient stratification (if extreme phenotype sampling is employed) and residual age at onset calculation, as this should minimise mis-sizing of the CAG repeat. Until sequencing of the *HTT* CAG repeat is standard for HD patients, however, re-genotyping individuals using MiSeq or Sanger sequencing is still necessary as (currently) centrally measured lengths do not consider *HTT* allele structure.

Further to additional WES, newer tools and analyses will improve downstream interpretability of sequence data. For instance, the newest version of Hail (Hail Team, v0.2) natively supports linear mixed models and several non-burden tests. Linear mixed modelling approaches may be especially useful in Enroll-HD as there are many more patients with non-European ancestries compared to Registry-HD (which is predominantly European). Similarly, various additions to the general sequence kernel association test (SKAT) framework have been proposed such as SMAAT/GMAAT ((Chen et al., 2019); https://github.com/lin-lab/GMMAT) and multi-SKAT ((Dutta et al., 2019); https://github.com/diptavo/MultiSKAT). Multi-SKAT allows the weighting of multiple disease phenotypes and would be very useful to consider where several HD phenotypes are available (see 6.4.1).

It is also worth considering whole-genome sequencing (WGS) in future work as the non-coding genome is better understood (reviewed by (Gloss and Dinger, 2018) and (Takata, 2019)). Although WGS is currently financially and computationally expensive, WGS does offer several advantages over WES: (1) globally more consistent coverage (as there are no hybridisation or capture steps), (2) coverage of non-coding regions, (3) PCR-free approaches and (4) the ability to investigate STRs occurring across the genome. Given the uniformity of WGS coverage, *HTT* allele sequence may be resolved more easily than in WES, and WGS has been suggested as a way to derive *HTT* allele sequence by others (Wright et al., 2019). Indeed, novel repeat diseases have recently been elucidated using

Illumina-based WGS (Cortese et al., 2019; LaCroix et al., 2019; van Kuilenburg et al., 2019). Interpretation of WGS data is still more difficult than that of coding data, but several non-coding annotation tools have been developed including EIGEN (Ionita-Laza et al., 2016), LINSIGHT (Huang et al., 2017), FUN-LDA (Backenroth et al., 2018) and PAFA (Zhou and Zhao, 2018). Being an unsupervised approach, CADD can also provide scores for non-coding regions (Rentzsch et al., 2019).

Improvements to STR calling technology will also benefit sequencing in HD, especially given STRs in *TCERG1* (as shown in this study) and in *MSH3* (Flower et al., 2019) are associated with altered HD onset. Newer STR genotyping tools such as HipSTR (Willems et al., 2017), STRetch (Dashnow et al., 2018), TREDPARSE (Tang et al., 2017), ExpansionHunter (Dolzhenko et al., 2017) and adVNTR (Bakhtiari et al., 2018) will improve STR calling from sequence. The adVNTR tool is particularly interesting given it can be applied to both short- and long-read WGS in the read-through of very long repeating units in DNA. These STR tools should ideally be used with WGS data. Additionally, the introduction of longer 100-250bp WES/WGS kits and compatible platforms (*e.g.* Illumina's NovaSeq) may enable shorter STRs to be read-through without the need for additional calling tools. We would advocate for the integration of standard WES/WGS variant calling pipelines with STR calling, as together these may find novel variation underpinning HD onset.

6.4.3 Considering other sequencing approaches

Although short-read second-generation NGS (*i.e.* Illumina platforms) are well-described and widely used, there are other sequencing modalities which should be considered. Third-generation sequencing using either Pacific Biosciences or Oxford Nanopore platforms routinely offer reads of >10kb (reviewed in (Mantere et al., 2019)). Long-reads offer a powerful way to dissect repetitive tracts of DNA, and have been used to describe repeat interruptions in DM1 (Ardui et al., 2018) and new repeat diseases (Ishiura et al., 2018; Mizuguchi et al., 2019; Sone et al., 2019; Tian et al., 2019). Long-read technologies will be especially useful in HD model systems (>100 CAGs) or other non-CAG repeat diseases (100s of repeats) where complete read-through of repeats is not possible using short reads due to the repeat size.

Single-cell sequencing approaches may also prove useful in future work. To our knowledge, single-cell sequencing is a relatively unaddressed area of research in HD, and is in the unique position to interrogate why certain cells are vulnerable to somatic instability and early degeneration in HD. A recent bulk RNA sequencing study using isogenic human HD pluripotent stem cells identified several CAG-associated transcriptional phenotypes in disease relevant neuronal cells (Ooi et al., 2019). Single-cell transcriptional RNA sequencing would offer further granularity (*i.e.* cell-specific RNA expression profiles) and could be applied to both cells and tissues. Furthermore, single-cell sequencing would allow for the investigation of cell-specific effects of disease modifiers identified by our WES (*e.g. FAN1* or *MSH3*) in a model system. Single-cell sequencing that captures both RNA and DNA from the same cell, as described by Macaulay and colleagues (Macaulay et al., 2015, 2016), is especially appealing in HD (and other repeat disease) as this may offer critical insight into how CAG length can affect the transcriptome on a single-cell basis.

6.4.4 Downstream analysis of implicated variants

A crucial advantage of sequencing over GWAS is the identification of specific coding variants associated with phenotype. This is advantageous as (1) variants can implicate particular regions in genes/proteins, helping to elucidate mechanism and (2) variants can be examined practically in the lab. For instance, the FAN1 variants we identify in this study could be investigated *in vitro* by engineering artificial DNA structures (5' flaps, interstrand crosslink containing DNA, etc.), as per the (MacKay et al., 2010) and (Pizzolato et al., 2015) studies. How efficiently the DNA substrates are processed might indicate how damaging the variant in question is. FAN1 variants could also be investigated in a cellular system (as done by (Goold et al., 2019)) by examining variant effects on CAG instability and vulnerability to crosslinking agents (*e.g.* mitomycin C or cisplatin). Similar DNA repair assays with other repair components identified in this study (*e.g.* the rare, late associated variants in *PMS1*) could also be explored using a similar strategy. It would be very useful to consider how MSH3:EXO1 interactions may be involved in repeat expansion in a model system, for instance.

Investigating how the glutamine-alanine (Q-A) variants in TCERG1 affect HD would also be of great interest. Given we are currently unsure how TCERG1 is acting (transcription, splicing or both), a chromatin immunoprecipitation (ChIP) experiment may elucidate the genes which TCERG1 helps regulate. Investigation of whether *HTT* incomplete splicing is modulated by the Q-A variants in TCERG1 could be achieved using a similar system as reported by (Neueder et al., 2018), where mouse *HTT* minigenes were expressed in a cellular model to examine incomplete splicing of *HTT* using quantitative PCR (qPCR). A high-throughput RNA sequencing approach (bulk or single-cell) could also be considered for an unbiased investigation of differential splicing in HD.

6.5 Final conclusions

The research presented in this thesis has described the successful identification of several pathologically relevant modifiers of Huntington's disease onset using whole-exome sequencing. Further to this, atypical *HTT* alleles, significant modifiers of HD onset, were confirmed and partially characterised using an independent next-generation sequencing technique. These data have profound implications for understanding the pathogenic mechanisms underlying HD, and reinforce the paradigm that *HTT* DNA itself is a prominent driver of disease. Characterisation of the disease relevant variants found in this study in future work will further refine our understanding of molecular mechanisms in HD, and this has direct pharmacological pertinence. DNA repair genes such as *MSH3*, *PMS1* or the postulated protein-protein interaction variants in *FAN1* could represent tractable drug targets. Some of the identified modifiers may additionally have relevance in other repeat diseases, although additional work is needed to explore this. As sequencing becomes ever more common place both in research and the clinic, a similar whole-exome or whole-genome sequencing strategy is likely to find additional modifiers of HD, and thereby inform both disease mechanism and therapeutic targets.

Appendices

Appendix 1 – Adjusted CCQ data for 5 and 10 year cut-offs

Below are the by-CAG mean summaries for CCQ-derived symptoms in Registry (see 3.3.1). (A) describes CCQ data without any sxrater adjustment; (B) describes symptoms using a <2 year cut-off for sxrater; (C) describes symptoms using a <5 year cut-off for sxrater; (D) describes symptoms using a <10 year cut-off for sxrater. The first number is the mean (in years); the second the N of HD patients.

A

CAG	MTR	COG	APT	DEP	POB	IRB	VAB	PSY
40	58.10 (425)	61.26 (233)	58.54 (220)	51.42 (309)	58.99 (135)	55.04 (267)	56.24 (118)	55.83 (47)
41	54.54 (717)	57.41 (386)	57.30 (387)	49.86 (494)	54.52 (233)	51.81 (444)	52.19 (203)	55.07 (86)
42	51.06 (942)	54.17 (539)	53.07 (483)	47.75 (652)	53.59 (313)	49.77 (568)	50.21 (281)	53.48 (108)
43	47.23 (834)	50.13 (494)	49.92 (453)	43.40 (618)	49.92 (326)	46.41 (543)	45.81 (288)	49.33 (94)
44	44.42 (659)	47.23 (408)	47.56 (355)	41.86 (488)	47.36 (268)	43.59 (447)	44.06 (218)	47.24 (92)
45	40.87 (532)	44.01 (294)	43.26 (279)	39.63 (384)	44.16 (197)	41.03 (311)	41.94 (167)	44.81 (64)
46	38.54 (399)	41.47 (235)	41.19 (204)	37.72 (261)	42.72 (151)	39.49 (253)	39.82 (131)	40.33 (39)
47	36.34 (336)	39.23 (215)	38.97 (180)	34.60 (227)	39.87 (129)	36.61 (193)	36.83 (108)	38.60 (45)
48	34.27 (205)	37.20 (123)	36.58 (108)	33.33 (126)	37.30 (73)	34.58 (136)	35.51 (74)	36.53 (17)
49	32.98 (166)	35.35 (103)	34.87 (83)	32.50 (102)	35.32 (63)	34.06 (89)	33.44 (48)	38.05 (22)
50	30.94 (132)	35.01 (74)	33.93 (80)	30.24 (98)	35.33 (51)	31.20 (75)	30.94 (48)	32.08 (12)
51	29.60 (87)	32.00 (50)	31.10 (41)	29.50 (50)	31.16 (25)	28.98 (56)	30.18 (34)	34.33 (6)
52	27.39 (64)	29.88 (41)	30.08 (36)	27.50 (38)	30.50 (26)	28.17 (36)	28.14 (29)	32.86 (7)
53	26.39 (46)	28.55 (31)	27.86 (29)	25.81 (27)	32.13 (16)	26.16 (31)	25.62 (21)	28.70 (10)
54	26.53 (32)	25.44 (16)	25.67 (18)	28.06 (16)	28.17 (12)	26.61 (18)	24.13 (8)	30.33 (3)
55	25.10 (39)	29.31 (26)	28.67 (18)	25.00 (23)	28.59 (17)	24.71 (17)	25.00 (11)	23.75 (4)
56	24.14 (21)	25.76 (17)	25.73 (11)	25.17 (12)	28.55 (11)	22.45 (11)	27.00 (7)	29.75 (4)
57	23.77 (13)	25.27 (11)	27.89 (9)	25.86 (7)	33.67 (3)	20.00 (5)	22.20 (5)	19.00 (1)
58	23.38 (13)	20.10 (10)	25.67 (9)	28.50 (6)	29.20 (5)	22.00 (7)	27.20 (5)	21.00 (1)
59	22.50 (16)	22.67 (12)	24.50 (8)	20.25 (8)	25.00 (5)	21.50 (6)	23.00 (4)	23.00 (2)
60	20.70 (10)	19.67 (6)	25.20 (5)	19.75 (4)	28.00 (4)	24.50 (6)	23.75 (4)	16.00 (2)

B

CAG	MTR	COG	APT	DEP	POB	IRB	VAB	PSY
40	58.79 (362)	62.89 (206)	61.69 (175)	57.76 (192)	62.20 (109)	59.80 (184)	60.36 (83)	61.03 (39)
41	54.94 (619)	57.70 (346)	58.66 (326)	55.02 (327)	57.17 (195)	55.56 (325)	56.07 (155)	57.35 (72)
42	51.61 (815)	54.65 (487)	54.41 (405)	51.87 (448)	55.4 (269)	52.27 (434)	53.06 (217)	54.72 (94)
43	47.69 (735)	50.98 (433)	51.84 (376)	47.75 (420)	52.02 (274)	49.56 (418)	49.72 (208)	51.22 (78)
44	44.49 (594)	47.68 (368)	48.68 (311)	45.19 (343)	49.35 (232)	45.92 (362)	46.64 (177)	48.62 (81)
45	41.14 (472)	44.81 (268)	44.63 (240)	42.79 (280)	45.91 (172)	43.24 (254)	44.69 (131)	44.37 (54)
46	38.82 (352)	41.8 (210)	41.98 (183)	40.10 (194)	44.12 (129)	40.87 (215)	41.60 (113)	41.78 (32)
47	36.71 (299)	39.74 (196)	39.69 (158)	37.11 (166)	40.97 (116)	37.67 (159)	38.46 (90)	41.08 (38)
48	34.42 (182)	37.14 (110)	36.98 (91)	34.98 (92)	37.83 (65)	35.77 (108)	37.56 (57)	36.53 (17)
49	33.35 (155)	35.67 (99)	36.01 (74)	35.18 (79)	36.44 (59)	35.04 (81)	34.91 (44)	38.54 (22)
50	30.85 (120)	34.91 (69)	34.22 (69)	32.24 (74)	35.64 (47)	32.14 (64)	32.15 (40)	34.00 (11)
51	29.38 (81)	32.27 (44)	31.64 (36)	30.6 (40)	32.35 (20)	29.73 (44)	30.50 (26)	36.33 (3)
52	27.52 (61)	29.88 (40)	30.11 (35)	29.82 (28)	31.46 (24)	28.97 (33)	29.69 (26)	32.86 (7)
53	26.33 (43)	29.19 (27)	29.65 (23)	27.55 (20)	32.77 (13)	29.04 (24)	29.47 (15)	31.63 (8)
54	26.32 (31)	24.67 (15)	26.76 (17)	29.92 (13)	28.17 (12)	27.06 (16)	24.50 (6)	27.00 (2)
55	25.31 (36)	29.24 (25)	29.24 (17)	26.05 (20)	28.59 (17)	26.07 (14)	25.90 (10)	23.75 (4)
56	23.83 (18)	25.47 (15)	25.22 (9)	24.8 (10)	28.00 (10)	23.20 (10)	25.83 (6)	29.75 (4)
57	25.08 (12)	25.27 (11)	27.89 (9)	28.17 (6)	33.67 (3)	20.00 (5)	22.20 (5)	19.00 (1)
58	24.18 (11)	23.63 (8)	26.50 (8)	28.50 (6)	29.20 (5)	24.33 (6)	27.20 (5)	N/A (0)
59	22.13 (15)	22.67 (12)	24.50 (8)	21.14 (7)	22.50 (4)	21.50 (6)	23.00 (4)	23.00 (2)
60	20.56 (9)	22.40 (5)	25.20 (5)	21.67 (3)	29.33 (3)	24.60 (5)	23.67 (3)	N/A (0)

C

CAG	MTR	COG	APT	DEP	POB	IRB	VAB	PSY
40	58.80 (369)	62.73 (210)	61.55 (181)	57.85 (198)	62.20 (109)	59.75 (188)	60.08 (86)	61.03 (39)
41	54.88 (634)	57.65 (353)	58.53 (335)	54.71 (348)	57.13 (198)	55.03 (342)	55.38 (162)	57.27 (73)
42	51.56 (843)	54.53 (495)	54.26 (415)	51.65 (476)	55.31 (271)	52.07 (448)	52.91 (221)	54.80 (95)
43	47.57 (752)	50.83 (440)	51.48 (388)	47.32 (448)	51.96 (275)	49.08 (438)	48.89 (225)	51.18 (79)
44	44.48 (608)	47.59 (375)	48.44 (321)	45.01 (363)	49.16 (237)	45.80 (375)	46.41 (184)	48.62 (81)
45	41.12 (480)	44.64 (272)	44.36 (249)	42.41 (298)	45.82 (173)	43.00 (263)	44.30 (138)	44.31 (55)
46	38.75 (362)	41.82 (213)	41.98 (184)	39.82 (207)	43.94 (134)	40.72 (221)	41.50 (115)	41.56 (34)
47	36.71 (307)	39.60 (199)	39.48 (164)	36.90 (177)	40.97 (116)	37.40 (167)	38.28 (95)	41.08 (38)
48	34.30 (185)	37.14 (110)	36.86 (93)	34.34 (99)	37.88 (68)	35.50 (114)	37.49 (59)	36.53 (17)
49	33.31 (158)	35.67 (99)	35.68 (77)	34.38 (86)	36.17 (60)	34.69 (85)	34.78 (45)	38.54 (22)
50	30.96 (124)	34.91 (69)	34.11 (70)	31.21 (84)	35.35 (48)	32.06 (68)	31.76 (42)	34.00 (11)
51	29.38 (81)	32.27 (44)	31.64 (36)	30.34 (41)	32.35 (20)	29.53 (47)	30.04 (28)	36.33 (3)
52	27.39 (62)	29.88 (40)	30.11 (35)	28.84 (31)	30.96 (25)	28.97 (33)	29.69 (26)	32.86 (7)
53	26.33 (43)	29.50 (28)	29.25 (24)	27.19 (21)	32.77 (13)	28.68 (25)	29.47 (15)	31.63 (8)
54	26.32 (31)	24.67 (15)	26.76 (17)	29.36 (14)	28.17 (12)	27.00 (17)	24.71 (7)	27.00 (2)
55	25.31 (36)	29.24 (25)	29.24 (17)	25.90 (21)	28.59 (17)	25.87 (15)	25.90 (10)	23.75 (4)
56	23.83 (18)	25.47 (15)	25.22 (9)	24.80 (10)	28.00 (10)	23.20 (10)	25.83 (6)	29.75 (4)
57	25.08 (12)	25.27 (11)	27.89 (9)	28.17 (6)	33.67 (3)	20.00 (5)	22.20 (5)	19.00 (1)
58	23.58 (12)	23.63 (8)	26.50 (8)	28.50 (6)	29.20 (5)	24.33 (6)	27.20 (5)	N/A (0)
59	22.13 (15)	22.67 (12)	24.50 (8)	20.25 (8)	22.50 (4)	21.50 (6)	23.00 (4)	23.00 (2)
60	20.56 (9)	22.40 (5)	25.20 (5)	21.67 (3)	29.33 (3)	24.60 (5)	23.67 (3)	18.00 (1)

D

CAG	MTR	COG	APT	DEP	POB	IRB	VAB	PSY
40	58.82 (373)	62.71 (212)	61.4 (184)	57.57 (205)	61.91 (111)	59.27 (197)	59.80 (89)	61.03 (39)
41	54.83 (641)	57.61 (355)	58.19 (344)	54.04 (371)	56.92 (200)	54.50 (355)	55.13 (165)	56.83 (75)
42	51.48 (853)	54.51 (497)	54.08 (419)	51.08 (505)	54.95 (276)	51.64 (469)	52.29 (233)	54.48 (96)
43	47.53 (755)	50.83 (441)	51.33 (392)	46.73 (478)	51.59 (282)	48.78 (451)	48.59 (230)	50.50 (82)
44	44.52 (614)	47.54 (377)	48.24 (326)	44.21 (395)	48.88 (241)	45.56 (383)	46.03 (189)	48.45 (82)
45	40.99 (488)	44.64 (273)	44.12 (254)	41.59 (324)	45.53 (177)	42.76 (268)	43.93 (143)	44.31 (55)
46	38.71 (364)	41.77 (214)	41.98 (184)	39.20 (221)	43.54 (137)	40.42 (226)	41.35 (117)	41.56 (34)
47	36.71 (307)	39.56 (201)	39.48 (165)	36.62 (186)	40.58 (118)	37.27 (172)	38.25 (97)	39.90 (40)
48	34.25 (186)	37.07 (111)	36.86 (93)	34.07 (104)	37.88 (68)	35.33 (117)	37.12 (60)	36.53 (17)
49	33.25 (159)	35.67 (99)	35.45 (79)	33.71 (91)	36.17 (60)	34.69 (85)	34.78 (45)	38.54 (22)
50	30.94 (125)	34.91 (69)	33.78 (72)	31.24 (85)	35.35 (48)	31.66 (70)	31.14 (44)	34.00 (11)
51	29.38 (81)	32.31 (45)	31.64 (36)	29.84 (43)	31.62 (21)	29.33 (49)	29.97 (29)	36.33 (3)
52	27.33 (63)	29.88 (40)	30.08 (36)	28.03 (35)	30.50 (26)	28.17 (36)	28.64 (28)	32.86 (7)
53	26.33 (43)	29.50 (28)	28.72 (25)	26.30 (23)	32.77 (13)	28.68 (25)	28.88 (16)	31.63 (8)
54	26.32 (31)	24.67 (15)	25.67 (18)	28.60 (15)	28.17 (12)	27.00 (17)	24.71 (7)	27.00 (2)
55	25.08 (37)	29.24 (25)	28.67 (18)	25.50 (22)	28.59 (17)	25.38 (16)	25.00 (11)	23.75 (4)
56	23.83 (18)	25.47 (15)	25.22 (9)	24.18 (11)	28.00 (10)	22.45 (11)	25.83 (6)	29.75 (4)
57	25.08 (12)	25.27 (11)	27.89 (9)	25.86 (7)	33.67 (3)	20.00 (5)	22.20 (5)	19.00 (1)
58	23.58 (12)	21.44 (9)	26.50 (8)	28.50 (6)	29.20 (5)	24.33 (6)	27.20 (5)	N/A (0)
59	22.13 (15)	22.67 (12)	24.50 (8)	20.25 (8)	22.50 (4)	21.50 (6)	23.00 (4)	23.00 (2)
60	20.56 (9)	22.40 (5)	25.20 (5)	21.67 (3)	29.33 (3)	24.60 (5)	23.67 (3)	18.00 (1)

Appendix 2 – Correlation matrices for CCQ onset data by sex

Below are the correlation matrices for males and females based on onset data derived from the CCQ for unadjusted data (A and B), 2 year adjusted data (C and D), 5 year adjusted data (E and F) and 10 year adjusted data (G and H) from 3.3.2. Correlation matrices for symptoms in males are presented first (A, C, E and G) and symptoms in females second (B, D, F and H). Total male N=4140; total female N=4753. MTR: Motor; COG: cognitive; APT: apathy; DEP: depression; POB: perseverative/obsessive behaviour; IRB: irritability; VAB: violent/aggressive behaviour; PSY: psychosis.

A	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.894	0.796	0.830	0.818	0.892	0.840	0.805
COG	0.894	1	0.824	0.842	0.823	0.904	0.879	0.853
DEP	0.796	0.824	1	0.815	0.791	0.870	0.815	0.838
IRB	0.830	0.842	0.815	1	0.939	0.858	0.829	0.793
VAB	0.818	0.823	0.791	0.939	1	0.840	0.833	0.786
APT	0.892	0.904	0.870	0.858	0.840	1	0.882	0.862
POB	0.840	0.879	0.815	0.829	0.833	0.882	1	0.853
PSY	0.805	0.853	0.838	0.793	0.786	0.862	0.853	1

B	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.903	0.761	0.853	0.847	0.883	0.821	0.807
COG	0.903	1	0.787	0.875	0.869	0.915	0.850	0.836
DEP	0.761	0.787	1	0.791	0.810	0.819	0.750	0.776
IRB	0.853	0.875	0.791	1	0.938	0.874	0.835	0.841
VAB	0.847	0.869	0.810	0.938	1	0.881	0.873	0.852
APT	0.883	0.915	0.819	0.874	0.881	1	0.872	0.845
POB	0.821	0.850	0.750	0.835	0.873	0.872	1	0.815
PSY	0.807	0.836	0.776	0.841	0.852	0.845	0.815	1

C	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.926	0.897	0.916	0.906	0.929	0.908	0.867
COG	0.926	1	0.912	0.911	0.909	0.931	0.926	0.901
DEP	0.897	0.912	1	0.926	0.914	0.929	0.909	0.893
IRB	0.916	0.911	0.926	1	0.973	0.918	0.925	0.869
VAB	0.906	0.909	0.914	0.973	1	0.914	0.939	0.877
APT	0.929	0.931	0.929	0.918	0.914	1	0.943	0.881
POB	0.908	0.926	0.909	0.925	0.939	0.943	1	0.915
PSY	0.867	0.901	0.893	0.869	0.877	0.881	0.915	1

D	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.929	0.904	0.909	0.893	0.923	0.905	0.846
COG	0.929	1	0.905	0.916	0.909	0.944	0.913	0.880
DEP	0.904	0.905	1	0.932	0.910	0.919	0.885	0.867
IRB	0.909	0.916	0.932	1	0.962	0.919	0.915	0.878
VAB	0.893	0.909	0.910	0.962	1	0.920	0.919	0.881
APT	0.923	0.944	0.919	0.919	0.920	1	0.936	0.899
POB	0.905	0.913	0.885	0.915	0.919	0.936	1	0.876
PSY	0.846	0.880	0.867	0.878	0.881	0.899	0.876	1

E	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.925	0.896	0.913	0.905	0.925	0.908	0.867
COG	0.925	1	0.909	0.907	0.906	0.929	0.926	0.902
DEP	0.896	0.909	1	0.919	0.909	0.925	0.906	0.893
IRB	0.913	0.907	0.919	1	0.969	0.917	0.922	0.873
VAB	0.905	0.906	0.909	0.969	1	0.914	0.940	0.883
APT	0.925	0.929	0.925	0.917	0.914	1	0.942	0.887
POB	0.908	0.926	0.906	0.922	0.940	0.942	1	0.914
PSY	0.867	0.902	0.893	0.873	0.883	0.887	0.914	1

F	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.928	0.906	0.908	0.892	0.921	0.901	0.844
COG	0.928	1	0.899	0.913	0.897	0.940	0.907	0.877
DEP	0.906	0.899	1	0.931	0.908	0.917	0.882	0.858
IRB	0.908	0.913	0.931	1	0.961	0.918	0.909	0.873
VAB	0.892	0.897	0.908	0.961	1	0.920	0.908	0.859
APT	0.921	0.940	0.917	0.918	0.920	1	0.932	0.895
POB	0.901	0.907	0.882	0.909	0.908	0.932	1	0.867
PSY	0.844	0.877	0.858	0.873	0.859	0.895	0.867	1

G	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.922	0.885	0.907	0.901	0.923	0.901	0.867
COG	0.922	1	0.898	0.902	0.900	0.924	0.921	0.897
DEP	0.885	0.898	1	0.911	0.900	0.916	0.899	0.874
IRB	0.907	0.902	0.911	1	0.964	0.909	0.916	0.856
VAB	0.901	0.900	0.900	0.964	1	0.908	0.931	0.866
APT	0.923	0.924	0.916	0.909	0.908	1	0.935	0.882
POB	0.901	0.921	0.899	0.916	0.931	0.935	1	0.910
PSY	0.867	0.897	0.874	0.856	0.866	0.882	0.910	1

H	MTR	COG	DEP	IRB	VAB	APT	POB	PSY
MTR	1	0.925	0.892	0.899	0.883	0.915	0.892	0.834
COG	0.925	1	0.886	0.906	0.892	0.936	0.899	0.873
DEP	0.892	0.886	1	0.909	0.886	0.906	0.868	0.847
IRB	0.899	0.906	0.909	1	0.957	0.913	0.905	0.866
VAB	0.883	0.892	0.886	0.957	1	0.915	0.904	0.853
APT	0.915	0.936	0.906	0.913	0.915	1	0.925	0.879
POB	0.892	0.899	0.868	0.905	0.904	0.925	1	0.859
PSY	0.834	0.873	0.847	0.866	0.853	0.879	0.859	1

Appendix 3 – Full binary symptom, HADS-SIS and covariate correlation matrix

Below, (A) is a correlation matrix for all the covariates used to construct GLMs using adjusted (ADJ) binary CCQ data for CAGs 36-99 in Registry data with a pairwise deletion method (see 3.4.4). The adjusted binary CCQ removed individuals with CCQ symptoms occurring > 2 years earlier than the clinician's estimate for onset (sxrater). Only individuals with sxrater were included in this analysis (N=6303). Number of individuals for each association used is available in the second table (B) and p values for the correlations in (C). Note *p* values shown as 0 are *p*<2E-16. Duration: disease duration; Alcohol: alcohol use in units per week; Tobacco: cigarettes per day; TFC: total functional capacity; TMS: total motor score; onset: onset defined by sxrater; TDS: total depression score; TAS: total anxiety score; TIS: total irritability score; MTR_ADJ: Motor; COG_ADJ: cognitive; APT_ADJ: apathy; DEP_ADJ: depression; POB_ADJ: perseverative/ obsessive behaviour; IRB_ADJ: irritability; VAB_ADJ: violent/aggressive behaviour; PSY_ADJ: psychosis.

A	Sex	CAG	Duration	Alcohol	Tobacco	Education	TFC	TMS	Age	TDS	TAS	TIS	MTR_ADJ	COG_ADJ	APT_ADJ	DEP_ADJ	POB_ADJ	IRB_ADJ	VAB_ADJ	PSY_ADJ
Sex	1	0.018	-0.016	-0.122	-0.001	0.005	-0.065	0.066	-0.005	-0.021	0.041	-0.007	0.005	0.007	-0.017	0.111	-0.008	-0.054	-0.076	-0.005
CAG	0.018	1	-0.056	-0.064	0.058	0.029	-0.157	0.193	-0.701	-0.039	-0.057	-0.006	0.019	0.057	0.005	-0.059	0.046	-0.005	0.057	0.001
Duration	-0.016	-0.056	1	-0.063	-0.106	-0.145	-0.449	0.381	-0.138	0.108	0.004	-0.006	0.011	0.217	0.153	0.158	0.166	0.115	0.176	0.170
Alcohol	-0.122	-0.064	-0.063	1	0.135	0.052	0.135	-0.138	0.035	-0.020	0.009	0.027	-0.013	-0.001	0.015	-0.028	0.016	0.048	0.022	-0.012
Tobacco	-0.001	0.058	-0.106	0.135	1	0.057	0.112	-0.110	-0.142	0.074	0.072	0.097	0.003	-0.023	0.026	0.020	-0.016	0.066	0.046	-0.018
Education	0.005	0.029	-0.145	0.052	0.057	1	0.175	-0.135	-0.076	-0.019	0.018	0.025	-0.021	-0.037	-0.061	-0.075	-0.068	-0.027	-0.034	-0.041
TFC	-0.065	-0.157	-0.449	0.135	0.112	0.175	1	-0.722	0.035	-0.280	-0.064	-0.015	-0.098	-0.402	-0.246	-0.159	-0.236	-0.118	-0.223	-0.216
TMS	0.066	0.193	0.381	-0.138	-0.110	-0.135	-0.722	1	-0.062	0.178	-0.020	-0.056	0.132	0.292	0.155	0.092	0.174	0.058	0.134	0.119
Age	-0.005	-0.701	-0.138	0.035	-0.142	-0.076	0.035	-0.062	1	0.021	-0.017	-0.105	0.074	-0.066	-0.038	-0.057	-0.063	-0.072	-0.117	-0.047
TDS	-0.021	-0.039	0.108	-0.020	0.074	-0.019	-0.280	0.178	0.021	1	0.551	0.463	-0.009	0.149	0.252	0.213	0.075	0.118	0.132	0.044
TAS	0.041	-0.057	0.004	0.009	0.072	0.018	-0.064	-0.020	-0.017	0.551	1	0.681	-0.017	0.043	0.173	0.220	0.077	0.168	0.148	0.053
TIS	-0.007	-0.006	-0.006	0.027	0.097	0.025	-0.015	-0.056	-0.105	0.463	0.681	1	-0.011	0.047	0.130	0.165	0.086	0.275	0.244	0.038
MTR_ADJ	0.005	0.019	0.011	-0.013	0.003	-0.021	-0.098	0.132	0.074	-0.009	-0.017	-0.011	1	0.070	0.010	-0.021	0.022	0.023	0.006	-0.026
COG_ADJ	0.007	0.057	0.217	-0.001	-0.023	-0.037	-0.402	0.292	-0.066	0.149	0.043	0.047	0.070	1	0.280	0.161	0.204	0.205	0.215	0.146
APT_ADJ	-0.017	0.005	0.153	0.015	0.026	-0.061	-0.246	0.155	-0.038	0.252	0.173	0.130	0.010	0.280	1	0.279	0.253	0.269	0.218	0.119
DEP_ADJ	0.111	-0.059	0.158	-0.028	0.020	-0.075	-0.159	0.092	-0.057	0.213	0.220	0.165	-0.021	0.161	0.279	1	0.133	0.211	0.178	0.119
POB_ADJ	-0.008	0.046	0.166	0.016	-0.016	-0.068	-0.236	0.174	-0.063	0.075	0.077	0.086	0.022	0.204	0.253	0.133	1	0.284	0.277	0.186
IRB_ADJ	-0.054	-0.005	0.115	0.048	0.066	-0.027	-0.118	0.058	-0.072	0.118	0.168	0.275	0.023	0.205	0.269	0.211	0.284	1	0.476	0.142
VAB_ADJ	-0.076	0.057	0.176	0.022	0.046	-0.034	-0.223	0.134	-0.117	0.132	0.148	0.244	0.006	0.215	0.218	0.178	0.277	0.476	1	0.244
PSY_ADJ	-0.005	0.001	0.170	-0.012	-0.018	-0.041	-0.216	0.119	-0.047	0.044	0.053	0.038	-0.026	0.146	0.119	0.119	0.186	0.142	0.244	1

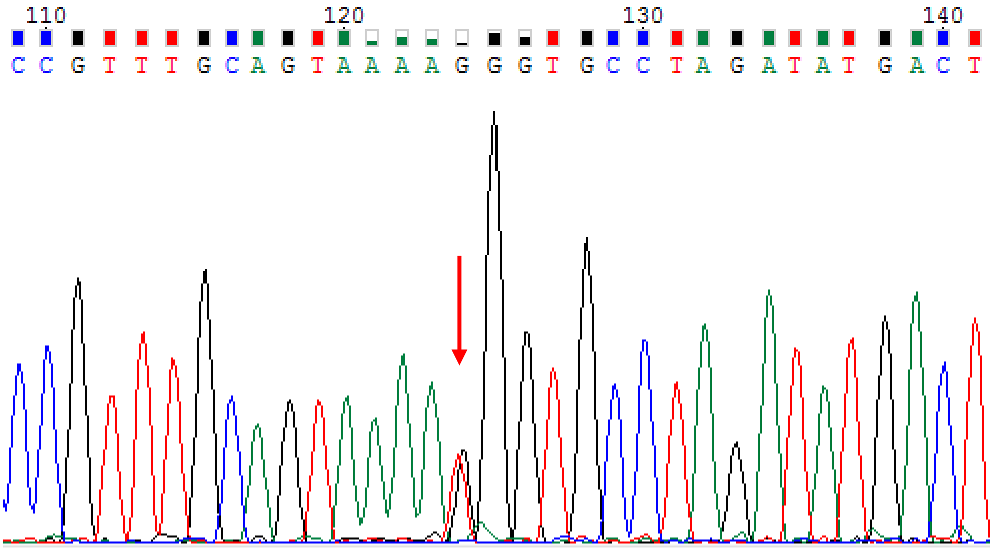
B	Sex	CAG	Duration	Alcohol	Tobacco	Education	TFC	TMS	Age	TDS	TAS	TIS	MTR_ADJ	COG_ADJ	APT_ADJ	DEP_ADJ	POB_ADJ	IRB_ADJ	VAB_ADJ	PSY_ADJ
Sex	6303	6303	5272	4470	4481	4577	4563	4549	6303	2374	2362	2366	5323	5407	5354	4875	5407	5162	5358	5530
CAG	6303	6303	5272	4470	4481	4577	4563	4549	6303	2374	2362	2366	5323	5407	5354	4875	5407	5162	5358	5530
Duration	5272	5272	5272	4468	4479	4575	4561	4547	5272	2373	2361	2365	4407	4479	4438	4042	4491	4288	4446	4577
Alcohol	4470	4470	4468	4470	4436	4470	4448	4434	4470	2300	2289	2294	4238	4306	4268	3882	4313	4121	4273	4396
Tobacco	4481	4481	4479	4436	4481	4481	4459	4445	4481	2317	2303	2309	4247	4318	4278	3888	4325	4131	4285	4409
Education	4577	4577	4575	4470	4481	4577	4549	4535	4577	2364	2351	2355	4337	4406	4365	3975	4415	4215	4370	4500
TFC	4563	4563	4561	4448	4459	4549	4563	4540	4563	2369	2357	2361	4323	4395	4353	3963	4404	4204	4361	4488
TMS	4549	4549	4547	4434	4445	4535	4540	4549	4549	2369	2357	2361	4309	4381	4340	3951	4391	4191	4348	4475
Age	6303	6303	5272	4470	4481	4577	4563	4549	6303	2374	2362	2366	5323	5407	5354	4875	5407	5162	5358	5530
TDS	2374	2374	2373	2300	2317	2364	2369	2369	2374	2374	2337	2342	2246	2300	2271	2077	2292	2190	2267	2339
TAS	2362	2362	2361	2289	2303	2351	2357	2357	2362	2337	2362	2334	2236	2289	2258	2067	2280	2177	2255	2327
TIS	2366	2366	2365	2294	2309	2355	2361	2361	2366	2342	2334	2366	2239	2293	2263	2072	2285	2184	2260	2333
MTR_ADJ	5323	5323	4407	4238	4247	4337	4323	4309	5323	2246	2236	2239	5323	5185	5133	4701	5169	4965	5128	5276
COG_ADJ	5407	5407	4479	4306	4318	4406	4395	4381	5407	2300	2289	2293	5185	5407	5233	4759	5254	5037	5213	5367
APT_ADJ	5354	5354	4438	4268	4278	4365	4353	4340	5354	2271	2258	2263	5133	5233	5354	4774	5225	5031	5184	5321
DEP_ADJ	4875	4875	4042	3882	3888	3975	3963	3951	4875	2077	2067	2072	4701	4759	4774	4875	4768	4627	4734	4851
POB_ADJ	5407	5407	4491	4313	4325	4415	4404	4391	5407	2292	2280	2285	5169	5254	5225	4768	5407	5052	5218	5370
IRB_ADJ	5162	5162	4288	4121	4131	4215	4204	4191	5162	2190	2177	2184	4965	5037	5031	4627	5052	5162	5110	5133
VAB_ADJ	5358	5358	4446	4273	4285	4370	4361	4348	5358	2267	2255	2260	5128	5213	5184	4734	5218	5110	5358	5324
PSY_ADJ	5530	5530	4577	4396	4409	4500	4488	4475	5530	2339	2327	2333	5276	5367	5321	4851	5370	5133	5324	5530

C	Sex	CAG	Duration	Alcohol	Tobacco	Education	TFC	TMS	Age	TDS	TAS	TIS	MTR_ADJ	COG_ADJ	APT_ADJ	DEP_ADJ	POB_ADJ	IRB_ADJ	VAB_ADJ	PSY_ADJ
Sex	NA	0.162506	0.254247	2.22E-16	0.929209	0.732307	1.09E-05	8.58E-06	0.678815	0.30893	0.047226	0.743098	0.731666	0.61531	0.217017	8.44E-15	0.543973	0.000105	3.12E-08	0.731052
CAG	0.162506	NA	4.43E-05	1.80E-05	0.000111	0.046809	0	0	0	0.057336	0.005287	0.777595	0.175095	2.59E-05	0.695461	3.62E-05	0.000738	0.704433	3.32E-05	0.943221
Duration	0.254247	4.43E-05	NA	2.35E-05	1.13E-12	0	0	0	0	1.22E-07	0.832463	0.776341	0.48501	0	0	0	0	4.49E-14	0	0
Alcohol	2.22E-16	1.80E-05	2.35E-05	NA	0	0.000534	0	0	0.018499	0.331453	0.652786	0.194791	0.39729	0.966221	0.313935	0.086445	0.294249	0.002036	0.157798	0.418501
Tobacco	0.929209	0.000111	1.13E-12	0	NA	0.00014	7.24E-14	2.21E-13	0	0.000398	0.000572	2.96E-06	0.848009	0.130985	0.088514	0.207665	0.286626	2.34E-05	0.002448	0.225387
Education	0.732307	0.046809	0	0.000534	0.00014	NA	0	0	2.78E-07	0.358544	0.3924	0.217194	0.165007	0.013622	6.23E-05	1.93E-06	6.45E-06	0.075977	0.024444	0.005546
TFC	1.09E-05	0	0	0	7.24E-14	0	NA	0	0.019691	0	0.001957	0.477871	8.79E-11	0	0	0	0	1.69E-14	0	0
TMS	8.58E-06	0	0	0	2.21E-13	0	0	NA	3.11E-05	0	0.329079	0.006436	0	0	0	7.89E-09	0	0.000191	0	1.33E-15
Age	0.678815	0	0	0.018499	0	2.78E-07	0.019691	3.11E-05	NA	0.316722	0.405602	3.09E-07	5.44E-08	1.04E-06	0.004983	6.43E-05	3.75E-06	2.25E-07	0	0.00042
TDS	0.30893	0.057336	1.22E-07	0.331453	0.000398	0.358544	0	0	0.316722	NA	0	0	0.685986	6.63E-13	0	0	0.000352	3.15E-08	2.46E-10	0.03159
TAS	0.047226	0.005287	0.832463	0.652786	0.000572	0.3924	0.001957	0.329079	0.405602	0	NA	0	0.410455	0.039907	0	0	0.000251	3.11E-15	1.74E-12	0.010621
TIS	0.743098	0.777595	0.776341	0.194791	2.96E-06	0.217194	0.477871	0.006436	3.09E-07	0	0	NA	0.603961	0.023283	5.68E-10	4.57E-14	3.73E-05	0	0	0.064397
MTR_ADJ	0.731666	0.175095	0.48501	0.39729	0.848009	0.165007	8.79E-11	0	5.44E-08	0.685986	0.410455	0.603961	NA	4.70E-07	0.489164	0.146689	0.111973	0.112399	0.681316	0.061325
COG_ADJ	0.61531	2.59E-05	0	0.966221	0.130985	0.013622	0	0	1.04E-06	6.63E-13	0.039907	0.023283	4.70E-07	NA	0	0	0	0	0	0
APT_ADJ	0.217017	0.695461	0	0.313935	0.088514	6.23E-05	0	0	0.004983	0	0	5.68E-10	0.489164	0	NA	0	0	0	0	0
DEP_ADJ	8.44E-15	3.62E-05	0	0.086445	0.207665	1.93E-06	0	7.89E-09	6.43E-05	0	0	4.57E-14	0.146689	0	0	NA	0	0	0	0
POB_ADJ	0.543973	0.000738	0	0.294249	0.286626	6.45E-06	0	0	3.75E-06	0.000352	0.000251	3.73E-05	0.111973	0	0	0	NA	0	0	0
IRB_ADJ	0.000105	0.704433	4.49E-14	0.002036	2.34E-05	0.075977	1.69E-14	0.000191	2.25E-07	3.15E-08	3.11E-15	0	0.112399	0	0	0	0	NA	0	0
VAB_ADJ	3.12E-08	3.32E-05	0	0.157798	0.002448	0.024444	0	0	0	2.46E-10	1.74E-12	0	0.681316	0	0	0	0	0	NA	0
PSY_ADJ	0.731052	0.943221	0	0.418501	0.225387	0.005546	0	1.33E-15	0.00042	0.03159	0.010621	0.064397	0.061325	0	0	0	0	0	0	NA

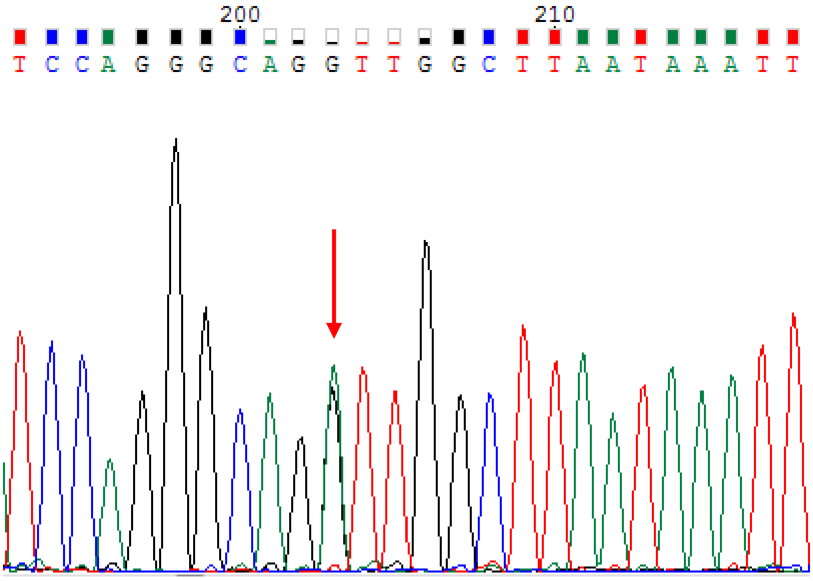
Appendix 4 – *FAN1* Sanger sequencing traces

Shown below are Sanger sequencing traces for every distinct *FAN1* variant sequenced successfully (17 variants), including 14 non-synonymous damaging (NSD) variants. Note the Gln717Arg variant was confirmed to be absent and is not included (as it is a negative trace). Traces visualised using Chromas (Technelysium). See 4.6.2 for the results of the Sanger sequencing in text, which are summarised in Table 4.4.

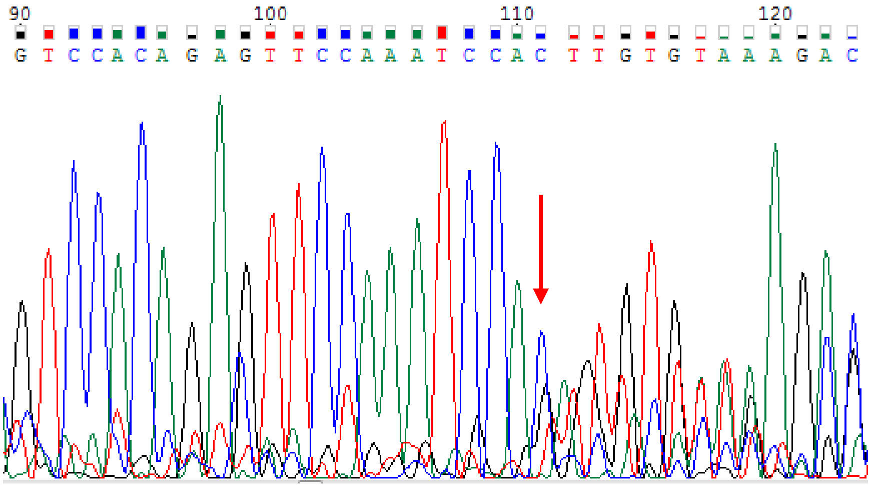
M50R: ATG (M) -> AGG (R)



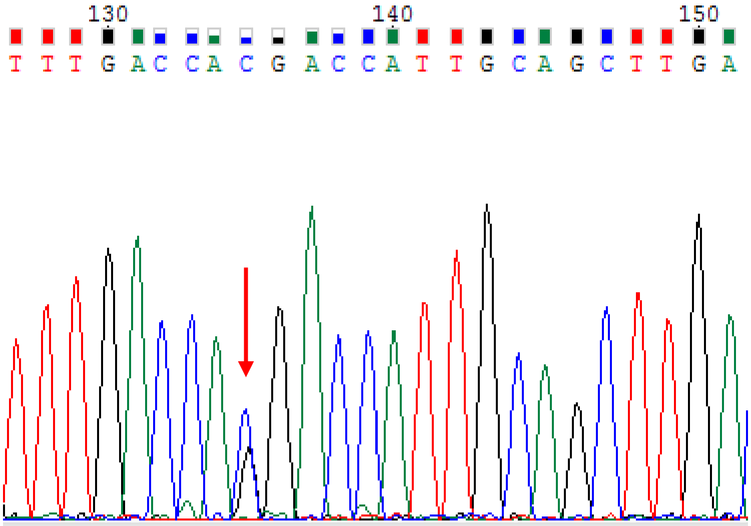
V77I: GTT (V) -> ATT (I)



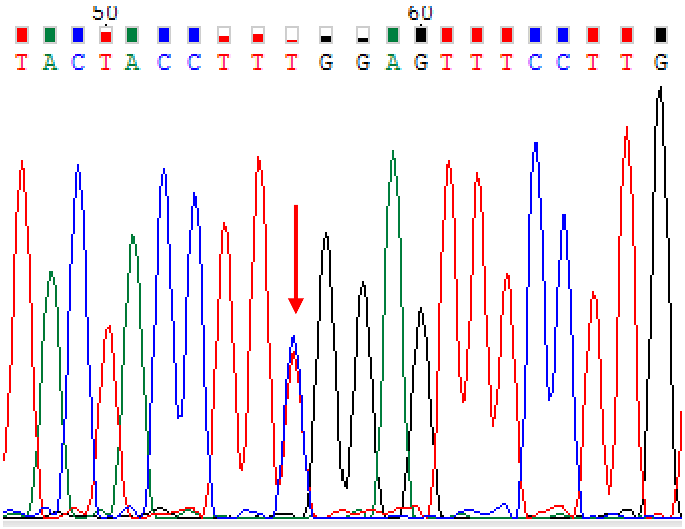
T187fs: CCA (T) -> Cfs



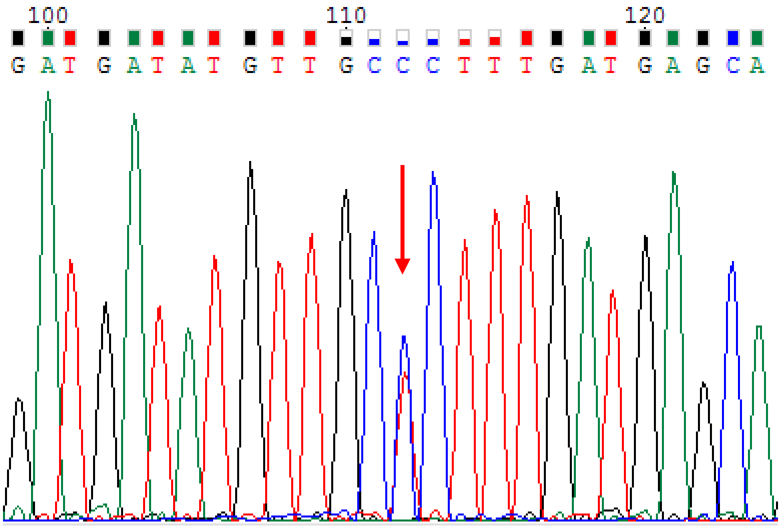
P366R: CCT (P) -> CGT (R) (reverse)



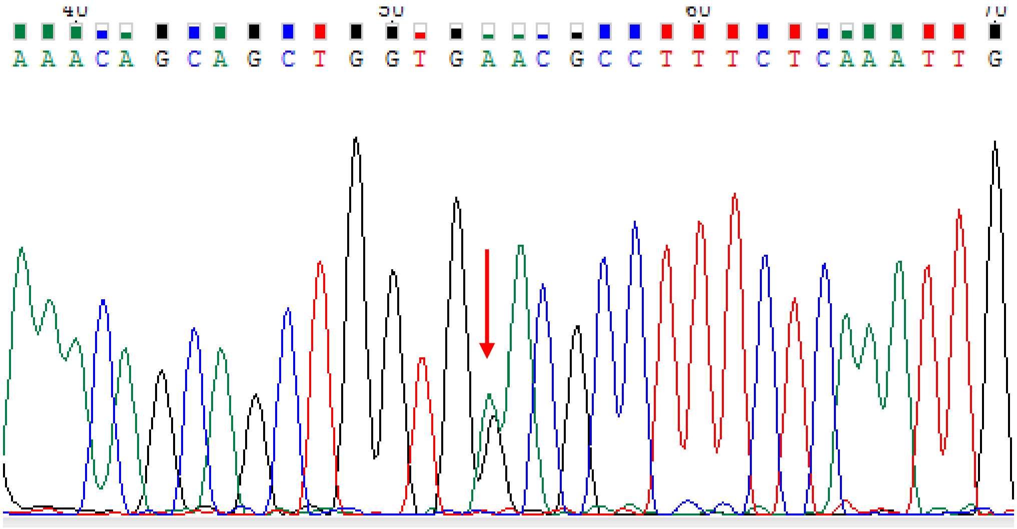
R377W: CGG (R) -> TGG (W)



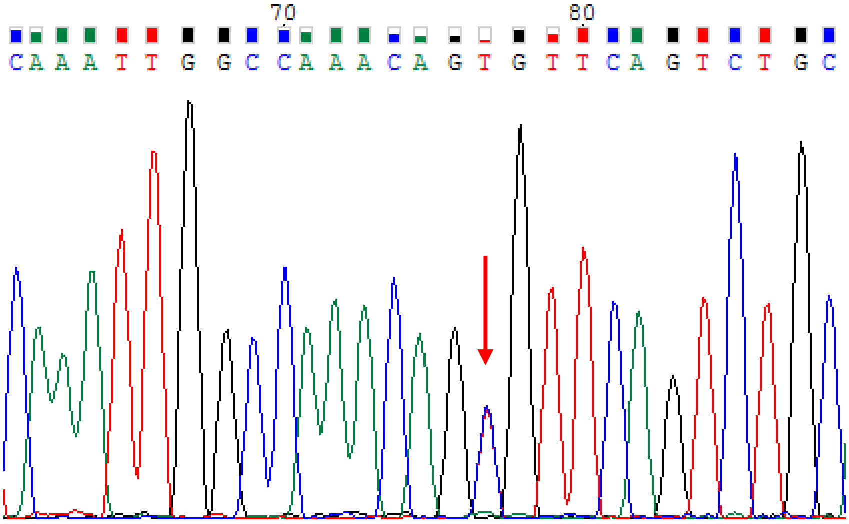
L395P: CTC (L) -> CCC (P)



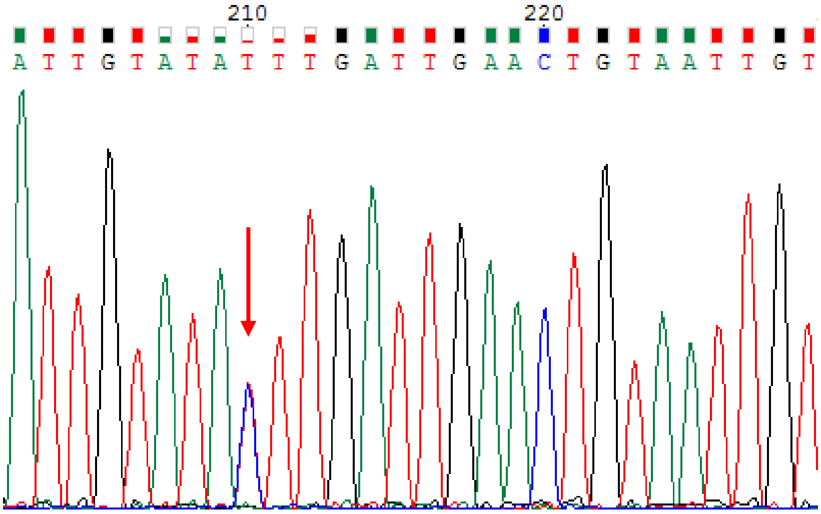
D498N: GAC (D) -> AAC (N)



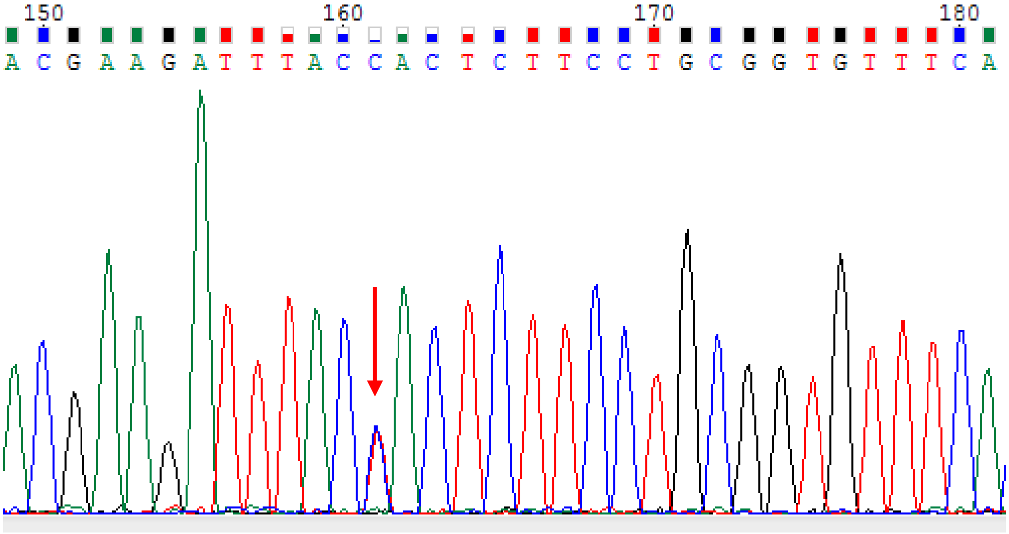
R507C: CGT (R) -> CAT (H)



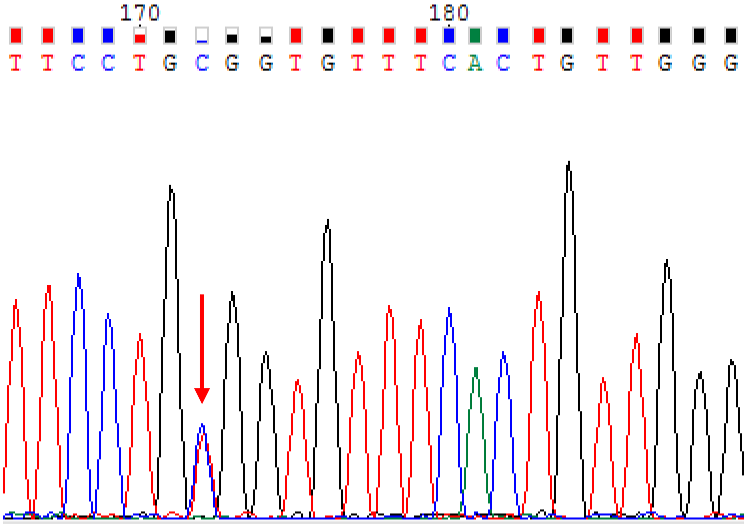
R507H: CGT (R) -> CAT (H)



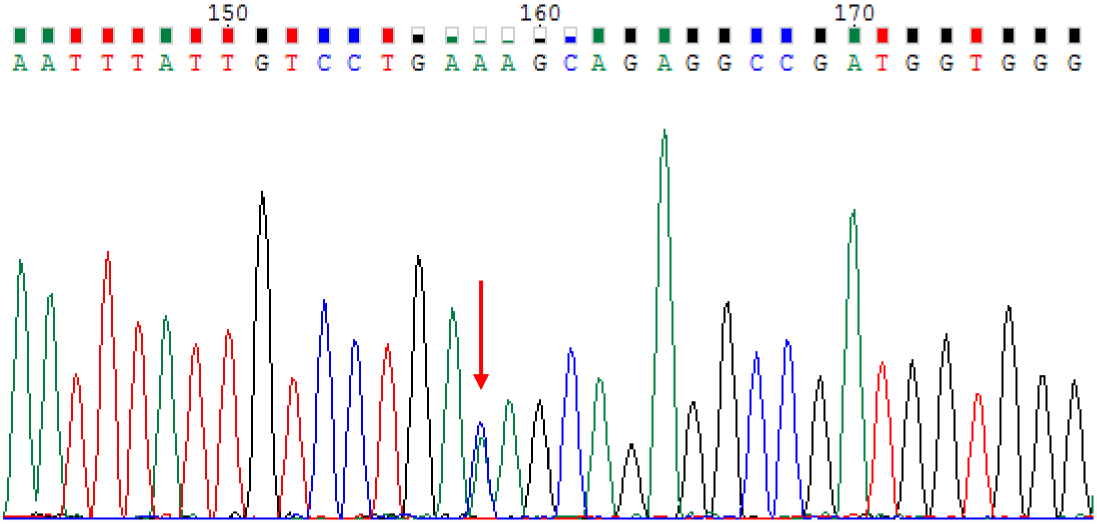
P654L: CCA (P) -> CTA (L)



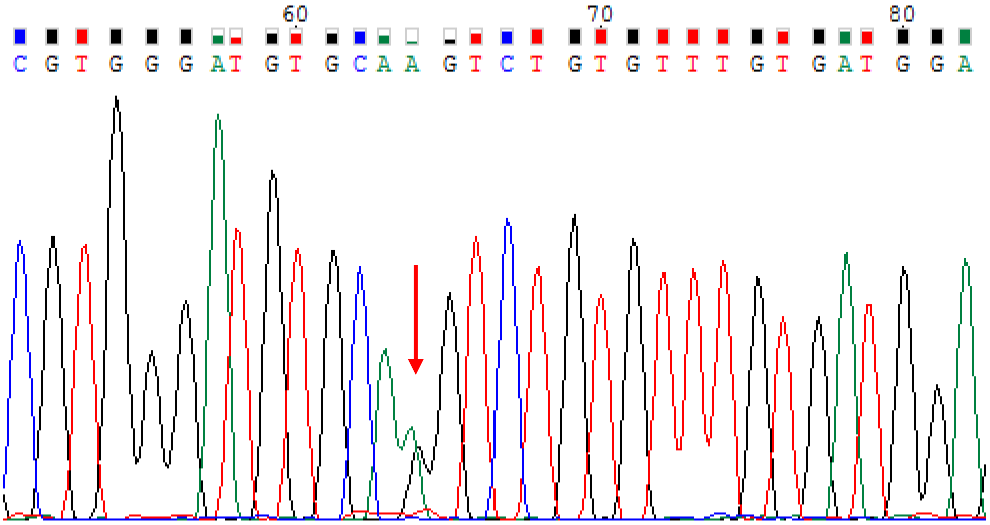
R658W: CGG (R) -> TGG (W)



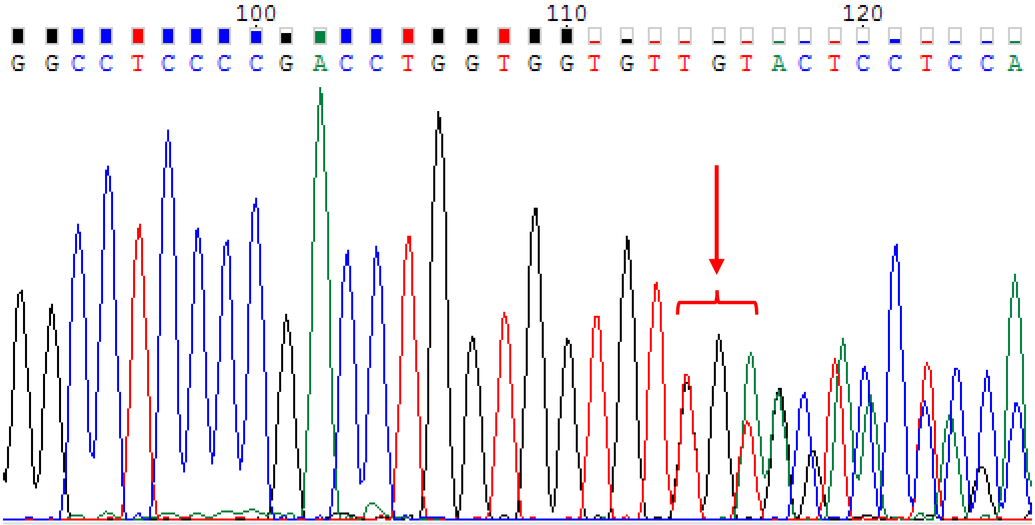
D702E: GAC (D) -> GAA (E)



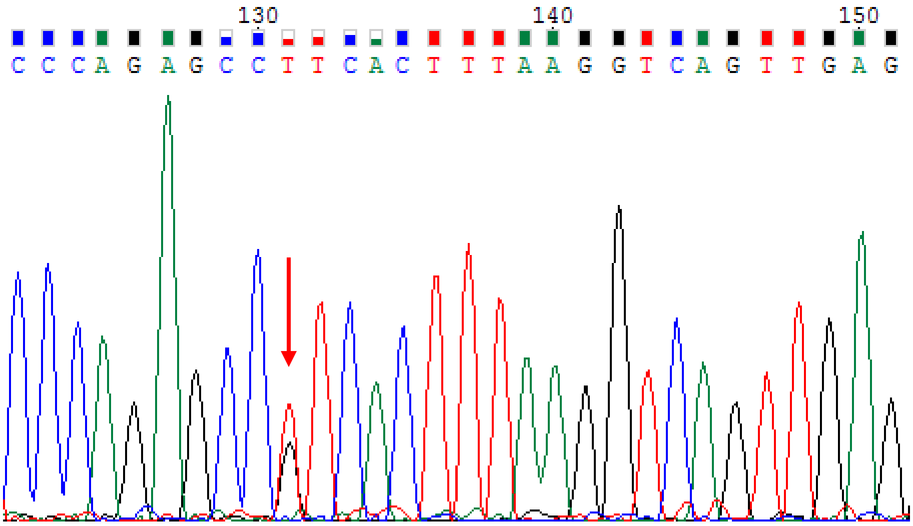
K794R: AAG (K) -> AGG (R)



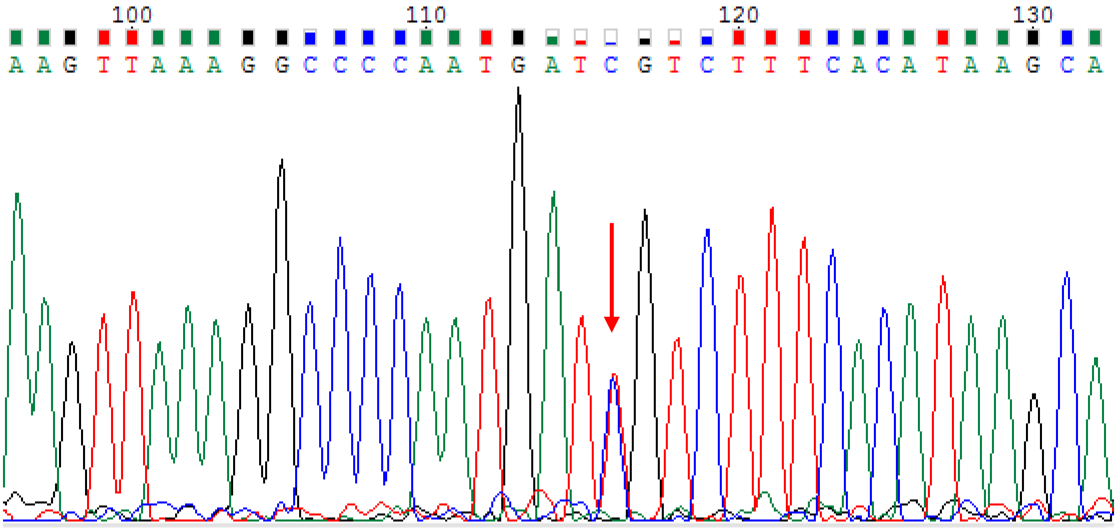
V963_W964insL: TGT insertion



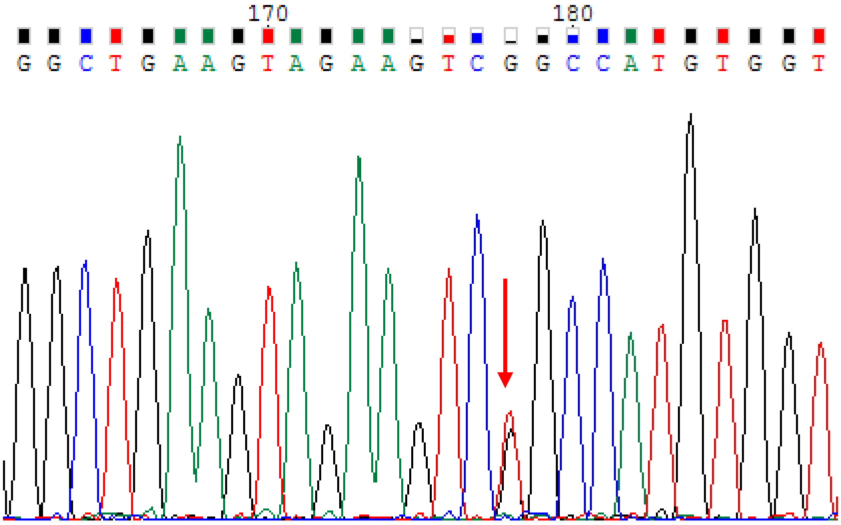
R969L: CGT (R) -> CTT (L)



R982C: CGT (R) -> TGT (C)



C1004G: TGC (C) -> GGC (G)



Appendix 5 – Non-coding variation identified by WES in other candidate genes

Listed below are non-coding variants for the genes described in 4.6. This includes synonymous mutations and variants in intronic/splice site regions. Genomic locations are based on hg19/GRCh37. MAF annotations taken from v2.0.2 of gnomAD. Total N=440 (225 early; 215 late). DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Arg58Arg	15:31197040:G:A	13.61	8.87E-04	134	89	2	0	143	72	0	0
Asp201Asp	15:31197469:C:T	23.46	1.53E-03	3	222	0	0	7	207	1	0
Asn214Asn	15:31197508:C:T	22.20	8.97E-06	16	209	0	0	23	191	1	0
Ala261Ala	15:31197649:G:A	35.62	4.39E-04	0	225	0	0	0	213	2	0
Thr370Thr	15:31197976:C:T	36.30	1.42E-02	0	209	16	0	0	206	9	0
Glu398Glu	15:31198060:G:A	15.37	1.02E-03	83	141	1	0	94	121	0	0
Intron	15:31198640:A:G	38.21	5.16E-04	0	224	1	0	0	215	0	0
Leu443Leu	15:31200415:A:G	21.29	1.53E-04	23	201	1	0	24	191	0	0
Ser617Ser	15:31210406:C:T	28.33	2.39E-02	4	212	9	0	0	204	10	1
Splice region	15:31212744:T:C	30.15	5.76E-03	0	225	0	0	3	208	4	0
His650His	15:31212754:C:T	30.53	0.00E+00	0	224	1	0	3	212	0	0
Arg706Arg	15:31214503:A:C	32.48	9.22E-04	0	224	1	0	0	214	1	0
Ala750Ala	15:31217407:C:T	28.54	0.00E+00	1	223	1	0	0	215	0	0
Pro757Pro	15:31217428:G:C	27.00	3.76E-04	2	222	1	0	0	215	0	0
Asp806Asp	15:31218072:C:T	37.04	4.52E-03	0	224	1	0	0	214	1	0
Thr905Thr	15:31221528:G:A	20.76	5.41E-05	9	215	1	0	9	206	0	0
His1005His	15:31229420:T:C	27.98	4.40E-01	11	68	105	41	9	88	88	30

FAN1

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Splice region	4:3107081:C:T	24.53	7.35E-04	1	222	2	0	3	212	0	0
Splice region	4:3107171:G:T	35.13	7.34E-04	5	220	0	0	1	213	1	0
Leu295Leu	4:3117168:C:G	15.17	3.29E-02	86	128	11	0	95	114	6	0
Thr396Thr	4:3123074:C:T	36.11	4.05E-02	0	217	8	0	0	199	16	0
Gly428Gly	4:3124626:T:C	15.26	4.49E-05	66	159	0	0	71	143	1	0
Ala539Ala	4:3129205:C:T	36.49	0.00E+00	0	225	0	0	1	213	1	0
Tyr861Tyr	4:3136217:T:C	36.83	1.79E-05	0	225	0	0	0	214	1	0
Thr969Thr	4:3142345:G:A	31.72	9.05E-04	2	222	1	0	4	211	0	0
Intron	4:3158974:G:A	24.95	4.36E-04	24	200	1	0	21	194	0	0
Leu1267Leu	4:3162056:C:T	44.92	2.98E-01	0	151	64	10	0	147	65	3
Thr1722Thr	4:3189554:G:A	19.99	7.16E-05	36	189	0	0	31	183	1	0
Val2016Val	4:3208683:A:G	14.73	9.04E-06	147	77	1	0	141	73	1	0
His2087His	4:3210608:C:T	27.10	1.62E-04	0	225	0	0	2	212	1	0
Glu2139Glu	4:3213658:G:A	16.73	2.42E-04	42	183	0	0	43	171	1	0
Glu2197Glu	4:3213832:G:A	13.49	3.02E-01	144	46	31	4	138	40	34	3
Leu2392Leu	4:3219613:A:C	33.05	1.00E+00	0	0	0	225	1	0	0	214
Leu2599Leu	4:3227419:A:G	37.34	3.03E-01	3	151	62	9	2	144	66	3
Ala2646Ala	4:3230431:C:G	12.57	1.08E-03	203	21	1	0	202	13	0	0
Leu2719Leu	4:3231661:G:A	32.55	6.92E-02	0	210	15	0	0	199	16	0
Thr2725Thr	4:3231679:A:C	34.43	2.68E-03	0	224	1	0	0	215	0	0
Leu2886Leu	4:3237376:C:T	39.71	2.16E-03	0	225	0	0	0	214	1	0
Asp2985Asp	4:3240237:C:T	28.69	1.88E-04	0	224	1	0	2	213	0	0
Gln3010Gln	4:3240312:G:A	18.69	5.14E-03	63	161	1	0	64	150	1	0
Thr3026Thr	4:3240568:C:T	16.86	9.13E-06	44	181	0	0	45	169	1	0
Phe3059Phe	4:3240667:C:T	22.41	6.83E-03	20	204	1	0	18	195	2	0

HTT (CAG tract variants are not shown)

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Intron	1:242015734:ATC:A	25.26	3.58E-05	3	221	1	0	4	211	0	0
Intron	1:242020804:A:G	27.00	8.25E-02	23	165	36	1	37	153	24	1
Intron	1:242023808:T:C	26.92	1.79E-05	2	223	0	0	2	213	0	0
Leu408Leu	1:242030312:T:C	36.80	0.00E+00	1	224	0	0	0	215	0	0
Intron	1:242030382:G:A	27.45	5.61E-05	2	222	1	0	1	214	0	0
Pro564Pro	1:242042228:G:A	39.20	2.69E-05	0	224	1	0	0	215	0	0
Pro622Pro	1:242042402:G:A	14.89	6.27E-05	115	109	1	0	121	94	0	0
Intron	1:242042690:A:G	15.53	8.07E-01	60	3	49	113	62	2	45	106
Ser725Ser	1:242045283:T:A	23.97	2.32E-02	13	201	11	0	22	186	7	0
Intron	1:242045336:A:AT	19.87	9.17E-04	120	105	0	0	125	90	0	0
Intron	1:242045336:A:T	19.78	9.17E-04	109	115	1	0	114	101	0	0
Intron	1:242045336:AT:A	19.70	9.17E-04	132	93	0	0	125	87	3	0
Intron	1:242048600:G:T	34.65	NA	0	224	1	0	0	215	0	0
Intron	1:242048896:C:T	33.33	5.53E-05	1	224	0	0	0	214	1	0
3'UTR	1:242052915:G:A	34.04	8.98E-06	1	223	1	0	0	215	0	0

EXO1

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Ser259Ser	5:145894900:A:G	39.83	8.96E-06	0	225	0	0	0	214	1	0
Thr97Thr	5:145895386:C:G	39.04	1.79E-05	0	224	1	0	0	215	0	0

GPR151

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
5'UTR	5:79950497:C:T	19.75	2.74E-01	177	23	22	3	150	31	25	9
5'UTR	5:79950508:C:T	20.36	1.77E-01	144	61	19	1	128	69	16	2
5'UTR	5:79950512:A:G	21.12	6.61E-01	150	5	37	33	130	12	31	42
Ala7Ala	5:79950567:G:A	21.01	7.49E-03	66	155	4	0	81	133	1	0
Thr37Thr	5:79950657:C:T	34.71	1.40E-04	2	223	0	0	0	214	1	0
Ala43Ala	5:79950675:A:G	28.45	9.88E-06	13	211	1	0	16	199	0	0
Intron	5:79952390:T:C	48.25	2.70E-01	0	113	98	14	0	100	89	26
Intron	5:79960929:A:G	22.51	3.61E-04	9	216	0	0	13	201	1	0
Splice region	5:79960955:G:A	27.92	2.76E-01	8	110	97	10	14	92	86	23
Pro231Pro	5:79966029:G:A	40.58	5.91E-02	0	199	26	0	0	185	30	0
Intron	5:79966178:G:A	25.32	5.49E-02	25	175	23	2	32	162	21	0
Intron	5:79968716:C:G	32.48	6.28E-05	11	214	0	0	10	204	1	0
Intron	5:79974930:T:G	35.82	2.70E-05	0	224	1	0	0	215	0	0
Intron	5:80021399:T:C	28.40	1.17E-03	1	223	1	0	1	214	0	0
Intron	5:80057353:CTT:C	41.65	2.39E-03	0	223	2	0	0	215	0	0
Ser598Ser	5:80057395:G:A	41.54	1.75E-03	0	225	0	0	0	214	1	0
Lys632Lys	5:80057497:A:G	41.31	9.87E-05	0	225	0	0	0	214	1	0
Intron	5:80063740:T:A	19.59	9.01E-06	13	212	0	0	16	198	1	0
Splice region	5:80063744:A:G	20.18	6.48E-04	20	205	0	0	18	194	3	0
Gln664Gln	5:80063847:G:A	36.26	1.77E-03	0	223	2	0	0	215	0	0
Intron	5:80071494:T:G	25.95	5.73E-03	8	213	4	0	9	206	0	0
Intron	5:80083371:G:T	41.79	1.54E-03	0	225	0	0	0	213	2	0
Intron	5:80109375:AT:A	25.82	NA	0	225	0	0	4	210	1	0
Intron	5:80160596:AAATG:A	13.29	5.52E-01	139	4	51	31	161	5	27	22
Intron	5:80160610:T:A	16.32	8.08E-02	47	148	29	1	61	129	23	2
Splice region	5:80160765:T:G	34.25	2.69E-05	0	225	0	0	0	214	1	0

MSH3

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Tyr47Tyr	2:190660503:T:C	32.40	6.27E-05	0	224	1	0	0	215	0	0
Gly58Gly	2:190660536:G:T	35.91	4.68E-03	1	223	1	0	0	214	1	0
Tyr90Tyr	2:190660632:C:T	35.49	1.16E-04	0	225	0	0	0	214	1	0
Splice region	2:190660683:G:A	34.30	3.74E-02	0	208	17	0	0	203	10	2
Intron	2:190660727:T:G	27.14	4.55E-05	10	215	0	0	14	200	1	0
Intron	2:190670332:C:T	24.20	9.05E-06	10	214	1	0	9	206	0	0
Asp115Asp	2:190670407:T:C	33.03	1.79E-04	1	223	1	0	1	214	0	0
Intron	2:190682952:T:TAAAC	18.15	2.33E-02	45	167	13	0	57	150	8	0
Intron	2:190717335:G:A	19.89	1.71E-04	21	204	0	0	24	190	1	0
Intron	2:190717358:C:A	28.38	5.74E-04	1	223	1	0	1	214	0	0
Intron	2:190717517:G:GA	16.74	1.33E-04	175	49	1	0	158	57	0	0
Ser490Ser	2:190719468:T:C	39.18	NA	0	225	0	0	0	214	1	0
Intron	2:190719895:A:G	23.93	3.77E-05	12	213	0	0	14	200	1	0
Intron	2:190722244:T:C	18.09	5.70E-04	46	179	0	0	35	179	1	0
Intron	2:190722245:A:G	18.04	1.28E-02	48	174	3	0	38	175	2	0
Intron	2:190729102:G:T	23.33	1.79E-05	6	218	1	0	15	200	0	0
Intron	2:190729105:C:T	23.98	7.31E-05	2	223	0	0	13	201	1	0
Intron	2:190729106:G:A	23.54	5.16E-02	1	204	19	1	15	176	24	0
Intron	2:190738196:A:G	22.29	2.74E-05	12	213	0	0	10	204	1	0
Intron	2:190738210:T:C	23.77	1.91E-03	7	218	0	0	5	208	2	0
Intron	2:190741968:G:A	12.84	8.97E-06	129	95	1	0	139	76	0	0

PMS1

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Leu822Leu	7:6013153:A:G	14.59	1.44E-01	209	7	9	0	204	5	6	0
Intron	7:6017189:T:C	17.05	7.94E-03	62	161	2	0	79	136	0	0
Intron	7:6018185:A:G	18.74	4.74E-03	44	179	2	0	52	161	2	0
Leu729Leu	7:6018315:G:C	28.45	1.12E-03	13	212	0	0	9	205	1	0
Asn683Asn	7:6022580:G:A	16.15	1.35E-04	138	85	2	0	160	55	0	0
Splice region	7:6022626:C:T	11.55	1.41E-01	209	14	2	0	200	11	4	0
Splice region	7:6022629:G:A	11.59	6.63E-02	215	8	2	0	208	5	2	0
Splice region	7:6026384:C:T	16.84	4.14E-02	50	165	10	0	50	159	6	0
Ser523Ser	7:6026827:G:C	31.03	3.70E-03	0	223	2	0	1	212	2	0
Tyr519Tyr	7:6026839:A:G	29.81	8.06E-05	1	224	0	0	2	212	1	0
Pro440Pro	7:6027076:T:C	34.95	1.52E-04	0	225	0	0	0	214	1	0
Intron	7:6027261:C:T	31.39	7.41E-05	0	225	0	0	1	213	1	0
Arg295Arg	7:6035183:C:T	38.01	NA	0	224	1	0	0	215	0	0
Ser260Ser	7:6036980:G:C	39.53	8.17E-01	3	8	69	145	0	3	78	134
Splice region	7:6037057:G:GA	17.31	4.27E-01	98	110	17	0	83	102	30	0
Splice region	7:6037057:G:GAA	17.76	4.27E-01	52	173	0	0	51	162	2	0
Splice region	7:6037057:GA:G	17.99	4.27E-01	96	20	109	0	88	31	95	1
Splice region	7:6037057:GAA:G	16.85	4.27E-01	113	97	15	0	89	109	17	0
Splice region	7:6037057:GAAA:G	17.82	4.27E-01	64	160	1	0	58	156	1	0
Intron	7:6038703:G:A	34.68	1.46E-02	0	221	4	0	0	210	5	0
Intron	7:6038714:C:T	35.97	0.00E+00	0	225	0	0	0	214	1	0
Intron	7:6038722:T:C	46.90	4.12E-01	1	83	109	32	3	63	111	38
Splice region	7:6042274:G:A	39.09	0.00E+00	0	224	1	0	0	215	0	0
Intron	7:6043295:G:A	21.71	9.35E-06	9	216	0	0	12	202	1	0
Ala96Ala	7:6043386:G:A	37.58	2.95E-02	0	214	11	0	0	204	10	1
Intron	7:6043443:A:C	33.81	1.62E-03	0	224	1	0	1	214	0	0
Intron	7:6048618:C:G	40.34	2.92E-03	0	223	2	0	0	214	1	0
5'-UTR	7:6048725:A:G	34.19	4.62E-05	2	222	1	0	0	215	0	0

PMS2

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
5'-UTR	3:37035011:A:G	44.71	1.24E-03	0	224	1	0	0	215	0	0
5'-UTR	3:37035032:C:T	44.78	1.13E-03	0	224	1	0	0	215	0	0
Intron	3:37038063:TTAGAG:T	35.58	1.79E-05	0	225	0	0	2	212	1	0
Intron	3:37038094:C:G	38.00	1.79E-05	0	225	0	0	0	214	1	0
Thr66Thr	3:37038191:C:T	39.59	5.11E-04	0	225	0	0	0	213	2	0
Intron	3:37045863:C:A	29.81	5.28E-03	1	223	1	0	0	213	2	0
Ala125Ala	3:37045960:A:G	35.88	5.38E-04	0	225	0	0	0	214	1	0
Intron	3:37048435:GTATC:G	28.34	7.19E-05	3	221	1	0	0	215	0	0
Intron	3:37048441:A:G	30.57	2.87E-04	2	223	0	0	3	211	1	0
Intron	3:37048579:A:G	29.44	5.28E-03	3	221	1	0	2	211	2	0
Intron	3:37050254:T:C	29.77	2.87E-02	0	214	11	0	0	202	12	1
Intron	3:37053271:T:C	30.80	8.96E-06	2	223	0	0	1	213	1	0
Intron	3:37053364:G:C	36.60	4.48E-05	0	225	0	0	0	214	1	0
Intron	3:37053401:AAG:A	30.08	NA	2	222	1	0	2	213	0	0
Thr212Thr	3:37053549:C:T	36.85	4.49E-05	0	225	0	0	0	214	1	0
Splice region	3:37055919:A:G	28.35	2.69E-05	0	223	2	0	1	214	0	0
Thr237Thr	3:37055956:C:T	35.28	NA	0	225	0	0	0	214	1	0
Intron	3:37056045:A:G	33.79	3.37E-03	0	224	1	0	0	214	1	0
Intron	3:37058951:C:T	50.65	NA	0	225	0	0	1	213	1	0
Intron	3:37059129:G:A	52.14	7.89E-04	0	224	1	0	0	215	0	0
Intron	3:37061777:T:A	36.6909	0.001934	0	223	2	0	0	215	0	0
His318His	3:37061870:C:T	34.70	2.69E-05	0	225	0	0	0	214	1	0

MLH1

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Gln900Gln	19:48618966:T:C	36.37	7.17E-05	0	225	0	0	0	214	1	0
Pro884Pro	19:48619154:C:T	36.83	1.35E-02	0	219	6	0	0	212	3	0
Intron	19:48619232:C:T	38.36	2.87E-03	0	222	3	0	0	215	0	0
Intron	19:48620861:A:G	19.67	1.71E-04	11	213	1	0	23	192	0	0
Intron	19:48620873:G:A	22.06	7.47E-05	4	221	0	0	10	204	1	0
Ala814Ala	19:48621036:C:G	55.42	4.29E-01	0	68	112	45	0	69	110	36
Intron	19:48622375:C:T	26.11	1.05E-01	4	182	37	2	3	163	42	7
Intron	19:48622376:G:A	26.14	0.00E+00	4	220	1	0	4	211	0	0
Intron	19:48622382:G:A	29.04	1.28E-04	2	223	0	0	2	212	1	0
Asp802Asp	19:48622427:A:G	40.69	5.29E-01	5	47	109	64	5	43	99	68
Intron	19:48622483:A:G	35.62	4.23E-01	9	67	107	42	5	67	107	36
Intron	19:48624625:C:T	10.67	4.42E-01	220	1	3	1	208	1	5	1
Intron	19:48626159:T:C	40.23	9.85E-05	0	225	0	0	0	214	1	0
Intron	19:48626181:C:T	40.73	1.16E-04	0	225	0	0	0	214	1	0
VAl729Val	19:48626236:A:G	41.79	3.56E-03	0	223	2	0	0	211	4	0
Intron	19:48626389:C:T	36.87	1.23E-02	0	218	7	0	0	206	9	0
Intron	19:48626604:G:A	35.30	5.89E-02	2	196	26	1	0	187	28	0
Asp661Asp	19:48630555:G:A	36.35	NA	0	224	1	0	0	215	0	0
Ala648Ala	19:48630594:C:T	34.51	4.83E-04	0	224	1	0	0	215	0	0
Intron	19:48631131:C:G	43.11	3.73E-02	0	214	11	0	0	196	18	1
Ala622Ala	19:48631233:G:A	42.38	0.00E+00	0	224	1	0	0	215	0	0
Val613Val	19:48631260:G:A	42.32	NA	0	224	1	0	0	215	0	0
Intron	19:48634298:C:T	37.66	1.12E-01	1	184	38	2	0	166	42	7
Intron	19:48634319:T:A	49.43	4.31E-01	2	66	113	44	0	68	111	36

LIG1 (part 1)

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Intron	19:48634473:A:G	22.73	NA	15	209	1	0	21	194	0	0
Intron	19:48640234:C:G	37.13	5.87E-02	1	198	26	0	0	187	28	0
Splice region	19:48640276:C:T	39.26	1.91E-04	0	225	0	0	0	214	1	0
Intron	19:48640363:C:T	28.61	1.03E-04	1	224	0	0	2	212	1	0
Intron	19:48640763:G:A	32.54	9.19E-05	1	223	1	0	1	214	0	0
Splice region	19:48643361:G:A	25.44	1.23E-03	5	219	1	0	10	203	2	0
Intron	19:48646775:C:G	48.56	7.99E-04	166	58	1	0	157	58	0	0
Intron	19:48646785:G:T	62.12	7.30E-03	134	89	2	0	135	78	2	0
Intron	19:48646883:T:C	74.99	5.42E-01	0	47	111	67	0	43	104	68
Splice region	19:48647225:G:A	40.80	1.07E-04	0	225	0	0	0	214	1	0
Intron	19:48647241:T:C	48.95	1.12E-01	0	185	38	2	0	166	42	7
Intron	19:48652975:G:A	32.94	3.64E-05	0	224	1	0	0	215	0	0
Intron	19:48654466:G:A	16.04	1.89E-04	60	165	0	0	60	154	1	0
Splice region	19:48654553:G:T	55.06	4.85E-01	0	58	116	51	0	57	103	55
Intron	19:48654606:G:T	42.61	1.25E-04	0	224	1	0	0	215	0	0
Intron	19:48660445:T:C	11.45	3.81E-01	202	8	12	3	196	9	8	2
Splice region	19:48664769:A:G	18.40	6.74E-04	30	192	3	0	30	185	0	0
Intron	19:48664798:G:A	15.61	3.81E-01	83	41	78	23	95	40	61	19
Intron	19:48665643:A:C	26.50	4.56E-04	3	222	0	0	6	208	1	0
Intron	19:48668795:G:A	42.29	3.81E-01	2	80	104	39	2	85	100	28
5'-UTR	19:48668830:G:A	40.58	1.23E-01	0	180	42	3	1	159	50	5
5'-UTR	19:48673428:C:T	21.03	NA	38	183	4	0	46	161	8	0
5'-UTR	19:48673458:G:A	27.98	3.84E-01	13	78	96	38	23	76	92	24

LIG1 (part 2)

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Phe14Phe	5:145826954:C:T	31.85	8.49E-04	3	222	0	0	1	212	2	0
Ala22Ala	5:145834625:C:T	17.73	2.09E-02	37	184	4	0	43	163	9	0
Pro32Pro	5:145834655:G:A	26.14	1.79E-05	2	222	1	0	0	215	0	0
Gln202Gln	5:145838614:A:G	36.79	1.63E-03	2	221	2	0	2	212	1	0
Gln224Gln	5:145838680:A:G	43.41	1.63E-04	4	221	0	0	2	212	1	0
Lys785Lys	5:145878222:A:G	18.06	9.01E-04	27	197	1	0	31	184	0	0
Thr976Thr	5:145887453:G:A	19.61	7.68E-02	26	175	23	1	20	167	28	0
Ser1040Ser	5:145890028:A:G	19.57	9.43E-01	12	0	16	197	14	1	29	171
Pro1087Pro	5:145890169:C:T	26.82	1.67E-02	5	212	8	0	1	208	6	0

TCERG1

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Asp117Asp	11:7324475:T:C	55.39	1.00E+00	0	0	0	225	0	0	0	215
Asp212Asp	11:7334764:C:T	36.45	2.69E-05	0	224	1	0	0	215	0	0
Intron	11:7335444:G:A	22.76	1.00E-04	18	207	0	0	9	205	1	0
Thr414Thr	11:7439264:C:T	42.40	1.39E-01	0	178	43	4	0	155	56	4
Splice region	11:7441737:T:C	22.67	1.22E-01	17	170	35	3	11	154	47	3
Glu451Glu	11:7441752:G:A	28.97	8.96E-06	2	222	1	0	2	213	0	0
Intron	11:7487978:G:A	14.94	3.78E-03	208	17	0	0	201	13	1	0

SYT9

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Intron	3:9792129:G:A	37.89	2.08E-03	0	225	0	0	0	214	1	0
Leu109Leu	3:9792818:G:C	38.45	4.48E-05	0	224	1	0	0	215	0	0
Ser143Ser	3:9793497:T:A	39.54	NA	0	224	1	0	0	215	0	0
Splice	3:9796384:G:A	34.63	2.06E-02	0	216	9	0	1	205	9	0
Gly200Gly	3:9796422:C:A	36.56	NA	0	225	0	0	0	214	1	0
Intron	3:9798140:C:G	34.14	2.24E-01	2	138	75	10	1	127	77	10
3'-UTR	3:9798572:G:A	40.66	0.00E+00	0	225	0	0	0	214	1	0
Intron	3:9800838:G:A	40.45	4.03E-04	0	224	1	0	0	215	0	0
3'-UTR	3:9807858:G:A	39.32	1.84E-05	0	225	0	0	0	214	1	0
Intron	3:9792129:G:A	37.89	2.08E-03	0	225	0	0	0	214	1	0
Leu109Leu	3:9792818:G:C	38.45	4.48E-05	0	224	1	0	0	215	0	0
Ser143Ser	3:9793497:T:A	39.54	NA	0	224	1	0	0	215	0	0
Splice	3:9796384:G:A	34.63	2.06E-02	0	216	9	0	1	205	9	0
Gly200Gly	3:9796422:C:A	36.56	NA	0	225	0	0	0	214	1	0
Intron	3:9798140:C:G	34.14	2.24E-01	2	138	75	10	1	127	77	10
3'-UTR	3:9798572:G:A	40.66	0.00E+00	0	225	0	0	0	214	1	0
Intron	3:9800838:G:A	40.45	4.03E-04	0	224	1	0	0	215	0	0

OGG1

				Early				Late			
Variant	Location	DP	gnomAD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Gln1421Gln	14:75483812:T:C	8.24E+01	5.33E-01	0	53	100	72	1	51	95	68
Intron	14:75485489:A:G	4.91E+01	1.00E+00	0	0	0	225	0	0	0	215
Intron	14:75485492:A:G	4.26E+01	1.80E-05	0	224	1	0	0	215	0	0
Intron	14:75485504:C:G	4.35E+01	1.62E-03	0	223	2	0	0	215	0	0
Intron	14:75485519:G:C	4.34E+01	7.40E-03	0	218	7	0	0	207	8	0
Intron	14:75489632:C:T	5.57E+01	4.67E-01	0	72	100	53	0	69	95	51
Intron	14:75489756:G:A	3.84E+01	9.20E-03	0	220	5	0	0	210	5	0
Intron	14:75497231:G:A	3.89E+01	7.87E-03	0	220	5	0	0	210	5	0
Intron	14:75497428:T:TA	3.09E+01	7.00E-02	97	127	1	0	87	125	3	0
Intron	14:75497428:TA:T	3.01E+01	7.00E-02	168	47	10	0	160	42	13	0
Intron	14:75498727:G:A	3.52E+01	2.91E-04	0	224	1	0	0	215	0	0
Intron	14:75498906:T:C	4.05E+01	8.96E-06	0	224	1	0	0	215	0	0
Intron	14:75505008:G:C	5.17E+01	4.84E-01	1	63	101	60	1	58	96	60
Intron	14:75505016:A:G	5.61E+01	1.00E+00	0	0	0	225	0	0	0	215
Intron	14:75506585:T:TA	2.08E+01	7.15E-04	45	179	1	0	43	171	1	0
Intron	14:75506586:T:A	2.18E+01	1.43E-02	71	152	2	0	81	130	4	0
Intron	14:75506586:T:TA	2.18E+01	NA	71	154	0	0	81	133	1	0
Intron	14:75506586:T:TATA	2.18E+01	1.43E-02	71	153	1	0	81	134	0	0
Intron	14:75506586:T:TATATATATA	2.18E+01	1.43E-02	71	153	1	0	81	134	0	0
Intron	14:75506721:A:G	2.90E+01	NA	8	217	0	0	17	197	1	0
Intron	14:75508280:G:GA	2.65E+01	6.70E-03	118	106	1	0	124	89	2	0
Intron	14:75508280:GA:G	2.69E+01	6.70E-03	154	67	4	0	143	65	7	0
Ser1137Ser	14:75508372:C:T	3.88E+01	0.00E+00	0	225	0	0	0	214	1	0
Glu1025Glu	14:75513284:T:C	3.86E+01	0.00E+00	0	225	0	0	0	214	1	0
Cys279Cys	14:75515522:G:A	2.43E+01	6.36E-04	16	208	1	0	16	199	0	0
Lys222Lys	14:75515693:C:T	2.55E+01	7.32E-03	4	214	7	0	3	204	8	0
Asp136Asp	14:75515951:A:G	2.93E+01	8.47E-03	0	222	3	0	0	212	3	0

MLH3

Appendix 6 – Coding variation identified in other candidate genes

Listed below are the coding variants for the genes described in 4.5.6. Non-synonymous damaging variants (CADD PHRED≥20 or LoF) are emboldened. Genomic locations are based on hg19/GRCh37. MAF annotations taken from v2.0.2 of gnomAD. Δ denotes variants where the most damaging consequence only occurs in the non-canonical transcript. Total N=440 (225 early; 215 late). LoF variants are marked with a [*]. DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Gly696Glu	**4:3133113:G:A**	**40.58**	**3.49E-03**	**22.7**	**0**	**221**	**4**	**0**	**0**	**213**	**2**	**0**
Gly893Arg	4:3137674:G:A	15.48	6.01E-02	24.4	90	115	18	2	83	107	24	1
Pro972del	4:3142351:TCCA:T	24.93	NA	NA	5	219	1	0	15	200	0	0
Leu989Val	4:3144512:C:G	27.74	3.58E-05	11.7	6	218	1	0	2	212	1	0
Val1064Ile	4:3148570:G:A	39.07	6.18E-02	3.8	0	191	31	3	0	177	35	3
Leu1074Pro	4:3148601:T:C	37.99	NA	25.8	0	224	1	0	0	215	0	0
Asp1082His	**4:3148624:G:C**	**38.40**	**2.34E-03**	**28.1**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ile1091Met	4:3148653:T:G	33.12	5.12E-02	16.7	2	215	8	0	1	201	13	0
Thr1260Met	**4:3162034:C:T**	**37.32**	**8.77E-04**	**22.2**	**0**	**223**	**2**	**0**	**0**	**211**	**4**	**0**
Met1306Ile	4:3174100:G:A	38.90	9.85E-05	18.5	0	225	0	0	0	214	1	0
Val1551Ala	**4:3182281:T:C**	**28.64**	**1.79E-05**	**26.6**	**1**	**223**	**1**	**0**	**0**	**214**	**1**	**0**
Cys1708Arg	4:3189510:T:C	37.83	4.57E-03	10.7	0	224	1	0	0	214	1	0
Thr1720Asn	4:3189547:C:A	22.71	1.11E-01	0.0	23	180	22	0	27	150	36	2
Arg2002His	**4:3208640:G:A**	**18.10**	**2.06E-04**	**24.8**	**30**	**194**	**1**	**0**	**27**	**186**	**2**	**0**
Tyr2309His	4:3215835:T:C	10.33	4.24E-01	15.23	224	0	1	0	213	0	2	0
Lys2337Arg	4:3216894:A:G	34.72	9.67E-04	0.1	0	225	0	0	0	214	1	0
Glu2444Asp	4:3221998:A:T	25.31	2.40E-03	5.9	2	222	1	0	3	211	1	0
Glu2643del	4:3230410:AGAG:A	13.03	6.99E-02	NA	181	28	14	2	188	17	10	0
Val2786Ile	4:3234980:G:A	14.71	3.01E-01	4.4	126	69	27	3	138	53	22	2

HTT

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Ala188Val	**3:37053328:C:T**	**36.90**	**8.97E-06**	**23.9**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Ile219Val	3:37053568:A:G	36.51	3.20E-01	13.9	2	123	84	16	3	86	98	28
Tyr379Cys	**3:37067225:A:G**	**40.33**	**0.00E+00**	**24.9**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ser406Asn	3:37067306:G:A	39.57	1.29E-03	7.3	1	224	0	0	0	214	1	0
Ala441Thr	3:37067410:G:A	37.49	5.29E-04	3.5	0	225	0	0	1	213	1	0
Lys618Glu	**3:37089130:A:G**	**42.06**	**5.63E-03**	**27.8**	**0**	**224**	**1**	**0**	**0**	**214**	**1**	**0**
Lys618Thr	**3:37089131:A:C**	**42.07**	**5.62E-03**	**28.0**	**0**	**224**	**1**	**0**	**0**	**214**	**1**	**0**
Gln689Arg	3:37090471:A:G	37.98	4.67E-04	17.0	0	224	1	0	0	215	0	0
Val716Met	**3:37092019:G:A**	**42.81**	**1.88E-03**	**24.6**	**0**	**224**	**1**	**0**	**0**	**214**	**1**	**0**
His718Tyr	**3:37092025:C:T**	**42.67**	**1.88E-04**	**29.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**

MLH1

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Arg46Gln	**3:9792107:G:A**	**37.93**	3.83E-03	**34.0**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Ala85Ser	3:9792744:G:T	43.12	2.79E-03	0.0	0	224	1	0	0	215	0	0
Arg197Trp	**3:9796411:A:T**	**35.69**	**2.60E-04**	**26.1**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg229Gln	**3:9796508:G:A**	**30.54**	**2.94E-04**	**23.4**	**3**	**222**	**0**	**0**	**2**	**212**	**1**	**0**
Ala288Val	3:9798270:C:T	17.40	2.46E-03	17.3	29	196	0	0	49	165	1	0
Gly308Glu	**3:9798475:G:A**	**37.75**	**6.57E-03**	**31.0**	**0**	**222**	**3**	**0**	**0**	**209**	**6**	**0**
Pro332Ala	3:9798773:C:G	31.77	2.23E-01	3.0	2	138	75	10	4	124	77	10
Gln362Ter[*],$^{\Delta}$	**3:9798863:C:T**	**20.00**	**NA**	**0.4**	**13**	**211**	**1**	**0**	**26**	**189**	**0**	**0**
Phe324Ser	3:9800893:T:C	40.63	1.74E-04	4.6	0	224	1	0	0	215	0	0
Gly348Glu	3:9807587:G:A	24.54	6.71E-04	13.1	4	221	0	0	3	211	1	0
Pro402Ala	3:9807748:C:G	31.26	NA	9.5	1	224	0	0	0	214	1	0

OGG1

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Val154Met	11:7324584:G:A	15.53	3.35E-02	0.9	81	133	11	0	81	124	10	0
Ser319Arg	11:7335424:C:G	23.94	5.71E-01	NA	15	43	97	70	24	41	86	64
Leu353Val	**11:7437285:C:G**	**34.65**	**9.84E-03**	**29.2**	**0**	**219**	**6**	**0**	**0**	**214**	**1**	**0**

SYT9

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Glu59Lys	2:190660537:G:A	35.95	4.68E-03	18.5	1	223	1	0	0	214	1	0
Thr75Ile	**2:190660586:C:T**	**35.75**	**1.08E-03**	**32.0**	**0**	**225**	**0**	**0**	**0**	**213**	**2**	**0**
Leu164ValfsTer4[*],$^{\Delta}$	**2:190670539:T:TA**	**28.09**	**2.44E-02**	**NA**	**63**	**86**	**74**	**2**	**94**	**75**	**46**	**0**
Lys163SerfsTer15[*],$^{\Delta}$	**2:190670539:TA:T**	**26.16**	**2.44E-02**	**NA**	**114**	**99**	**12**	**0**	**120**	**82**	**13**	**0**
Arg202Lys	2:190708712:G:A	38.67	1.38E-02	19.2	0	223	2	0	0	208	7	0
Lys433Arg	2:190719296:A:G	28.05	1.80E-05	0.0	6	218	1	0	4	211	0	0
Gly501Arg	**2:190719499:G:A**	**32.60**	**5.84E-04**	**25.9**	**4**	**221**	**0**	**0**	**7**	**207**	**1**	**0**
Leu524Ser	2:190719569:T:C	20.97	0.00E+00	0.0	17	207	1	0	24	191	0	0
Glu537Lys	2:190719607:G:A	17.31	2.97E-03	10.4	66	159	0	0	76	138	1	0
Arg569Gln	**2:190719704:G:A**	**26.05**	**1.26E-04**	**22.6**	**1**	**224**	**0**	**0**	**0**	**214**	**1**	**0**
Tyr793His	2:190732559:T:C	31.12	5.66E-04	0.1	0	225	0	0	1	213	1	0

PMS1

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Gly857Ala	7:6013049:C:G	16.14	2.67E-01	9.8	121	58	44	2	112	49	51	3
Leu729fs[*]	**7:6018314:TGA:T**	**28.43**	**1.56E-04**	**NA**	**11**	**213**	**1**	**0**	**9**	**204**	**2**	**0**
Thr728Ala	7:6018320:T:C	27.34	1.71E-04	11.6	14	210	1	0	16	197	2	0
Lys647Ter	**7:6026457:T:A**	**23.28**	**NA**	**38.0**	**24**	**200**	**1**	**0**	**29**	**186**	**0**	**0**
Met622Ile	7:6026530:C:T	36.90	2.34E-02	18.1	0	216	9	0	0	212	3	0
Thr597Ser	7:6026607:T:A	38.99	1.47E-02	0.0	0	219	6	0	0	209	6	0
Arg563Leu	7:6026708:C:A	39.18	8.77E-03	13.1	0	217	8	0	0	208	7	0
Lys541Glu	7:6026775:T:C	45.28	8.50E-01	0.1	3	8	59	155	4	3	59	149
Thr511Met	7:6026864:G:A	23.97	2.78E-04	5.4	3	221	1	0	7	207	1	0
Thr511Ala	7:6026865:T:C	23.85	3.02E-02	0.0	3	210	12	0	6	201	8	0
Thr485Lys	7:6026942:G:T	40.11	4.24E-02	0.0	0	207	18	0	1	201	13	0
His479Gln	7:6026959:G:C	41.14	3.63E-03	0.6	0	223	2	0	1	214	0	0
Pro470Ser	7:6026988:G:A	74.13	4.16E-01	0.1	1	83	108	33	0	63	112	40
Thr458Ser[◊]	7:6027024:T:A	41.12	NA	0.0	0	225	0	0	0	214	1	0
Val415Met[◊]	7:6027153:C:T	40.48	2.15E-04	10.0	0	225	0	0	0	214	1	0
Asn371His	7:6029464:T:G	32.77	NA	19.7	0	224	1	0	1	214	0	0
Asn335Ser	**7:6029571:T:C**	**20.50**	**4.66E-04**	**26.7**	**61**	**164**	**0**	**0**	**68**	**146**	**1**	**0**
Val159Met	**7:6042146:C:T**	**36.99**	**1.44E-04**	**24.1**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ala116Thr	**7:6043328:C:T**	**35.19**	**9.33E-06**	**33.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Phe80Phe	7:6043613:G:A	36.60	1.44E-04	10.9	0	224	1	0	1	214	0	0
Arg20Gln	7:6045627:C:T	42.62	7.60E-02	21.7	0	192	32	1	0	190	24	1
Ile18Val	7:6045634:T:C	39.86	1.15E-02	25.6	0	223	2	0	0	211	4	0

PMS2

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Glu387Gln	**5:145894518:C:G**	**36.78**	**2.69E-05**	**26.1**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ser366Thr	5:145894580:C:G	34.52	8.95E-06	0.0	0	225	0	0	0	214	1	0
Leu304fs[*]	**5:145894764:TGA:T**	**30.51**	**1.79E-05**	**NA**	**1**	**224**	**0**	**0**	**0**	**214**	**1**	**0**
Ile288Thr	5:145894814:A:G	28.27	NA	9.8	1	223	1	0	3	212	0	0
Pro284Ser	5:145894827:G:A	28.72	NA	23.9	2	222	1	0	0	215	0	0
Leu261Val	5:145894896:G:C	44.20	2.00E-01	0.0	0	138	78	9	0	140	68	7
Phe175fs[*]	**5:145895150:CTA:C**	**25.98**	**1.70E-04**	**NA**	**2**	**223**	**0**	**0**	**6**	**208**	**1**	**0**
Ala144Val	**5:145895246:G:A**	**38.58**	**1.37E-03**	**22.8**	**0**	**225**	**0**	**0**	**0**	**213**	**2**	**0**
Arg95Ter	**5:145895394:G:A**	**39.59**	**7.07E-03**	**36.0**	**0**	**216**	**9**	**0**	**0**	**211**	**4**	**0**
Pro40Leu	5:145895558:G:A	42.65	7.78E-02	16.4	0	196	28	1	0	182	33	0
Tyr27Ter	**5:145895596:G:T**	**37.28**	**8.93E-04**	**36.0**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Phe23Leu	**5:145895608:A:C**	**34.43**	**8.23E-04**	**25.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**

GPR151

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Met393Thr	8:103220407:A:G	26.96	NA	12.4	9	216	0	0	8	207	0	0
Arg56fs[*]	8:103250839:C:CG	23.45	8.33E-02	NA	48	149	25	3	53	141	16	5
Ala26Thr	8:103250930:C:T	26.58	8.66E-02	NA	19	170	31	5	29	161	19	6
Arg16Pro	8:103250959:C:G	30.92	2.19E-03	NA	1	224	0	0	1	209	5	0

RRM2B

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Glu1451Lys	**14:75483796:C:T**	**44.48**	**1.50E-03**	**23.6**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Ile1391fs[*]	**14:75483904:T:TCTATGGGAAGAAAGAATAACTTCAATTAGCAATATGA**	**54.05**	**8.95E-06**	**NA**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ser1167Asn	**14:75506684:C:T**	**32.09**	**8.95E-06**	**21.0**	**0**	**225**	**0**	**0**	**1**	**213**	**1**	**0**
Val1156Ile	**14:75506718:C:T**	**29.31**	**1.52E-04**	**24.7**	**7**	**217**	**1**	**0**	**16**	**199**	**0**	**0**
Asn1147Ile	**14:75508343:T:A**	**39.27**	**6.27E-04**	**31.0**	**0**	**224**	**1**	**0**	**0**	**213**	**2**	**0**
Asp1105Glu	14:75509146:G:T	40.45	3.69E-03	15.4	0	223	2	0	0	214	1	0
Asp1073Asn	**14:75513142:C:T**	**28.02**	**2.06E-04**	**26.9**	**1**	**224**	**0**	**0**	**2**	**212**	**1**	**0**
Pro1070Arg	**14:75513150:G:C**	**25.65**	**NA**	**23.9**	**2**	**223**	**0**	**0**	**7**	**207**	**1**	**0**
Val971Ile	14:75513448:C:T	31.94	5.29E-04	13.5	1	224	0	0	1	213	1	0
Ser966Pro	14:75513463:A:G	31.54	1.69E-02	0.2	1	216	8	0	1	209	5	0
Gln943Pro	**14:75513531:T:G**	**39.11**	**NA**	**23.4**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ser845Gly	14:75513826:T:C	44.38	7.30E-03	0.0	0	218	7	0	0	208	7	0
Pro844Leu	14:75513828:G:A	95.60	4.60E-01	5.1	0	77	97	51	0	74	93	48
Asn826Asp	14:75513883:T:C	125.36	1.00E+00	0.0	0	0	0	225	0	0	0	215
Arg797His	14:75513969:C:T	44.38	8.96E-05	0.0	0	223	2	0	0	215	0	0
Val741Phe	**14:75514138:C:A**	**43.98**	**7.37E-03**	**21.4**	**0**	**218**	**7**	**0**	**0**	**207**	**8**	**0**
Tyr720Cys	14:75514200:T:C	37.87	1.52E-04	0.0	0	224	1	0	0	215	0	0
Glu624Gln	14:75514489:C:G	27.10	1.09E-02	17.0	7	211	7	0	10	201	4	0
Arg546Ile	14:75514722:C:A	18.14	5.47E-04	1.5	32	192	1	0	35	180	0	0
Val420Ile	14:75515101:C:T	14.21	1.60E-02	0.0	104	117	4	0	117	97	1	0
Tyr238Ser	**14:75515646:T:G**	**17.32**	**4.50E-05**	**24.8**	**35**	**190**	**0**	**0**	**52**	**162**	**1**	**0**
Lys231Gln	14:75515668:T:G	19.70	1.88E-02	12.4	17	198	10	0	19	188	8	0
Phe50Tyr	14:75516210:A:T	39.45	5.19E-04	14.4	0	224	1	0	0	215	0	0
Lys3Arg	14:75516351:T:C	35.54	2.88E-04	8.9	0	224	1	0	0	215	0	0

MLH3

Appendix 7 – Coding variation in *CUBN*, *MUT* and *NOP14*

Listed below are the coding variants for *MUT*, *CUBN* and *NOP14* (see 4.7-4.9 which identifies these genes in unbiased exome-wide tests). Non-synonymous damaging variants (CADD PHRED≥20 or LoF) are emboldened. Genomic locations are based on hg19/GRCh37. MAF annotations taken from v2.0.2 of gnomAD. Total N=440 (225 early; 215 late). Δ denotes variants where the most damaging consequence only occurs in the non-canonical transcript. LoF variants are marked with a [*]. DP: Mean depth of variant site in early and late samples; NS: non-synonymous; N/C: not called (failed by-variant DP/GQ check); HomR: homozygote reference; Het: heterozygote; HomV: homozygote variant.

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Arg694Trp	**6:49403213:G:A**	**38.08**	**1.79E-05**	**35.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Ala676Thr	**6:49403267:C:T**	**36.03**	**8.96E-05**	**34.0**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Ile671Val	6:49403282:T:C	45.96	6.21E-01	3.0	0	40	111	74	0	27	124	64
Ala664Val	**6:49403302:G:A**	**35.42**	**6.46E-04**	**34.0**	**0**	**223**	**2**	**0**	**0**	**215**	**0**	**0**
Met622Thr	**6:49408010:A:G**	**37.08**	**8.96E-05**	**27.4**	**1**	**224**	**0**	**0**	**0**	**214**	**1**	**0**
Arg532His	6:49412433:C:T	21.87	3.65E-01	21.9	15	93	93	24	17	90	87	21
Ala499Thr	6:49415448:C:T	36.69	1.10E-01	16.0	1	175	45	4	0	174	40	1
Met375Ile	**6:49419386:C:T**	**25.25**	**2.78E-03**	**24.8**	**1**	**218**	**6**	**0**	**2**	**213**	**0**	**0**
Thr359Ser	6:49421305:G:C	18.16	NA	27.6	23	201	1	0	24	191	0	0
Lys335Asn	**6:49421376:T:A**	**17.96**	**0.00E+00**	**23.7**	**36**	**189**	**0**	**0**	**39**	**175**	**1**	**0**
Gln293Arg	6:49423826:T:C	24.24	5.11E-04	15.7	7	217	1	0	7	207	1	0
Asn189Ile	**6:49425591:T:A**	**34.86**	**0.00E+00**	**28.3**	**3**	**222**	**0**	**0**	**1**	**213**	**1**	**0**
Arg154His	**6:49425696:C:T**	**18.51**	**8.98E-06**	**23.1**	**20**	**204**	**1**	**0**	**21**	**194**	**0**	**0**
Ala141Val	**6:49425735:G:A**	**22.34**	**0.00E+00**	**31.0**	**4**	**220**	**1**	**0**	**8**	**207**	**0**	**0**
Thr78Ala	6:49426948:T:C	37.85	0.00E+00	12.3	0	224	1	0	0	215	0	0
Ile69Val	6:49426975:T:C	37.68	3.49E-03	0.2	0	224	1	0	0	213	2	0

MUT

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Arg3612Trp	10:16867012:G:A	44.32	2.35E-04	13.4	0	225	0	0	0	214	1	0
Val3596Met	10:16867060:C:T	43.78	2.73E-05	10.7	0	225	0	0	0	214	1	0
Gly3587Arg	**10:16870809:C:T**	**37.66**	**1.26E-04**	**29.5**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Asn3552Lys	10:16870912:G:T	48.73	8.40E-02	12.6	0	197	28	0	0	186	24	5
Thr3432Ser	10:16877080:G:C	37.63	5.29E-04	1.9	0	225	0	0	0	214	1	0
Thr3422Ile	10:16877110:G:A	36.74	2.75E-02	12.7	2	216	7	0	3	205	7	0
Gly3347Arg	10:16878375:C:T	33.46	4.22E-04	4.6	0	224	1	0	0	215	0	0
Ser3293Leu	**10:16882483:G:A**	**36.62**	**9.04E-06**	**28.3**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Thr3253Met	**10:16882952:G:A**	**37.12**	**5.38E-05**	**24.2**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Pro3242Ser	**10:16882986:G:A**	**35.09**	**5.38E-05**	**25.7**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ile3189Val	**10:16893332:T:C**	**44.80**	**4.48E-05**	**23.9**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Arg3148Trp	**10:16911647:G:A**	**39.20**	**3.59E-05**	**21.9**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Gly3114Ser	10:16911749:C:T	40.18	1.10E-02	32.0	0	224	1	0	0	212	3	0
Asp3100Asn	**10:16911791:C:T**	**36.13**	**1.79E-04**	**31.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Thr3069Ile	10:16916403:G:A	38.33	7.17E-05	0.3	0	225	0	0	0	214	1	0
Glu3002Gly	10:16918997:T:C	49.56	1.03E-01	11.0	0	170	53	2	0	167	47	1
Val2990Ile	10:16919034:C:T	45.29	1.97E-04	0.1	0	224	1	0	0	215	0	0
Ile2984Val	10:16919052:T:C	46.13	1.03E-01	6.4	0	173	50	2	0	168	47	0
Glu2968Gln	10:16930419:C:G	32.32	2.28E-02	23.3	0	222	3	0	1	205	9	0
Phe2965Ser	**10:16930427:A:G**	**31.62**	**2.71E-03**	**27.2**	**2**	**221**	**2**	**0**	**2**	**213**	**0**	**0**
Ala2914Val	10:16932384:G:A	44.26	1.42E-02	34.0	0	218	7	0	0	209	6	0
Val2891Ile	10:16932454:C:T	42.83	7.34E-04	0.0	0	224	1	0	0	215	0	0
Leu2879Ile	10:16932490:G:T	45.92	3.84E-02	24.1	0	209	16	0	0	197	18	0
Thr2800Ile	**10:16942635:G:A**	**41.50**	**1.20E-03**	**21.8**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ile2613Leu	**10:16948277:T:G**	**43.82**	**1.88E-04**	**23.7**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Pro2575Arg	10:16948390:G:C	43.95	1.60E-02	0.2	0	215	10	0	0	207	8	0
Asp2550Gly	**10:16949563:T:C**	**39.33**	**3.58E-05**	**26.3**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg2489Gln	10:16955877:C:T	37.27	8.95E-06	6.6	0	225	0	0	0	214	1	0

CUBN (part 1)

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Asp2455Glu	10:16955978:A:T	38.67	8.80E-03	6.4	0	221	4	0	0	210	5	0
Met2449Thr	10:16957036:A:G	36.71	2.91E-03	0.0	0	222	3	0	0	213	2	0
Ser2344Arg	10:16957998:A:T	33.82	NA	22.1	1	224	0	0	0	214	1	0
Glu2310fs	**10:16960686:ATAACCTC:A**	**36.43**	**2.78E-04**	**NA**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Phe2263Cys	**10:16961995:A:C**	**39.78**	**5.21E-03**	**25.0**	**0**	**219**	**6**	**0**	**0**	**215**	**0**	**0**
Val2221Ile	10:16962122:C:T	33.02	2.51E-04	1.2	0	225	0	0	0	214	1	0
Asn2157Asp	**10:16967417:T:C**	**32.74**	**6.75E-03**	**26.4**	**1**	**218**	**6**	**0**	**0**	**212**	**3**	**0**
Leu2153Phe	10:16967586:C:G	40.96	1.21E-01	24.9	0	168	55	2	0	165	49	1
Gly2093Ser	**10:16967768:C:T**	**28.56**	**1.80E-05**	**34.0**	**0**	**224**	**1**	**0**	**1**	**214**	**0**	**0**
Gln2048Glu	10:16970285:G:C	41.45	9.00E-06	12.6	0	225	0	0	0	214	1	0
Thr2031Ala	10:16975119:T:C	43.19	5.65E-04	0.0	0	224	1	0	0	215	0	0
Arg2030Gln	10:16975121:C:T	43.19	1.79E-05	0.6	0	224	1	0	0	215	0	0
Pro1975Leu	10:16979593:G:A	37.34	1.03E-02	21.6	0	217	8	0	0	210	5	0
Pro1971Thr	10:16979606:G:T	37.65	1.75E-02	25.0	0	216	9	0	0	207	8	0
Gly1953Arg	**10:16979660:C:G**	**38.34**	**8.98E-06**	**23.3**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Ser1935Gly	10:16979714:T:C	45.05	1.84E-01	6.7	0	145	75	5	0	148	61	6
Gln1930Ter[*]	**10:16979729:G:A**	**40.81**	**NA**	**37.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Ile1850Leu	10:16982031:T:A	32.20	6.30E-05	0.2	0	225	0	0	0	214	1	0
Gly1840Ser	10:16982061:C:T	34.92	1.81E-02	19.8	0	216	9	0	0	206	9	0
Arg1810Ter[*]	**10:16982151:G:A**	**37.99**	**2.06E-04**	**43.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Val1769Ile	10:16989271:C:T	36.98	9.16E-04	12.7	0	224	1	0	0	215	0	0
Thr1730Met	10:16990497:G:A	35.95	2.42E-04	14.0	0	225	0	0	0	214	1	0
Ala1690Val	10:16992011:G:A	32.45	3.33E-03	28.3	1	222	2	0	0	215	0	0
Cys1620Ter[*]	**10:16994384:G:T**	**41.08**	**NA**	**35.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Ser1606Cys	10:16996426:G:C	40.83	NA	23.7	0	224	1	0	0	215	0	0
Pro1559Ser	10:17024503:G:A	36.32	8.89E-01	10.5	2	3	47	173	0	3	51	161
Leu1484Gln	**10:17026178:A:T**	**37.49**	**8.95E-06**	**26.5**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Leu1465Met	10:17026236:G:T	37.46	NA	21.8	0	224	1	0	0	215	0	0

CUBN (part 2)

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Gly1390Ser	**10:17061832:C:T**	**40.15**	**0.00E+00**	**29.5**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg1349Cys	**10:17061955:G:A**	**40.16**	**0.00E+00**	**33.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Arg1311Gln	10:17083117:C:T	37.27	1.79E-05	13.9	0	224	1	0	0	215	0	0
Ser1251Thr	10:17085903:C:G	42.07	8.96E-06	12.4	0	225	0	0	0	212	3	0
Ala1202Thr	10:17087074:C:T	36.23	1.02E-03	10.1	0	225	0	0	0	214	1	0
Leu1119Ser	10:17088067:A:G	24.53	2.04E-03	0.4	9	214	2	0	6	208	1	0
His919Arg	10:17110639:T:C	25.22	6.70E-03	6.6	7	215	3	0	13	196	6	0
Ser865Asn	10:17113456:C:T	38.30	1.15E-02	8.6	0	219	6	0	0	213	1	1
Phe831Cys	**10:17113558:A:C**	**24.91**	**2.69E-05**	**24.3**	**6**	**218**	**1**	**0**	**6**	**209**	**0**	**0**
His730Tyr	10:17126383:G:A	37.23	5.63E-03	0.0	0	219	6	0	0	210	5	0
Leu648Phe	10:17130168:G:A	38.69	NA	27.9	0	224	1	0	0	215	0	0
Pro389Thr	10:17147521:G:T	74.58	6.66E-01	23.6	0	23	106	96	0	21	109	85
Phe253Ser	10:17156151:A:G	45.12	6.61E-01	0.1	2	25	97	101	0	16	97	102
Gly126Ala	10:17168770:C:G	37.90	NA	0.0	0	225	0	0	0	214	1	0
Gly66Arg	**10:17171176:C:T**	**37.58**	**1.07E-03**	**34.0**	**0**	**225**	**0**	**0**	**0**	**214**	**1**	**0**
Ile37Leu	10:17171656:T:G	37.95	2.95E-04	0.0	0	225	0	0	0	214	1	0

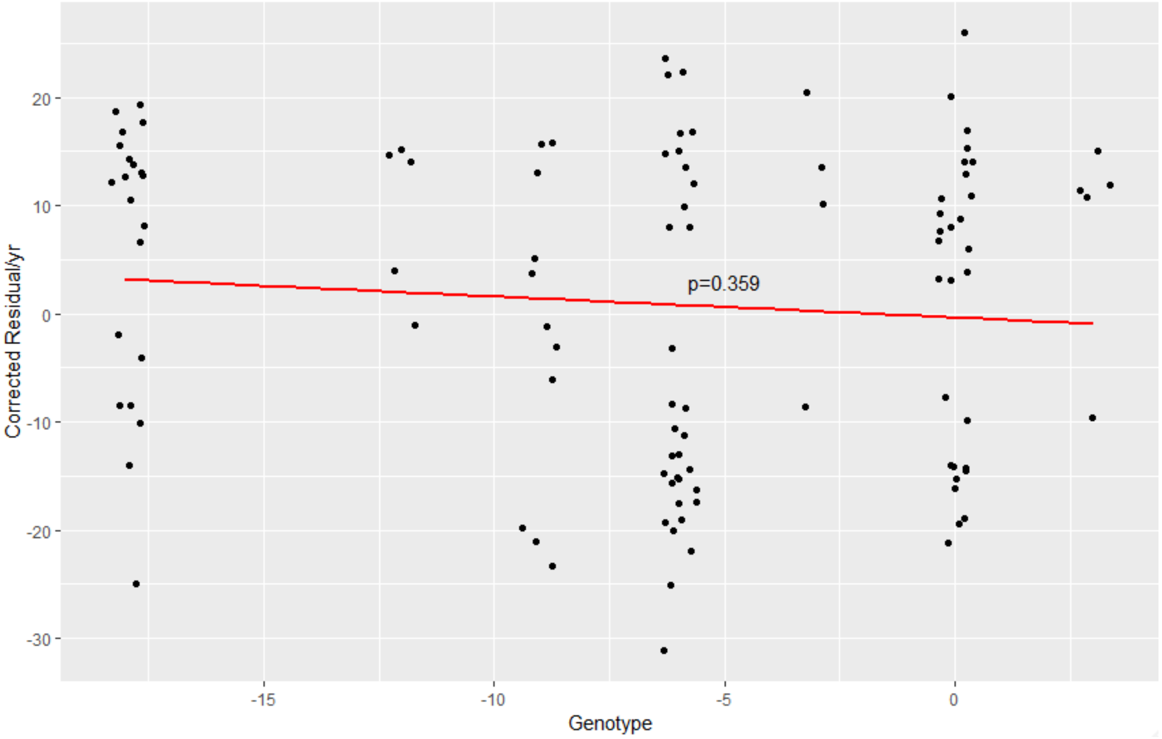
CUBN (part 3)

					Early				Late			
Variant	Location	DP	gnomAD	CADD	N/C	HomR	Het	HomV	N/C	HomR	Het	HomV
Asn780Asp$^{\Delta}$	4:2940027:T:C	13.53	5.90E-04	0.3	142	83	0	0	149	66	0	0
Glu775Lys	**4:2941066:C:T**	**26.81**	**0.00E+00**	**34.0**	**1**	**223**	**1**	**0**	**3**	**212**	**0**	**0**
Splice donor[*],$^{\Delta}$	4:2941265:C:T	24.63	2.69E-05	9.6	2	222	1	0	4	211	0	0
Leu749Phe	4:2941327:G:A	36.83	6.09E-04	10.9	0	225	0	0	0	214	1	0
Gln716Arg	4:2943361:T:C	45.99	4.16E-01	0.0	4	103	101	17	3	102	89	21
Arg697Cys	**4:2943419:G:A**	**37.98**	**5.29E-03**	**26.5**	**1**	**224**	**0**	**0**	**0**	**197**	**18**	**0**
Thr558Ala	**4:2946920:T:C**	**34.12**	**5.00E-03**	**22.8**	**0**	**218**	**7**	**0**	**0**	**208**	**7**	**0**
Arg537Gln	**4:2948164:C:T**	**29.97**	**4.48E-05**	**33.0**	**0**	**224**	**1**	**0**	**0**	**215**	**0**	**0**
Met525Thr	4:2948200:A:G	36.10	0.00E+00	12.9	0	225	0	0	0	214	1	0
Ala524Val	**4:2948203:G:A**	**36.12**	**4.12E-04**	**24.7**	**1**	**223**	**1**	**0**	**0**	**214**	**1**	**0**
Tyr424Cys	4:2951672:T:C	36.86	2.69E-05	10.0	0	225	0	0	1	214	0	0
Leu380Ser	4:2951804:A:G	70.73	4.26E-01	9.5	0	99	108	18	0	99	94	22
Arg295Gln	**4:2952959:C:T**	**21.04**	**8.06E-05**	**22.4**	**17**	**208**	**0**	**0**	**20**	**194**	**1**	**0**
Asp251Asn	**4:2954121:C:T**	**16.85**	**2.07E-04**	**34.0**	**100**	**125**	**0**	**0**	**119**	**95**	**1**	**0**
Leu171Pro	4:2956251:A:G	25.07	2.78E-04	11.5	11	213	1	0	11	204	0	0
Asp149Glu	4:2958422:A:T	20.59	NA	0.0	18	206	1	0	20	195	0	0
Asn141Ser	4:2958447:T:C	27.77	9.85E-05	8.4	2	222	1	0	0	215	0	0
Lys103Gln	4:2959356:T:G	25.37	5.37E-05	12.5	2	223	0	0	7	207	1	0
Gly8Trp	4:2965025:C:A	21.74	0.00E+00	19.8	149	76	0	0	132	83	0	0

NOP14

Appendix 8 – Modelling estimated STR length in MSH3

Each point represents one HD patient using the continous phenotype group where at least two of the four STRs observed in *MSH3* are called (N=108; 45 early, 63 late) (methods 2.7.2.3). Note that ‘called’ here can be a HomR, Het or HomV call. The y-axis uses corrected residual from MiSeq (pure CAG length to calculate age at motor onset residual). *MSH3* genotype is calculated from the short tandem Pro/Ala repeat where homozygotes are treated as having twice the genotype. Canonical genotype lengths are given 0 values. The x-axis has been jittered to improved readability. See 4.6.4 for this figure in context.



Appendix 9 – Average coverage of target genes from WES

Shown is the coverage for exons of canonical transcripts for 12 of the 13 candidate genes examined in 4.6 (note *MLH3* is not included). Calculations used bedtools as described (methods 2.7.2.4). N=483 (2 samples, 1 early and 1 late, failed coverage determination).

Gene	Transcript (canonical)	Plate 1	Plate 2	Plate 3	Plate 4	Plate 5	Plate 6	Plate 7	Average
EXO1	ENST00000366548.3	43.02	40.83	41.56	43.75	42.26	39.54	29.47	39.63
FAN1	ENST00000362065.4	32.75	29.33	31.01	30.10	29.65	29.54	24.74	29.08
GPR151	ENST00000311104.2	27.63	21.61	24.25	21.73	20.82	23.66	23.95	22.66
HTT	ENST00000355072.5	30.03	24.21	29.03	27.01	26.83	26.67	24.19	26.25
LIG1	ENST00000263274.7	70.91	53.21	66.40	56.94	52.64	57.91	59.50	57.40
MLH1	ENST00000231790.2	90.65	70.57	82.70	82.02	78.07	75.82	64.27	75.50
MSH3	ENST00000265081.6	50.07	44.54	48.46	48.91	48.40	44.23	33.61	44.69
OGG1	ENST00000302036.7	50.72	39.86	47.41	41.41	38.76	42.45	38.98	41.31
PMS1	ENST00000441310.2	39.43	36.85	38.17	42.08	39.08	35.74	25.95	36.33
PMS2	ENST00000265849.7	85.89	67.21	81.74	80.62	78.36	77.86	58.93	74.05
RRM2B	ENST00000251810.3	15.86	14.32	15.44	15.84	15.23	14.26	11.23	14.38
SYT9	ENST00000318881.6	23.96	18.02	21.24	19.25	17.62	18.41	17.37	18.59
TCERG1	ENST00000296702.5	32.71	25.95	33.32	25.59	23.88	23.97	22.84	25.60

Appendix 10 – Variant-by-variant logistic regression in exomes

The tables below show the top 20 (lowest *p* value) variant sites identified by variant-by-variant logistic regression analyses (Wald) in the dichotomous (N=440; 225 early, 215 late) HD exomes (see 4.7). The top table (A) shows results for nonsynonymous damaging (NSD) variants (CADD ≥20) or where CADD was missing, and the bottom table (B) for predicted loss-of-function (LoF) variants. No MAF or other filters were used. Covariates used: PC1-5 and mean variant depth. The top 20 variants are shown in each table.

A	Variant	DP	gnomAD	Gene	CADD	β	*SE*	*p*
	14:106330446:T:C	3.12E+01	2.05E-04	IGHJ4	NA	-66.15	15.38	1.71E-05
	5:43502596:G:A	1.85E+01	9.61E-06	C5orf34	23.00	-129.24	30.19	1.87E-05
	1:148342488:T:C	6.45E+01	3.51E-01	NBPF20	NA	1.05	0.26	5.59E-05
	5:140762659:C:T	1.64E+01	4.66E-04	PCDHGA7	34.00	-180.12	46.51	1.08E-04
	17:65716059:A:G	1.78E+01	1.87E-05	NOL11	28.90	-162.77	44.21	2.31E-04
	2:219868693:C:T	1.62E+01	3.52E-04	CCDC108	23.70	-281.95	77.92	2.96E-04
	19:53571679:T:C	1.61E+01	1.21E-03	ZNF160	23.60	-208.44	57.85	3.14E-04
	19:12358222:A:T	1.68E+01	NA	ZNF44	NA	-308.86	86.60	3.61E-04
	9:95237024:C:CTCA	1.78E+01	NA	ASPN	NA	-0.81	0.23	3.92E-04
	9:94495611:G:A	2.08E+01	5.49E-05	ROR2	32.00	336.65	95.84	4.44E-04
	3:108118027:C:A	3.73E+01	1.17E-04	MYH15	28.20	-225.62	65.85	6.12E-04
	7:156433296:A:G	1.50E+01	9.93E-04	C7orf13	NA	70.40	20.60	6.30E-04
	1:223991119:G:T	1.65E+01	8.99E-02	TP53BP2	28.90	1.22	0.36	6.68E-04
	22:38483165:T: TCATGGGGGA	2.17E+01	1.39E-04	BAIAP2L2	NA	-452.61	133.95	7.28E-04
	22:50969647:C:G	4.25E+01	4.01E-02	ODF3B	23.00	1.39	0.41	8.03E-04
	2:11295712:C:G	1.73E+01	0.00E+00	PQLC3	22.80	98.59	29.42	8.04E-04
	14:106330445:G:C	3.09E+01	0.00E+00	IGHJ4	NA	-363.57	108.83	8.36E-04
	14:92480795:T:A	2.42E+01	4.89E-04	TRIP11	20.80	-272.80	81.93	8.69E-04
	1:180148012:G:C	3.91E+01	1.59E-01	QSOX1	31.00	0.65	0.20	1.09E-03

Non-synonymous (NS) variants, CADD PHRED≥20 (or missing)

B Variant	DP	gnomAD	Gene	CADD	β	*SE*	*p*
19:38055666:TC:T	1.64E+01	NA	ZNF571	NA	-296.06	72.24	4.16E-05
6:154567863:C:T	2.56E+01	7.06E-02	OPRM1	34.00	-1.18	0.31	1.23E-04
11:31811483:TACTGTAA:T	2.32E+01	1.05E-03	PAX6	NA	-58.72	16.68	4.32E-04
16:81012303:C:A	2.52E+01	2.06E-01	CMC2	NA	0.64	0.19	6.03E-04
10:124214484:TTC:T	2.03E+01	0.00E+00	ARMS2	NA	-441.27	133.18	9.22E-04
X:46360423:G:A	2.11E+01	6.45E-03	ZNF674	32.00	-79.32	24.15	1.02E-03
15:55722882:C:A	2.19E+01	8.62E-02	DYX1C1	42.00	-0.88	0.27	1.28E-03
17:67190117:AAT:A	1.64E+01	1.04E-03	ABCA10	NA	-114.05	35.42	1.28E-03
2:107074109:G:A	1.35E+01	9.21E-04	RGPD3	35.00	-69.70	22.20	1.69E-03
1:153907305:G:GC	4.30E+01	1.30E-02	DENND4B	NA	1.35	0.45	2.62E-03
17:28887134:G:A	8.09E+01	6.83E-02	TBC1D29	0.00	0.86	0.29	2.66E-03
5:39316049:G:GCATAAAA	1.61E+01	2.73E-05	C9	NA	-182.71	61.09	2.78E-03
11:31811483:TA:T	2.25E+01	1.05E-03	PAX6	NA	-105.26	35.26	2.83E-03
6:17601125:C:CA	1.77E+01	NA	FAM8A1	NA	63.42	21.29	2.89E-03
7:99722497:C:T	1.48E+01	1.53E-04	CNPY4	38.00	-191.81	64.88	3.11E-03
14:106330441:CA:C	3.13E+01	NA	IGHJ4	NA	-342.87	117.02	3.39E-03
2:98810932:C:T	1.98E+01	4.48E-05	VWA3B	35.00	-239.04	83.55	4.22E-03
6:17601127:G:GCCTAAC	1.76E+01	NA	FAM8A1	NA	60.78	21.35	4.41E-03
5:149374879:CT:C	3.22E+01	6.57E-01	TIGD6	NA	0.44	0.15	4.42E-03
14:22690191:G:A	3.39E+01	7.23E-02	TRAV35	NA	-0.78	0.27	4.55E-03

Loss-of-function (LoF) variants

Appendix 11 – Whole-exome logistic regression weighting on MAF

Indicated are the β, *SE* and *p* values for the top 15 genes using logistic burden regression (Wald) in Hail for the dichotomous population weighting on minor allele frequency (MAF) from gnomAD (N=440; 225 early, 215 late) (see 4.7). Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Weighting used MAF from gnomAD (1/MAF); for the imputed values (bottom table), imputed 1/MAF used 1000000. Indicated at the top of each column are the adjusted (adj) values for each MAF cut-off, which is a count of how many genes/ORFs were tested that resulted in a *p* value and have >5 variants at the MAF tested. β: standardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

MAF = 0.1% (N=698)				MAF = 1% (N=2768)				MAF = 2% (N=3424)			
Gene	β	*SE*	*p*	Gene	β	*SE*	*p*	Gene	β	*SE*	*p*
DENND4B	9.16E-02	2.99E-02	2.21E-03	CUBN	5.89E-02	1.64E-02	3.29E-04	DENND4B	9.33E-02	3.00E-02	1.89E-03
CUBN	7.12E-02	2.58E-02	5.72E-03	DENND4B	9.18E-02	2.99E-02	2.17E-03	SIPA1L2	7.11E-02	2.45E-02	3.72E-03
MYO18B	4.63E-02	1.95E-02	1.75E-02	SIPA1L2	7.10E-02	2.45E-02	3.81E-03	ERAP2	-7.87E-02	2.75E-02	4.24E-03
MMP21	-9.75E-02	4.35E-02	2.49E-02	ERAP2	-7.92E-02	2.75E-02	3.98E-03	CUBN	2.39E-02	8.45E-03	4.60E-03
MACF1	-4.57E-02	2.04E-02	2.50E-02	NOP14	-3.94E-02	1.52E-02	9.58E-03	AKR1C3	6.37E-02	2.28E-02	5.23E-03
GLDC	-5.94E-02	2.70E-02	2.79E-02	GLI3	4.89E-02	1.89E-02	9.68E-03	ZNF462	-1.22E-01	4.36E-02	5.25E-03
STAB1	-4.54E-02	2.08E-02	2.88E-02	C9	7.42E-02	2.89E-02	1.02E-02	PGC	5.18E-01	1.91E-01	6.70E-03
ITGB4	-3.85E-02	1.78E-02	3.05E-02	CACNA1I	-7.01E-02	2.82E-02	1.30E-02	TRIM66	-4.06E-02	1.54E-02	8.25E-03
TENM2	-5.38E-02	2.52E-02	3.33E-02	ATP1A4	-5.02E-02	2.02E-02	1.32E-02	GLI3	4.89E-02	1.89E-02	9.58E-03
TRPM1	5.49E-02	2.59E-02	3.40E-02	ANXA11	7.86E-02	3.19E-02	1.37E-02	C9	7.42E-02	2.89E-02	1.01E-02
FBRSL1	9.24E-02	4.36E-02	3.41E-02	FAN1	4.02E-02	1.64E-02	1.45E-02	DNAJA4	-7.01E-02	2.73E-02	1.02E-02
CEP350	6.00E-02	2.91E-02	3.89E-02	ZNF462	-1.07E-01	4.40E-02	1.54E-02	NLRP1	-1.13E-01	4.42E-02	1.03E-02
INTS1	-6.11E-02	3.00E-02	4.14E-02	KIAA0556	-5.25E-02	2.19E-02	1.63E-02	RNMTL1	3.14E-02	1.23E-02	1.05E-02
SYT10	-9.07E-02	4.44E-02	4.14E-02	DENND2C	6.93E-02	2.90E-02	1.68E-02	NOP14	-3.89E-02	1.53E-02	1.07E-02
EIF4G1	-4.61E-02	2.26E-02	4.14E-02	MYO1A	4.08E-02	1.72E-02	1.76E-02	PLEKHA7	-2.58E-02	1.03E-02	1.21E-02

MAF not imputed, weighted (1/MAF)

MAF = 0.1% (N=736)				MAF = 1% (N=2796)				MAF = 2% (N=3450)			
Gene	β	SE	p	Gene	β	SE	p	Gene	β	SE	p
DENND4B	1.38E-06	4.53E-07	2.32E-03	DENND4B	1.38E-06	4.52E-07	2.28E-03	DENND4B	1.40E-06	4.54E-07	1.98E-03
GAPDHS	-1.24E-05	6.80E-06	6.86E-02	TTN-AS1	1.91E-02	9.14E-03	3.64E-02	PGC	3.91E-02	1.44E-02	6.70E-03
FJX1	-7.29E-04	5.11E-04	1.54E-01	ENTHD2	-1.45E-02	6.96E-03	3.73E-02	DNAJA4	-2.67E-02	1.14E-02	1.95E-02
PRKRIR	-7.31E-05	5.20E-05	1.60E-01	ENPP7	4.81E-03	2.36E-03	4.17E-02	ATG4B	-1.93E-02	8.46E-03	2.28E-02
GPR98	1.34E-06	1.00E-06	1.81E-01	NSG1	8.63E-03	4.32E-03	4.58E-02	SAMD10	-1.92E-02	8.55E-03	2.48E-02
RRH	7.76E-05	5.81E-05	1.81E-01	NDUFA10	1.62E-02	8.20E-03	4.77E-02	CENPBD1	1.19E-02	5.49E-03	3.06E-02
OR4K1	-2.74E-04	2.06E-04	1.83E-01	CYR61	1.44E-02	7.40E-03	5.11E-02	KIAA1644	-3.40E-02	1.59E-02	3.29E-02
GPR142	7.07E-04	5.33E-04	1.84E-01	OR5H2	-1.25E-02	6.50E-03	5.35E-02	VPS37C	-2.61E-02	1.24E-02	3.53E-02
PLXND1	-1.06E-04	8.08E-05	1.89E-01	TNFRSF13C	1.08E-02	5.61E-03	5.36E-02	ENTHD2	-1.44E-02	6.96E-03	3.80E-02
SLIT3	4.58E-04	3.50E-04	1.91E-01	MCEE	7.09E-03	3.69E-03	5.48E-02	TTN-AS1	1.90E-02	9.14E-03	3.81E-02
GOLGB1	-1.44E-06	1.12E-06	1.98E-01	GAPVD1	6.41E-03	3.43E-03	6.17E-02	OR1I1	-1.16E-02	5.65E-03	3.97E-02
IFT122	3.96E-05	3.08E-05	1.99E-01	TXNDC17	1.65E-02	8.88E-03	6.32E-02	ENPP7	4.83E-03	2.37E-03	4.13E-02
C19orf44	-3.05E-04	2.39E-04	2.03E-01	TTC6	-2.32E-03	1.26E-03	6.56E-02	RAP1B	-2.93E-02	1.44E-02	4.26E-02
USH1C	1.04E-04	8.13E-05	2.03E-01	OXR1	-1.29E-02	7.04E-03	6.67E-02	KIF9	1.87E-02	9.32E-03	4.51E-02
GLI1	-6.09E-04	4.83E-04	2.08E-01	GAPDHS	-1.24E-05	6.79E-06	6.70E-02	NSG1	8.59E-03	4.32E-03	4.66E-02

Imputed MAF, weighted (1/MAF)

Appendix 12 – Whole-exome logistic regression weighted on deleteriousness

Indicated are the β, *SE* and *p* values for the top 15 genes using logistic burden regression (Wald) in Hail weighting on deleteriousness (CADD PHRED score) (see 4.7). Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were weighted using CADD PHRED; missing CADD scores were imputed as 15. Indicated at the top of each column are the adjusted (adj) values for each MAF cut-off, which is a count of how many genes/ORFs were tested that resulted in a *p* value and have >5 variants at the MAF tested. β: standardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

MAF = 0.1% (N=1590 adj)				MAF = 1% (N=4261 adj)				MAF = 2% (N=5084 adj)			
Gene	β	*SE*	*p*	Gene	β	*SE*	*p*	Gene	β	*SE*	*p*
DENND4B	9.16E-02	2.99E-02	2.21E-03	CUBN	5.89E-02	1.64E-02	3.29E-04	DENND4B	9.33E-02	3.00E-02	1.89E-03
CUBN	7.12E-02	2.58E-02	5.72E-03	DENND4B	9.18E-02	2.99E-02	2.17E-03	SIPA1L2	7.11E-02	2.45E-02	3.72E-03
MYO18B	4.63E-02	1.95E-02	1.75E-02	SIPA1L2	7.10E-02	2.45E-02	3.81E-03	ERAP2	-7.87E-02	2.75E-02	4.24E-03
MMP21	-9.75E-02	4.35E-02	2.49E-02	ERAP2	-7.92E-02	2.75E-02	3.98E-03	CUBN	2.39E-02	8.45E-03	4.60E-03
MACF1	-4.57E-02	2.04E-02	2.50E-02	NOP14	-3.94E-02	1.52E-02	9.58E-03	AKR1C3	6.37E-02	2.28E-02	5.23E-03
GLDC	-5.94E-02	2.70E-02	2.79E-02	GLI3	4.89E-02	1.89E-02	9.68E-03	ZNF462	-1.22E-01	4.36E-02	5.25E-03
STAB1	-4.54E-02	2.08E-02	2.88E-02	C9	7.42E-02	2.89E-02	1.02E-02	PGC	5.18E-01	1.91E-01	6.70E-03
ITGB4	-3.85E-02	1.78E-02	3.05E-02	CACNA1I	-7.01E-02	2.82E-02	1.30E-02	TRIM66	-4.06E-02	1.54E-02	8.25E-03
TENM2	-5.38E-02	2.52E-02	3.33E-02	ATP1A4	-5.02E-02	2.02E-02	1.32E-02	GLI3	4.89E-02	1.89E-02	9.58E-03
TRPM1	5.49E-02	2.59E-02	3.40E-02	ANXA11	7.86E-02	3.19E-02	1.37E-02	C9	7.42E-02	2.89E-02	1.01E-02
FBRSL1	9.24E-02	4.36E-02	3.41E-02	FAN1	4.02E-02	1.64E-02	1.45E-02	DNAJA4	-7.01E-02	2.73E-02	1.02E-02
CEP350	6.00E-02	2.91E-02	3.89E-02	ZNF462	-1.07E-01	4.40E-02	1.54E-02	NLRP1	-1.13E-01	4.42E-02	1.03E-02
INTS1	-6.11E-02	3.00E-02	4.14E-02	KIAA0556	-5.25E-02	2.19E-02	1.63E-02	RNMTL1	3.14E-02	1.23E-02	1.05E-02
SYT10	-9.07E-02	4.44E-02	4.14E-02	DENND2C	6.93E-02	2.90E-02	1.68E-02	NOP14	-3.89E-02	1.53E-02	1.07E-02
EIF4G1	-4.61E-02	2.26E-02	4.14E-02	MYO1A	4.08E-02	1.72E-02	1.76E-02	PLEKHA7	-2.58E-02	1.03E-02	1.21E-02

Appendix 13 – Whole-exome linear regression weighted on deleteriousness

Indicated are the β, *SE* and *p* values for the top 15 genes using linear burden regression in Hail weighting on deleteriousness (CADD PHRED) (see 4.8). Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were weighted using CADD PHRED; missing CADD scores were imputed as 15 for both tables. Indicated at the top of each column are the adjusted (adj) values for each MAF cut-off, which is a count of how many genes/ORFs were tested that resulted in a *p* value and have >5 variants at the MAF tested. β: standardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate.

MAF = 0.1% (N=1614)				MAF = 1% (N=4307)				MAF = 2% (N=5129)			
Gene	β	*SE*	*p*	Gene	β	*SE*	*p*	Gene	β	*SE*	*p*
CUBN	-0.374	0.102	2.70E-04	CUBN	-0.305	0.069	1.30E-05	ZNF462	0.495	0.118	3.24E-05
CACNA1G	-0.616	0.174	4.31E-04	SIPA1L2	-0.429	0.116	2.37E-04	SIPA1L2	-0.430	0.116	2.27E-04
ZNF462	0.479	0.140	6.83E-04	TEKT1	0.806	0.232	5.55E-04	PGC	-1.315	0.376	5.22E-04
FBRSL1	-0.543	0.162	8.54E-04	ERAP2	0.391	0.115	7.06E-04	TEKT1	0.804	0.232	5.68E-04
CGN	0.673	0.208	1.26E-03	KIAA0319	-0.536	0.158	7.48E-04	KIAA0319	-0.542	0.158	6.57E-04
DENND4B	-0.517	0.162	1.54E-03	ZNF462	0.450	0.133	8.13E-04	ERAP2	0.389	0.115	7.77E-04
GLRA4	-0.536	0.170	1.77E-03	CACNA1G	-0.524	0.157	9.22E-04	NLRP1	0.419	0.126	9.14E-04
KIAA0319	-0.573	0.200	4.30E-03	ANXA11	-0.464	0.144	1.34E-03	MUT	-0.397	0.121	1.10E-03
GLDC	0.387	0.137	5.01E-03	GLRA4	-0.542	0.171	1.61E-03	ANXA11	-0.465	0.143	1.26E-03
DMRT2	-0.681	0.244	5.55E-03	SLC38A2	0.563	0.180	1.91E-03	SLC38A2	0.565	0.180	1.85E-03
PRKRIR	0.471	0.169	5.64E-03	ENPP7	-0.403	0.130	2.02E-03	AKR1C3	-0.388	0.124	1.85E-03
COL17A1	0.424	0.154	6.24E-03	GRTP1	0.682	0.222	2.26E-03	CUBN	-0.157	0.050	1.92E-03
NUP210L	-0.472	0.174	6.90E-03	AMPD2	-1.209	0.402	2.80E-03	GRTP1	0.685	0.221	2.09E-03
MMP21	0.413	0.152	7.01E-03	NMUR2	0.551	0.185	3.09E-03	OR1I1	0.315	0.102	2.24E-03
SON	0.464	0.172	7.04E-03	NUP210L	-0.478	0.161	3.20E-03	NCF2	0.447	0.146	2.36E-03

Weighted (deleteriousness), uncorrected AMO residual (polyglutamine length – 2)

MAF = 0.1% (N=1614)				MAF = 1% (N=4307)				MAF = 2% (N=5129)			
Gene	β	*SE*	*p*	Gene	β	*SE*	*p*	Gene	β	*SE*	*p*
CUBN	-0.366	0.097	1.94E-04	CUBN	-0.292	0.066	1.35E-05	ZNF462	0.458	0.113	6.05E-05
FBRSL1	-0.544	0.155	4.84E-04	SIPA1L2	-0.411	0.111	2.32E-04	SIPA1L2	-0.412	0.111	2.24E-04
CACNA1G	-0.582	0.167	5.16E-04	ERAP2	0.373	0.110	7.43E-04	RNMTL1	-0.220	0.063	5.81E-04
CGN	0.665	0.199	8.74E-04	KIAA0319	-0.507	0.151	8.62E-04	CUBN	-0.166	0.048	6.25E-04
GLRA4	-0.545	0.163	9.01E-04	DENND4B	-0.521	0.155	8.66E-04	DENND4B	-0.527	0.155	7.60E-04
DENND4B	-0.518	0.155	9.20E-04	GLRA4	-0.547	0.163	8.90E-04	ERAP2	0.373	0.110	7.80E-04
ZNF462	0.425	0.134	1.68E-03	CACNA1G	-0.497	0.151	1.03E-03	KIAA0319	-0.512	0.151	7.83E-04
GLDC	0.412	0.131	1.79E-03	ANXA11	-0.443	0.138	1.37E-03	NLRP1	0.401	0.120	9.30E-04
DMRT2	-0.681	0.234	3.74E-03	MUT	-0.369	0.116	1.58E-03	CACNA1G	-0.498	0.150	1.01E-03
SON	0.462	0.164	5.07E-03	ZNF708	-0.703	0.223	1.72E-03	GLRA4	-0.535	0.163	1.09E-03
TEKT2	-0.639	0.227	5.11E-03	NCF2	0.440	0.140	1.79E-03	AKR1C3	-0.389	0.119	1.13E-03
KIAA0319	-0.528	0.191	5.94E-03	ZNF462	0.402	0.128	1.79E-03	ANXA11	-0.445	0.137	1.29E-03
PRKRIR	0.448	0.162	6.01E-03	GRTP1	0.660	0.213	2.03E-03	PGC	-1.152	0.361	1.54E-03
ZNF530	-0.709	0.261	6.75E-03	SLC38A2	0.533	0.173	2.17E-03	ZNF708	-0.704	0.223	1.68E-03
NUP210L	-0.447	0.167	7.54E-03	SON	0.433	0.143	2.69E-03	NCF2	0.442	0.140	1.69E-03

Weighted (deleteriousness), corrected AMO residual (pure *HTT* CAG length)

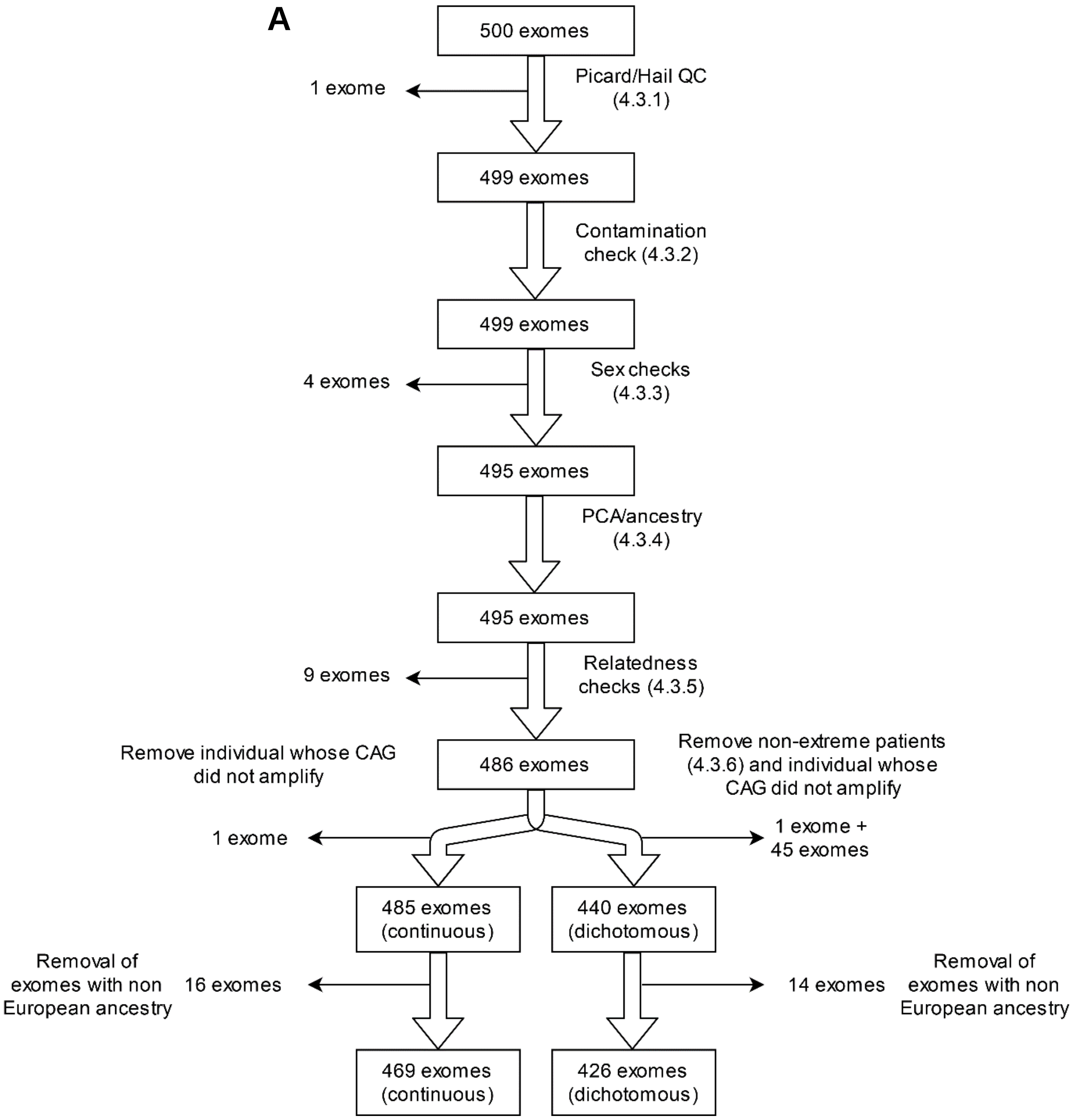
Appendix 14 – Whole-exome continuous SKAT-O test weighted on deleteriousness

Indicated are the *p* values for the top 15 genes from a continuous whole-exome SKAT-O test using the corrected AMO residual (pure CAG length) and weighted on deleteriousness (see 4.9). Filters used (for variants): MAF (1%), NS damaging (LoF or CADD PHRED ≥20), missingness ≤25%. Covariates used (for samples): PC1-5, BVR, mean variant depth. Variants were weighted on CADD PHRED score, and missing scores were imputed as 30 for LoF variants and 20 for other NS variants missing CADD PHRED scores. Only genes with >5 variants were tested, leaving 4737 genes (Bonferroni threshold *p*=1.06E-05). N=485 exomes from the continuous HD group. BVR: baseline variant rate.

Gene	*p*
CUBN	4.48E-05
SIPA1L2	7.17E-04
ERAP2	8.12E-04
MUT	8.79E-04
GLRA4	9.92E-04
DENND4B	1.31E-03
ANXA11	1.87E-03
KIAA0319	1.95E-03
ZNF708	1.96E-03
NOP14	2.02E-03
CACNA1G	2.04E-03
GRTP1	2.20E-03
NCF2	2.69E-03
SLC38A2	3.04E-03
FBP2	3.06E-03

Appendix 15 – Whole-exome auxiliary analysis excluding non-European individuals

Shown in this section are the results from whole-exome logistic and linear burden regression, and dichotomous and continuous SKAT(-O) tests where individuals with non-European ancestries (as determined by Peddy in 4.3.4) were excluded. (A) An extended flowchart showing the QC process used (this is a modified version of Fig 4.9 in 4.3.6 that shows the removal of individuals with non-European ancestry); (B) Logistic burden regression (Wald); (C) Linear burden regression using the uncorrected residual; (D) Linear burden regression using the corrected residual; (E) SKAT and SKAT-O tests at MAF≤1%; (F) Logistic burden regression (Wald) for candidate genes; (G) Linear burden regression (corrected residual) for candidate genes; (H) SKAT(-O) analyses (both logistic and linear regression) for candidate genes. All logistic tests had N=426 exomes and all linear/continuous tests had N=469 individuals. For B-E, the 15 genes with the lowest *p* value and >5 variants at the tested MAF are shown, and whole-exome genes are emboldened. For F-H, nominally significant *p* values (*p*<0.05) are emboldened. Logistic Bonferroni *p* thresholds: MAF 0.1% *p*=3.57E-05 (1399 genes); MAF 1% *p*=1.23E-05 (4052 genes); MAF 2% *p*=1.02E-05 (4906 genes). For linear Bonferroni *p* thresholds: MAF 0.1% *p*=2.98E-05 (1679 genes); MAF 1% *p*=1.12E-05 (4482 genes); MAF 2% *p*=9.43E-06 (5302 genes). Filters used (for variants): VQSR≥98.5, MAF (0.1, 1 and 2%), NS damaging (LoF or CADD PHRED ≥20), call rate ≥75%. Covariates used (for samples): PC1-5, BVR, mean variant depth. No weighting of variants was used. *B*: unstandardised beta; SE: standard error; MAF: minor allele frequency; PC: Principal component; BVR: baseline variant rate; Uncor: linear regression on the uncorrected (polyglutamine-2) AMO residual; Cor: Linear regression on the corrected (pure CAG) AMO residual.

A
500 exomes
1 exome
Picard/Hail QC (4.3.1)
499 exomes
Contamination check (4.3.2)
499 exomes
4 exomes
Sex checks (4.3.3)
495 exomes
PCA/ancestry (4.3.4)
495 exomes
9 exomes
Relatedness checks (4.3.5)
Remove individual whose CAG did not amplify
486 exomes
Remove non-extreme patients (4.3.6) and individual whose CAG did not amplify
1 exome
1 exome + 45 exomes
485 exomes (continuous)
440 exomes (dichotomous)
Removal of exomes with non European ancestry
16 exomes
14 exomes
Removal of exomes with non European ancestry
469 exomes (continuous)
426 exomes (dichotomous)

B

MAF≤0.1% (N=1399 adj)				MAF≤1% (N=4052 adj)				MAF≤2% (4906 adj)			
Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*
DENND4B	1.51E+00	4.74E-01	1.38E-03	CUBN	1.57E+00	4.36E-01	3.27E-04	DENND4B	1.55E+00	4.75E-01	1.08E-03
CUBN	1.78E+00	6.58E-01	6.74E-03	DENND4B	1.52E+00	4.73E-01	1.31E-03	SIPA1L2	1.87E+00	6.39E-01	3.39E-03
PCDH15	-1.59E+00	6.46E-01	1.41E-02	SIPA1L2	1.87E+00	6.38E-01	3.45E-03	ZNF462	-2.98E+00	1.04E+00	4.08E-03
MYO18B	1.24E+00	5.28E-01	1.93E-02	GLI3	1.59E+00	5.57E-01	4.24E-03	PCDH15	-1.11E+00	3.85E-01	4.12E-03
TRPM1	1.85E+00	8.09E-01	2.23E-02	ERAP2	-2.90E+00	1.04E+00	5.27E-03	GLI3	1.59E+00	5.56E-01	4.34E-03
DNHD1	1.10E+00	4.96E-01	2.60E-02	PCDH15	-1.52E+00	5.67E-01	7.37E-03	ERAP2	-2.88E+00	1.04E+00	5.55E-03
TENM2	-1.48E+00	6.68E-01	2.68E-02	MYO1A	1.29E+00	5.01E-01	9.96E-03	PGC	2.12E+00	7.73E-01	6.02E-03
MMP21	-2.36E+00	1.07E+00	2.76E-02	NOP14	-9.94E-01	3.90E-01	1.09E-02	CUBN	6.47E-01	2.42E-01	7.52E-03
EIF4G1	-1.36E+00	6.18E-01	2.79E-02	FAN1	1.05E+00	4.14E-01	1.10E-02	DNAJA4	-2.00E+00	7.67E-01	9.08E-03
LRIG1	1.71E+00	7.90E-01	3.08E-02	SPECC1L	-1.22E+00	4.83E-01	1.12E-02	RNMTL1	7.77E-01	2.99E-01	9.37E-03
MYO1A	2.25E+00	1.04E+00	3.13E-02	C9	1.95E+00	7.72E-01	1.17E-02	WDR64	-1.29E+00	5.00E-01	9.64E-03
FBRSL1	2.31E+00	1.08E+00	3.18E-02	ZNF462	-2.63E+00	1.05E+00	1.23E-02	SPECC1L	-1.25E+00	4.83E-01	9.78E-03
GLDC	-1.67E+00	7.89E-01	3.40E-02	UNC5B	-1.60E+00	6.50E-01	1.36E-02	MYO1A	1.29E+00	5.01E-01	1.02E-02
SYT10	-2.23E+00	1.07E+00	3.66E-02	ANXA11	1.87E+00	7.67E-01	1.49E-02	NLRP1	-2.65E+00	1.05E+00	1.14E-02
MEGF8	2.23E+00	1.08E+00	3.97E-02	CACNA1I	-1.62E+00	6.71E-01	1.56E-02	C9	1.95E+00	7.72E-01	1.15E-02

Logistic burden regression

C

MAF≤0.1% (N=1679)				MAF≤1% (N=4482)				MAF≤2% (N=5305)			
Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*
CACNA1G	-16.100	4.439	3.19E-04	CUBN	-8.745	1.985	1.31E-05	ZNF462	13.868	3.247	2.36E-05
CUBN	-10.882	3.096	4.84E-04	ERAP2	12.198	3.284	2.28E-04	ERAP2	12.147	3.291	2.51E-04
ZNF462	14.350	4.094	5.01E-04	SIPA1L2	-10.933	3.107	4.76E-04	SIPA1L2	-10.950	3.108	4.68E-04
DENND4B	-8.466	2.521	8.50E-04	CACNA1G	-14.152	4.079	5.71E-04	CACNA1G	-14.186	4.077	5.49E-04
FBRSL1	-14.086	4.469	1.73E-03	TEKT1	18.336	5.330	6.35E-04	PGC	-11.606	3.352	5.85E-04
CGN	16.792	5.369	1.88E-03	KIAA0319	-14.581	4.257	6.69E-04	KIAA0319	-14.705	4.263	6.13E-04
DLGAP2	-17.359	5.727	2.58E-03	DENND4B	-8.544	2.518	7.49E-04	DENND4B	-8.683	2.520	6.22E-04
KIAA0319	-17.037	5.750	3.20E-03	ZNF462	12.863	3.800	7.73E-04	TEKT1	18.344	5.331	6.32E-04
PCDH15	8.977	3.102	3.99E-03	PCDH15	9.455	2.850	9.81E-04	PCDH15	7.547	2.238	8.09E-04
GLDC	11.487	4.105	5.36E-03	ENPP7	-10.898	3.329	1.14E-03	ENPP7	-10.904	3.330	1.14E-03
SON	11.820	4.305	6.27E-03	FBP2	12.389	3.797	1.19E-03	RNMTL1	-5.242	1.607	1.19E-03
CCHCR1	12.667	4.705	7.35E-03	SCYL1	-14.874	4.683	1.59E-03	FBP2	12.390	3.800	1.20E-03
PRKRIR	10.927	4.063	7.41E-03	ANXA11	-11.045	3.551	1.99E-03	SCYL1	-14.887	4.684	1.58E-03
SYNPO2	13.327	4.984	7.76E-03	NCF2	12.495	4.096	2.42E-03	ANXA11	-11.122	3.549	1.84E-03
MIOX	15.318	5.770	8.21E-03	GRTP1	14.252	4.753	2.86E-03	NLRP1	9.583	3.064	1.88E-03

Linear burden regression (uncorrected residual)

D

MAF≤0.1% (N=1679)				MAF≤1% (N=4482)				MAF≤2% (N=5305)			
Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*	Gene	*B*	*SE*	*p*
CUBN	-10.637	2.961	3.64E-04	CUBN	-8.335	1.900	1.43E-05	ZNF462	12.793	3.113	4.70E-05
CACNA1G	-15.179	4.250	3.93E-04	ERAP2	11.714	3.144	2.18E-04	ERAP2	11.709	3.151	2.27E-04
DENND4B	-8.520	2.410	4.48E-04	DENND4B	-8.591	2.408	3.97E-04	DENND4B	-8.688	2.410	3.46E-04
FBRSL1	-14.134	4.273	1.01E-03	SIPA1L2	-10.551	2.974	4.29E-04	SIPA1L2	-10.555	2.975	4.28E-04
CGN	16.511	5.135	1.40E-03	CACNA1G	-13.418	3.906	6.46E-04	RNMTL1	-5.383	1.536	5.03E-04
ZNF462	12.557	3.926	1.48E-03	KIAA0319	-13.940	4.076	6.81E-04	PGC	-11.078	3.209	6.09E-04
GLDC	12.210	3.921	1.96E-03	FBP2	12.253	3.633	8.07E-04	CACNA1G	-13.460	3.904	6.18E-04
PCDH15	8.905	2.967	2.84E-03	PCDH15	8.703	2.731	1.54E-03	KIAA0319	-14.014	4.082	6.51E-04
DLGAP2	-16.345	5.483	3.03E-03	NCF2	12.475	3.918	1.55E-03	FBP2	12.283	3.636	7.91E-04
KIAA0319	-15.840	5.506	4.20E-03	ENPP7	-10.079	3.190	1.69E-03	PCDH15	7.196	2.143	8.52E-04
SON	11.707	4.117	4.66E-03	ANXA11	-10.624	3.400	1.89E-03	CUBN	-4.595	1.423	1.33E-03
TMC6	-14.212	5.120	5.74E-03	ZNF462	11.385	3.645	1.90E-03	NCF2	12.542	3.914	1.45E-03
SYNPO2	13.175	4.767	5.94E-03	SCYL1	-13.897	4.486	2.07E-03	ENPP7	-10.074	3.191	1.70E-03
MIOX	15.093	5.520	6.49E-03	FAN1	-6.951	2.265	2.28E-03	ANXA11	-10.680	3.398	1.78E-03
PRKRIR	10.364	3.889	7.96E-03	SLC38A2	14.133	4.791	3.34E-03	NLRP1	9.134	2.934	1.97E-03

Linear burden regression (corrected residual)

E Cor (SKAT-O)		Cor (SKAT)		Uncor (SKAT-O)		Uncor (SKAT)		Logistic (SKAT-O)		Logistic (SKAT)	
Gene	*p*	Gene	*p*	Gene	*p*	Gene	*p*	Gene	*p*	Gene	*p*
CUBN	2.41E-05	DENND4B	7.60E-04	NOP14	1.37E-05	**NOP14**	**5.19E-06**	NOP14	1.83E-05	**NOP14**	**4.17E-06**
ERAP2	5.44E-04	CRAMP1L	1.30E-03	CUBN	2.12E-05	TEKT1	7.04E-04	NUP210L	6.18E-05	TEKT1	2.98E-04
DENND4B	9.63E-04	SLC22A14	2.15E-03	ERAP2	5.46E-04	DENND4B	1.41E-03	CUBN	7.55E-05	ST7L	1.19E-03
FBP2	1.36E-03	ST7L	2.89E-03	TEKT1	7.76E-04	CRAMP1L	2.26E-03	ERAP2	4.61E-04	DENND4B	1.42E-03
KIAA0319	1.39E-03	MUT	3.58E-03	CACNA1G	1.30E-03	NRIP3	2.67E-03	KIAA0319	1.03E-03	FBP2	1.91E-03
CACNA1G	1.48E-03	NRIP3	3.84E-03	KIAA0319	1.33E-03	FBP2	2.71E-03	ZNF462	1.14E-03	NRIP3	2.55E-03
SIPA1L2	1.66E-03	NOP14	3.97E-03	FBP2	1.62E-03	MUT	3.67E-03	MAMDC2	1.81E-03	PRKRIR	3.18E-03
ANXA11	2.24E-03	FGL1	4.23E-03	SIPA1L2	1.64E-03	GNA15	3.90E-03	CACNA1G	2.01E-03	CDC20B	4.05E-03
CRAMP1L	2.37E-03	FBP2	4.41E-03	ZNF462	1.71E-03	ST7L	4.05E-03	C9	2.35E-03	CRAMP1L	4.09E-03
NCF2	2.71E-03	HGFAC	4.46E-03	DENND4B	1.83E-03	FGL1	4.10E-03	NDOR1	2.38E-03	GNA15	4.94E-03
ENPP7	2.86E-03	UNC5B	4.66E-03	ENPP7	1.93E-03	SLC38A2	4.81E-03	GRTP1	2.80E-03	MKI67	5.10E-03
PCDH15	3.03E-03	SLC38A2	5.00E-03	PCDH15	2.07E-03	ERAP2	4.86E-03	SIPA1L2	2.93E-03	UNC5B	5.38E-03
SLC22A14	3.56E-03	OR2B11	5.02E-03	ANXA11	2.19E-03	SLC22A14	4.96E-03	TEKT1	2.98E-03	MUT	6.54E-03
FAN1	3.83E-03	UNK	5.12E-03	SCYL1	2.87E-03	PRKRIR	5.06E-03	DENND4B	3.10E-03	FGL1	6.86E-03
MUT	3.86E-03	TEKT1	5.20E-03	MUT	3.32E-03	UNC5B	5.22E-03	TGM3	3.12E-03	SHARPIN	7.28E-03

SKAT(-O) analyses

F	MAF≤0.1%			MAF≤1%			MAF≤2%		
	B	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*
EXO1	-0.647	0.883	4.64E-01	-0.598	0.348	8.52E-02	-0.600	0.347	8.43E-02
FAN1	0.721	0.717	3.14E-01	**1.052**	**0.414**	**1.10E-02**	**1.035**	**0.413**	**1.22E-02**
HTT	-0.439	0.669	5.11E-01	0.178	0.507	7.25E-01	0.186	0.507	7.13E-01
LIG1	0.244	0.779	7.54E-01	-0.110	0.514	8.31E-01	-0.098	0.514	8.49E-01
MLH1	-0.359	1.271	7.78E-01	-0.152	0.575	7.92E-01	-0.148	0.575	7.97E-01
MLH3	-0.386	0.659	5.58E-01	0.103	0.426	8.09E-01	0.117	0.426	7.83E-01
MSH3	-0.927	0.868	2.86E-01	-0.388	0.668	5.62E-01	-0.375	0.668	5.74E-01
OGG1	-0.647	0.883	4.64E-01	0.188	0.584	7.47E-01	0.192	0.581	7.41E-01
PMS1	NA	NA	NA	NA	NA	NA	NA	NA	NA
PMS2	NA	NA	NA	NA	NA	NA	-0.048	0.721	9.47E-01
RRM2B	NA	NA	NA	NA	NA	NA	NA	NA	NA
SYT9	NA	NA	NA	1.510	1.105	1.72E-01	1.481	1.106	1.80E-01
TCERG1	NA	NA	NA	-0.236	1.442	8.70E-01	-0.235	1.445	8.71E-01

Logistic burden regression, candidate genes

G	MAF≤0.1%			MAF≤1%			MAF≤2%		
	B	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*
EXO1	3.328	5.136	5.17E-01	**4.978**	**2.088**	**1.76E-02**	**4.993**	**2.088**	**1.72E-02**
FAN1	-4.034	4.527	3.73E-01	**-6.951**	**2.265**	**2.28E-03**	**-6.875**	**2.264**	**2.53E-03**
HTT	1.383	4.140	7.38E-01	-0.934	3.170	7.68E-01	-0.869	3.167	7.84E-01
LIG1	-2.335	5.139	6.50E-01	1.138	3.329	7.33E-01	1.084	3.329	7.45E-01
MLH1	3.799	6.095	5.33E-01	1.558	3.531	6.59E-01	1.570	3.531	6.57E-01
MLH3	2.945	4.308	4.95E-01	-0.117	2.743	9.66E-01	-0.134	2.747	9.61E-01
MSH3	1.558	4.576	7.34E-01	1.306	3.799	7.31E-01	1.306	3.800	7.31E-01
OGG1	**-24.224**	**9.566**	**1.17E-02**	-2.138	3.841	5.78E-01	-2.263	3.828	5.55E-01
PMS1	4.770	6.846	4.86E-01	**11.651**	**5.568**	**3.69E-02**	**11.532**	**5.565**	**3.88E-02**
PMS2	-14.283	7.778	6.69E-02	-14.379	7.782	6.53E-02	-0.630	4.531	8.90E-01
RRM2B	NA	NA	NA	NA	NA	NA	NA	NA	NA
SYT9	NA	NA	NA	-5.629	5.538	3.10E-01	-5.558	5.547	3.17E-01
TCERG1	14.430	13.514	2.86E-01	-0.976	9.615	9.19E-01	-0.940	9.617	9.22E-01

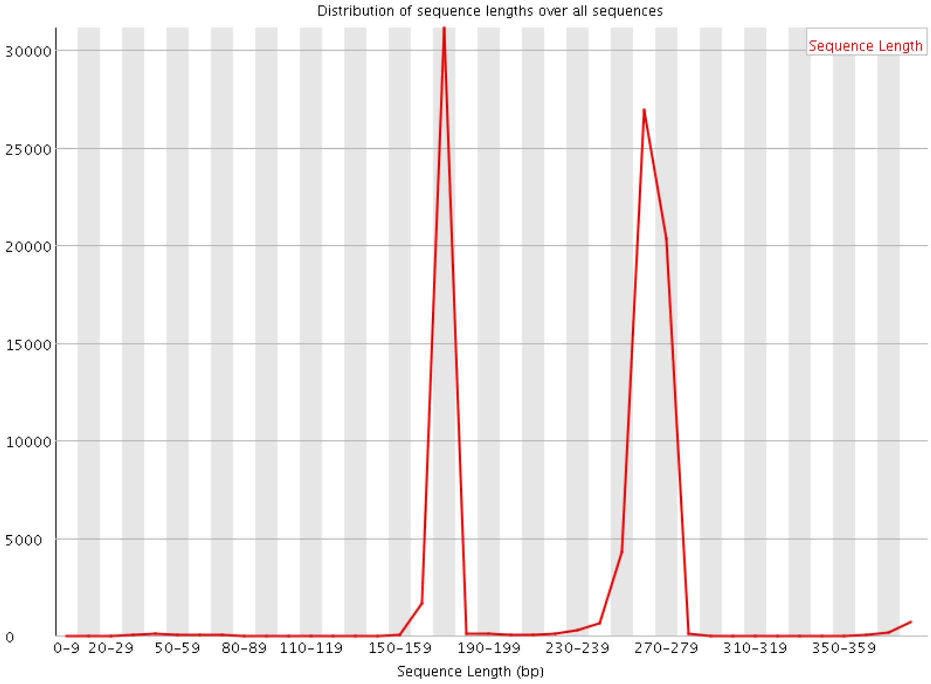
Linear burden regression (corrected residual), candidate genes

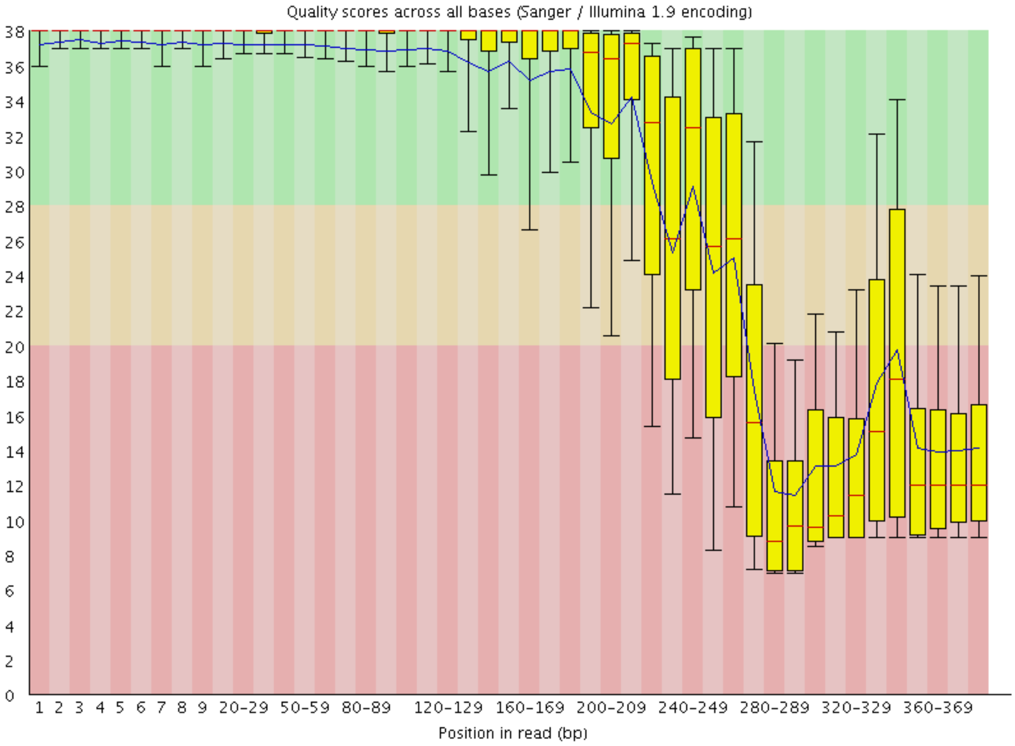
H	Cor (SKAT-O)	Uncor (SKAT-O)	Cor (SKAT)	Uncor (SKAT)	Logistic (SKAT-O)	Logistic (SKAT)
	p	*p*	*p*	*p*	*p*	*p*
EXO1	**3.05E-02**	6.77E-02	**4.74E-02**	6.85E-02	9.54E-02	7.24E-02
FAN1	**3.83E-03**	**5.75E-03**	**5.90E-03**	**8.41E-03**	**8.95E-03**	**1.24E-02**
HTT	8.99E-01	8.75E-01	7.61E-01	8.87E-01	1.00E+00	8.96E-01
LIG1	**4.74E-02**	**3.37E-02**	**2.70E-02**	**1.91E-02**	**3.35E-02**	**1.93E-02**
MLH1	8.34E-01	6.84E-01	8.32E-01	7.99E-01	7.36E-01	7.13E-01
MLH3	1.00E+00	8.64E-01	8.25E-01	6.78E-01	8.60E-01	6.71E-01
MSH3	8.42E-01	8.99E-01	6.40E-01	7.19E-01	8.63E-01	6.72E-01
OGG1	1.22E-01	1.18E-01	7.74E-02	7.50E-02	1.31E-01	8.40E-02
PMS1	5.62E-02	6.48E-02	6.48E-02	7.40E-02	**1.69E-02**	**2.30E-02**
PMS2	5.39E-02	**3.25E-02**	4.20E-01	3.07E-01	**4.40E-02**	3.69E-01
RRM2B	NA	NA	NA	NA	NA	NA
SYT9	3.10E-01	3.24E-01	3.10E-01	3.24E-01	3.54E-01	3.54E-01
TCERG1	3.74E-01	4.05E-01	2.73E-01	2.98E-01	4.45E-01	3.31E-01

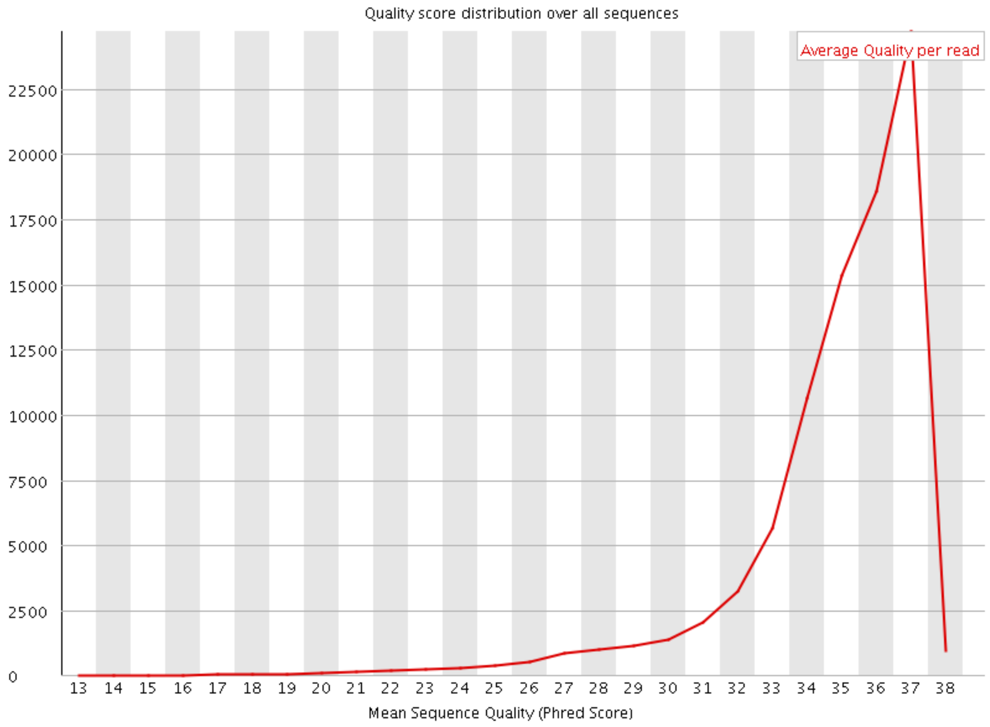
SKAT(-O), candidate genes

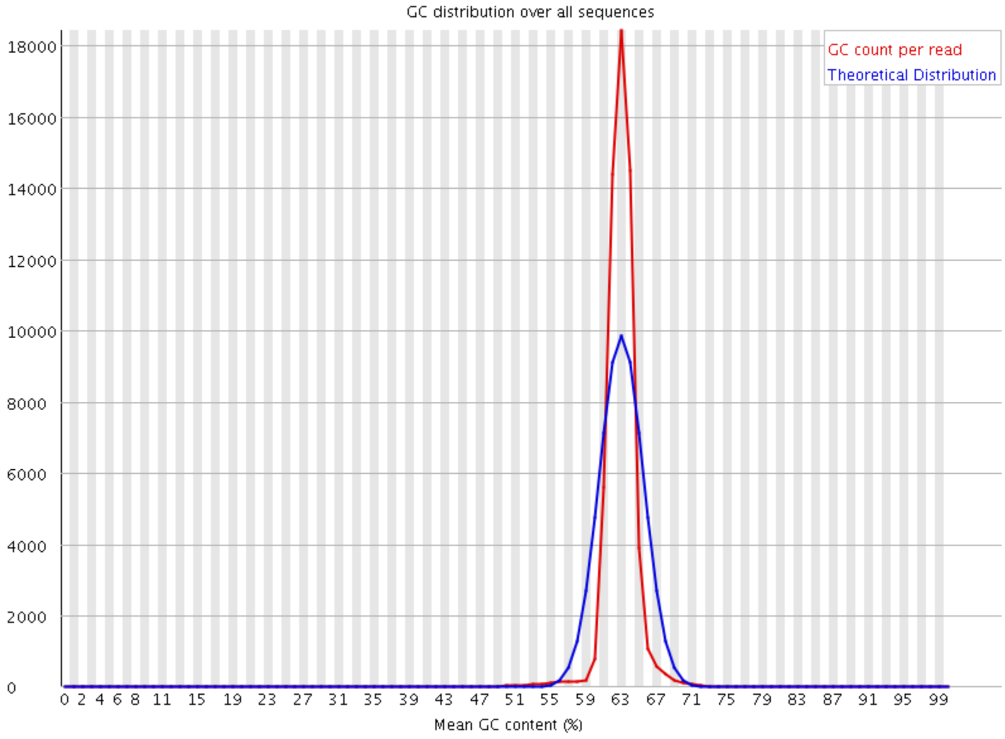
Appendix 16 – FastQC QC metrics for MiSeq

Shown are select FastQC QC metrics: sequence length, per-base sequence quality, mean sequence quality and GC content per base (this relates to the MiSeq sequencing from chapter 5). These data are taken from E86, the longest pure CAG length measured (wildtype: $(CAG)_{17}$**CAA**CAGCCGCCA$(CCG)_{10}(CCT)_2$; expanded: $(CAG)_{52}$**CAA**CAGCCGCCA$(CCG)_7(CCT)_2$. Note in the per-base plot, only lengths up to ~270 should be considered as indicated by the sequence length plot.



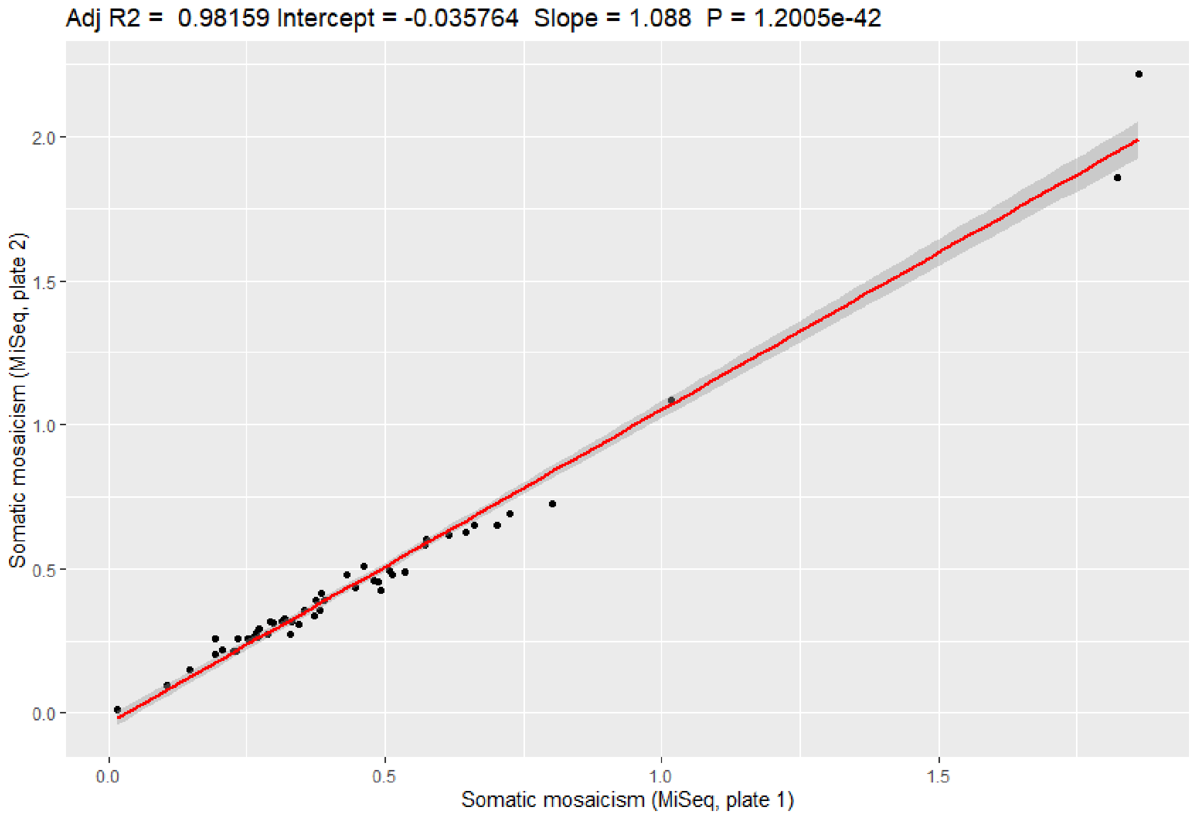



Quality score distribution over all sequences
Average Quality per read
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Mean Sequence Quality (Phred Score)


GC distribution over all sequences
GC count per read
Theoretical Distribution
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

Appendix 17 – Inter-run consistency of MiSeq

Shown are MiSeq somatic mosaicism measures for N=49 individuals for whom blood DNA was available (see chapter 5; 5.2). These DNA were run on two separate plates to investigate batch effects. The plot for CAG length determined for both plates is not shown as these were identical between plates (*i.e.* R^2=1, *p*=0).



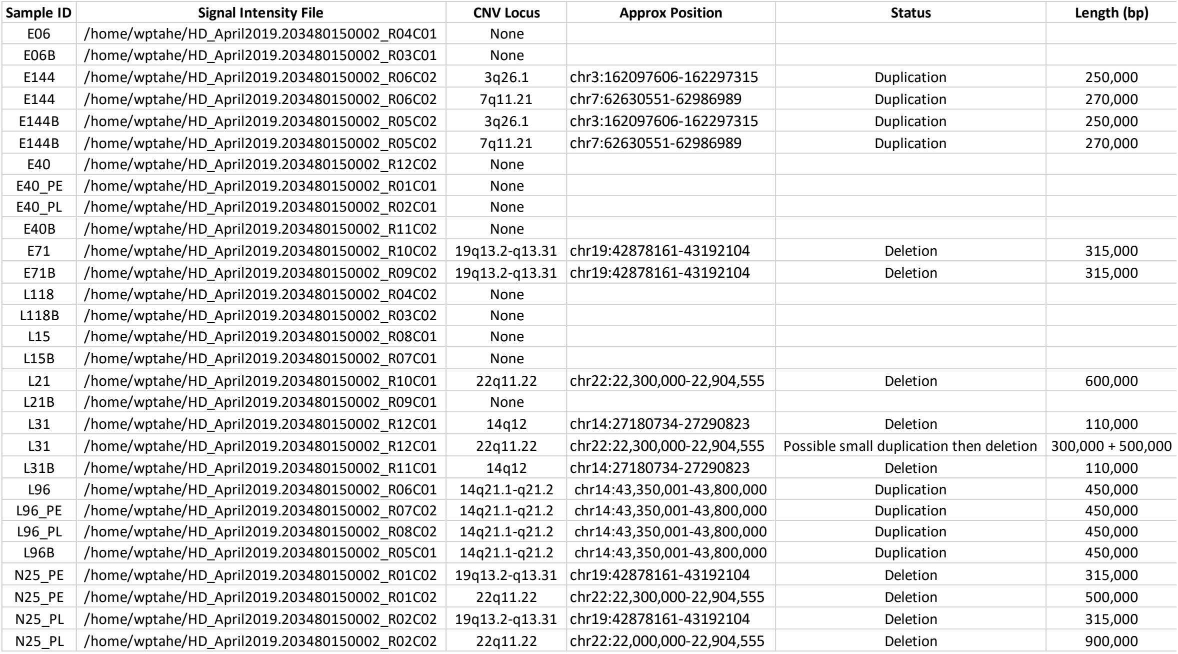
Appendix 18 – Alternative GLMs for allele structure on residual HD age at motor onset

Shown are generalised linear models (GLMs) for different structural features of HTT in both expanded and wild-type alleles, regressing uncorrected (polyglutamine-2) and corrected (pure CAG length) age at motor onset residuals on various covariates. Individuals previously quality controlled from 4.3.6 are used; N=483 as two individuals failed MiSeq sequencing. This is a sister Table to Table 5.4 in 5.3.2 where the number of interruptions are coded on a scale (0-3); here alleles are either loss-of-interruption (LOI), canonical or possess additional interruptions (INT) (see 2.9.4). Significant *p* values are emboldened. EXP: Expanded allele; WT: wild-type allele; PolyP: polyproline. B = unstandardised coefficient; β = standardised coefficient; SE = standard error.

	Uncorrected (N=483)				Corrected (N=483)			
	B	β	SE	*p*	*B*	β	SE	*p*
CAG tract interruption (EXP, INT)	16.383	0.238	2.976	**6.07E-08**	9.543	0.145	2.979	**1.45E-03**
CAG tract interruption (EXP, LOI)	-13.441	-0.195	4.243	**1.63E-03**	-5.693	-0.086	4.247	1.81E-01
CAG tract interruption (WT)	-2.791	-0.031	3.846	4.68E-01	-2.777	-0.032	3.850	4.71E-01
CCG tract interruption (EXP)	6.126	0.083	3.189	5.53E-02	6.086	0.086	3.193	5.72E-02
CCG tract interruption (WT)	5.978	0.074	5.689	2.94E-01	6.158	0.080	5.695	2.80E-01
PolyP length (EXP)	-0.470	-0.048	0.426	2.70E-01	-0.453	-0.048	0.426	2.88E-01
PolyP length (WT)	0.034	0.002	0.784	9.66E-01	0.081	0.006	0.785	9.18E-01

Appendix 19 – SNP genotyping array of lymphoblastoid cells

Genotyping and analysis were carried as described by the MRC core team (see methods 2.12) using an Illumina global screening array v2.0. These are the CNVs passing PennCNV QC (excluding CNVs <100kb or <10 total SNPs). Note these use the GRCh38 reference assembly. This relates to results in section 5.5.



Sample ID	Signal Intensity File	CNV Locus	Approx Position	Status	Length (bp)
E06	/home/wptahe/HD_April2019.203480150002_R04C01	None			
E06B	/home/wptahe/HD_April2019.203480150002_R03C01	None			
E144	/home/wptahe/HD_April2019.203480150002_R06C02	3q26.1	chr3:162097606-162297315	Duplication	250,000
E144	/home/wptahe/HD_April2019.203480150002_R06C02	7q11.21	chr7:62630551-62986989	Duplication	270,000
E144B	/home/wptahe/HD_April2019.203480150002_R05C02	3q26.1	chr3:162097606-162297315	Duplication	250,000
E144B	/home/wptahe/HD_April2019.203480150002_R05C02	7q11.21	chr7:62630551-62986989	Duplication	270,000
E40	/home/wptahe/HD_April2019.203480150002_R12C02	None			
E40_PE	/home/wptahe/HD_April2019.203480150002_R01C01	None			
E40_PL	/home/wptahe/HD_April2019.203480150002_R02C01	None			
E40B	/home/wptahe/HD_April2019.203480150002_R11C02	None			
E71	/home/wptahe/HD_April2019.203480150002_R10C02	19q13.2-q13.31	chr19:42878161-43192104	Deletion	315,000
E71B	/home/wptahe/HD_April2019.203480150002_R09C02	19q13.2-q13.31	chr19:42878161-43192104	Deletion	315,000
L118	/home/wptahe/HD_April2019.203480150002_R04C02	None			
L118B	/home/wptahe/HD_April2019.203480150002_R03C02	None			
L15	/home/wptahe/HD_April2019.203480150002_R08C01	None			
L15B	/home/wptahe/HD_April2019.203480150002_R07C01	None			
L21	/home/wptahe/HD_April2019.203480150002_R10C01	22q11.22	chr22:22,300,000-22,904,555	Deletion	600,000
L21B	/home/wptahe/HD_April2019.203480150002_R09C01	None			
L31	/home/wptahe/HD_April2019.203480150002_R12C01	14q12	chr14:27180734-27290823	Deletion	110,000
L31	/home/wptahe/HD_April2019.203480150002_R12C01	22q11.22	chr22:22,300,000-22,904,555	Possible small duplication then deletion	300,000 + 500,000
L31B	/home/wptahe/HD_April2019.203480150002_R11C01	14q12	chr14:27180734-27290823	Deletion	110,000
L96	/home/wptahe/HD_April2019.203480150002_R06C01	14q21.1-q21.2	chr14:43,350,001-43,800,000	Duplication	450,000
L96_PE	/home/wptahe/HD_April2019.203480150002_R07C02	14q21.1-q21.2	chr14:43,350,001-43,800,000	Duplication	450,000
L96_PL	/home/wptahe/HD_April2019.203480150002_R08C02	14q21.1-q21.2	chr14:43,350,001-43,800,000	Duplication	450,000
L96B	/home/wptahe/HD_April2019.203480150002_R05C01	14q21.1-q21.2	chr14:43,350,001-43,800,000	Duplication	450,000
N25_PE	/home/wptahe/HD_April2019.203480150002_R01C02	19q13.2-q13.31	chr19:42878161-43192104	Deletion	315,000
N25_PE	/home/wptahe/HD_April2019.203480150002_R01C02	22q11.22	chr22:22,300,000-22,904,555	Deletion	500,000
N25_PL	/home/wptahe/HD_April2019.203480150002_R02C02	19q13.2-q13.31	chr19:42878161-43192104	Deletion	315,000
N25_PL	/home/wptahe/HD_April2019.203480150002_R02C02	22q11.22	chr22:22,000,000-22,904,555	Deletion	900,000

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526: 68–74.

Acharya S., Wilson T., Gradia S., Kane M.F., Guerrette S., Marsischky G.T., et al. (1996). hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. Proceedings of the National Academy of Sciences of the United States of America 93: 13629–13634.

Adegbuyiro A., Sedighi F., Pilkington A.W., Groover S., and Legleiter J. (2017). Proteins Containing Expanded Polyglutamine Tracts and Neurodegenerative Disease. Biochemistry 56: 1199–1217.

Adzhubei I.A., Schmidt S., Peshkin L., Ramensky V.E., Gerasimova A., Bork P., et al. (2010). A method and server for predicting damaging missense mutations. Nature Methods 7: 248–9.

Agarwala V., Flannick J., Sunyaev S., GoT2D Consortium, and Altshuler D. (2013). Evaluating empirical bounds on complex disease genetic architecture. Nature Genetics 45: 1418–27.

Airik R., Schueler M., Airik M., Cho J., Porath J.D., Mukherjee E., et al. (2016). A FANCD2/FANCI-Associated Nuclease 1-Knockout Model Develops Karyomegalic Interstitial Nephritis. Journal of the American Society of Nephrology 27: 3552–3559.

Albin R.L., Reiner A., Anderson K.D., Dure L.S., Handelin B., Balfour R., et al. (1992). Preferential loss of striato-external pallidal projection neurons in presymptomatic Huntington’s disease. Annals of Neurology 31: 425–430.

Albin R.L., Young A.B., and Penney J.B. (1989). The functional anatomy of basal ganglia disorders. Trends in Neurosciences 12: 366–75.

Alexander G.E. and Crutcher M.D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends in Neurosciences 13: 266–71.

Almaguer-Mederos L.E., Mesa J.M.L., González-Zaldívar Y., Almaguer-Gotay D., Cuello-Almarales D., Aguilera-Rodríguez R., et al. (2018). Factors associated with ATXN2 CAG/CAA repeat intergenerational instability in Spinocerebellar ataxia type 2. Clinical Genetics 94: 346–350.

Altshuler D., Daly M.J., and Lander E.S. (2008). Genetic mapping in human disease. Science 322: 881–8.

Amoli M.M., Carthy D., Platt H., and Ollier W.E.R. (2008). EBV Immortalization of human B lymphocytes separated from small volumes of cryo-preserved whole blood. International

Journal of Epidemiology 37 Suppl 1: i41-5.

Andrade M.A. and Bork P. (1995). HEAT repeats in the Huntington's disease protein. Nature Genetics 11: 115–6.

Andresen J.M., Gayán J., Cherny S.S., Brocklebank D., Alkorta-Aranburu G., Addis E.A., et al. (2007a). Replication of twelve association studies for Huntington's disease residual age of onset in large Venezuelan kindreds. Journal of Medical Genetics 44: 44–50.

Andresen J.M., Gayán J., Djoussé L., Roberts S., Brocklebank D., Cherny S.S., et al. (2007b). The relationship between CAG repeat length and age of onset differs for Huntington's disease patients with juvenile onset or adult onset. Annals of Human Genetics 71: 295–301.

Andrew S.E., Goldberg Y.P., Kremer B., Telenius H., Theilmann J., Adam S., et al. (1993). The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nature Genetics 4: 398–403.

Andrich J.E., Wobben M., Klotz P., Goetze O., and Saft C. (2009). Upper gastrointestinal findings in Huntington's disease: patients suffer but do not complain. Journal of Neural Transmission 116: 1607–1611.

Arango M., Holbert S., Zala D., Brouillet E., Pearson J., Régulier E., et al. (2006). CA150 expression delays striatal cell death in overexpression and knock-in conditions for mutant huntingtin neurotoxicity. The Journal of Neuroscience 26: 4649–59.

Ardui S., Race V., de Ravel T., Van Esch H., Devriendt K., Matthijs G., and Vermeesch J.R. (2018). Detecting AGG Interruptions in Females With a FMR1 Premutation by Long-Read Single-Molecule Sequencing: A 1 Year Clinical Experience. Frontiers in Genetics 9: 150.

Aretouli E. and Brandt J. (2010). Episodic memory in dementia: Characteristics of new learning that differentiate Alzheimer's, Huntington's, and Parkinson's diseases. Archives of Clinical Neuropsychology 25: 396–409.

Arning L., Kraus P.H., Saft C., Andrich J., and Epplen J.T. (2005). Age at onset of Huntington disease is not modulated by the R72P variation in TP53 and the R196K variation in the gene coding for the human caspase activated DNase (hCAD). BMC Medical Genetics 6: 35.

Arnulf I., Nielsen J., Lohmann E., Schieffer J., Wild E., Jennum P., et al. (2008). Rapid Eye Movement Sleep Disturbances in Huntington Disease. Archives of Neurology 65: 482.

Arran N., Craufurd D., and Simpson J. (2014). Illness perceptions, coping styles and psychological distress in adults with Huntington's disease. Psychology, Health & Medicine

19: 169–179.

Arrasate M., Mitra S., Schweitzer E.S., Segal M.R., and Finkbeiner S. (2004). Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. Nature 431: 805–10.

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25: 25–9.

Auer P.L. and Lettre G. (2015). Rare variant association studies: considerations, challenges and opportunities. Genome Medicine 7: 16.

Augood S.J., Faull R.L., Love D.R., and Emson P.C. (1996). Reduction in enkephalin and substance P messenger RNA in the striatum of early grade Huntington's disease: a detailed cellular in situ hybridization study. Neuroscience 72: 1023–36.

Van der Auwera G.A., Carneiro M.O., Hartl C., Poplin R., Del Angel G., Levy-Moonshine A., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics 43: 11.10.1-33.

Aylward E., Mills J., Liu D., Nopoulos P., Ross C.A., Pierson R., and Paulsen J.S. (2011). Association between Age and Striatal Volume Stratified by CAG Repeat Length in Prodromal Huntington Disease. PLOS Currents 3: RRN1235.

Aylward E.H., Anderson N.B., Bylsma F.W., Wagster M. V, Barta P.E., Sherr M., et al. (1998). Frontal lobe volume in patients with Huntington's disease. Neurology 50: 252–8.

Aylward E.H., Li Q., Stine O.C., Ranen N., Sherr M., Barta P.E., et al. (1997). Longitudinal change in basal ganglia volume in patients with Huntington's disease. Neurology 48: 394–9.

Aylward E.H., Sparks B.F., Field K.M., Yallapragada V., Shpritz B.D., Rosenblatt A., et al. (2004). Onset and rate of striatal atrophy in preclinical Huntington disease. Neurology 63: 66–72.

Aziz N.A., Anguelova G. V., Marinus J., Van Dijk J.G., and Roos R.A.C. (2010). Autonomic symptoms in patients and pre-manifest mutation carriers of Huntington's disease. European Journal of Neurology 17: 1068–1074.

Aziz N.A., van Belzen M.J., Coops I.D., Belfroid R.D.M., and Roos R.A.C. (2011). Parent-of-origin differences of mutant HTT CAG repeat instability in Huntington's disease. European Journal of Medical Genetics 54: e413–e418.

Aziz N.A., van der Burg J.M.M., Landwehrmeyer G.B., Brundin P., Stijnen T., EHDI Study Group, and Roos R.A.C. (2008). Weight loss in Huntington disease increases with higher

CAG repeat number. Neurology 71: 1506–13.

Aziz N.A., van der Burg J.M.M., Tabrizi S.J., and Landwehrmeyer G.B. (2018). Overlap between age-at-onset and disease-progression determinants in Huntington disease. Neurology 90: e2099–e2106.

Aziz N.A., Jurgens C.K., Landwehrmeyer G.B., EHDN Registry Study Group, van Roon-Mom W.M.C., van Ommen G.J.B., et al. (2009). Normal and mutant HTT interact to affect clinical severity and progression in Huntington disease. Neurology 73: 1280–5.

Baake V., Coppen E.M., van Duijn E., Dumas E.M., van den Bogaard S.J.A., Scahill R.I., et al. (2018). Apathy and atrophy of subcortical brain structures in Huntington's disease: A two-year follow-up study. NeuroImage. Clinical 19: 66–70.

Baake V., Reijntjes R.H.A.M., Dumas E.M., Thompson J.C., REGISTRY Investigators of the European Huntington's Disease Network, and Roos R.A.C. (2017). Cognitive decline in Huntington's disease expansion gene carriers. Cortex 95: 51–62.

Bachinski L.L., Czernuszewicz T., Ramagli L.S., Suominen T., Shriver M.D., Udd B., et al. (2009). Premutation allele pool in myotonic dystrophy type 2. Neurology 72: 490–7.

Bachoud-Lévi A.-C., Ferreira J., Massart R., Youssov K., Rosser A., Busse M., et al. (2019). International Guidelines for the Treatment of Huntington's Disease. Frontiers in Neurology 10: 710.

Backenroth D., He Z., Kiryluk K., Boeva V., Pethukova L., Khurana E., et al. (2018). FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. American Journal of Human Genetics 102: 920–942.

Baine F.K., Kay C., Ketelaar M.E., Collins J.A., Semaka A., Doty C.N., et al. (2013). Huntington disease in the South African population occurs on diverse and ethnically distinct genetic haplotypes. European Journal of Human Genetics 21: 1120–1127.

Baine F.K., Krause A., and Greenberg L.J. (2016). The Frequency of Huntington Disease and Huntington Disease-Like 2 in the South African Population. Neuroepidemiology 46: 198–202.

Bakhtiari M., Shleizer-Burko S., Gymrek M., Bansal V., and Bafna V. (2018). Targeted genotyping of variable number tandem repeats with adVNTR. Genome Research 28: 1709–1719.

Bañez-Coronel M., Ayhan F., Tarabochia A.D., Zu T., Perez B.A., Tusi S.K., et al. (2015). RAN Translation in Huntington Disease. Neuron 88: 667–77.

Bañez-Coronel M., Porta S., Kagerbauer B., Mateu-Huertas E., Pantano L., Ferrer I., et al. (2012). A pathogenic mechanism in Huntington's disease involves small CAG-repeated RNAs with neurotoxic activity. PLOS Genetics 8: e1002481.

Bansal V., Mitjans M., Burik C.A.P., Linnér R.K., Okbay A., Rietveld C.A., et al. (2018). Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia. Nature Communications 9: 3078.

Barbaro B.A., Lukacsovich T., Agrawal N., Burke J., Bornemann D.J., Purcell J.M., et al. (2015). Comparative study of naturally occurring huntingtin fragments in Drosophila points to exon 1 as the most pathogenic species in Huntington's disease. Human Molecular Genetics 24: 913–25.

Barnett I.J., Lee S., and Lin X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genetic Epidemiology 37: 142–51.

Bastarache L., Hughey J.J., Hebbring S., Marlo J., Zhao W., Ho W.T., et al. (2018). Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. Science 359: 1233–1239.

Bates G.P., Dorsey R., Gusella J.F., Hayden M.R., Kay C., Leavitt B.R., et al. (2015). Huntington disease. Nature Reviews Disease Primers 1: 15005.

Bateup H.S., Santini E., Shen W., Birnbaum S., Valjent E., Surmeier D.J., et al. (2010). Distinct subclasses of medium spiny neurons differentially regulate striatal motor behaviors. Proceedings of the National Academy of Sciences of the United States of America 107: 14845–50.

Bäuerlein F.J.B., Saha I., Mishra A., Kalemanov M., Martínez-Sánchez A., Klein R., et al. (2017). In Situ Architecture and Cellular Interactions of PolyQ Inclusions. Cell 171: 179-187.e10.

Beagan K., Armstrong R.L., Witsell A., Roy U., Renedo N., Baker A.E., et al. (2017). Drosophila DNA polymerase theta utilizes both helicase-like and polymerase domains during microhomology-mediated end joining and interstrand crosslink repair. PLOS Genetics 13: e1006813.

Bellosta Diago E., Pérez Pérez J., Santos Lasaosa S., Viloria Alebesque A., Martínez Horta S., Kulisevsky J., and López del Val J. (2017). Circadian rhythm and autonomic dysfunction in presymptomatic and early Huntington's disease. Parkinsonism & Related Disorders 44: 95–100.

Bertier G., Hétu M., and Joly Y. (2016). Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. BMC Medical Genomics 9: 52.

Bettencourt C., Hensman-Moss D., Flower M., Wiethoff S., Brice A., Goizet C., et al. (2016). DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Annals of Neurology 79: 983–990.

Biglan K.M., Ross C.A., Langbehn D.R., Aylward E.H., Stout J.C., Queller S., et al. (2009). Motor abnormalities in premanifest persons with Huntington's disease: the PREDICT-HD study. Movement Disorders 24: 1763–72.

Bis J.C., Jian X., Kunkle B.W., Chen Y., Hamilton-Nelson K.L., Bush W.S., et al. (2018). Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Molecular Psychiatry [ahead of print].

Bjelland I., Dahl A.A., Haug T.T., and Neckelmann D. (2002). The validity of the Hospital Anxiety and Depression Scale. An updated literature review. Journal of Psychosomatic Research 52: 69–77.

Blekher T., Johnson S.A., Marshall J., White K., Hui S., Weaver M., et al. (2006). Saccades in presymptomatic and early stages of Huntington disease. Neurology 67: 394–9.

Blekher T.M., Yee R.D., Kirkwood S.C., Hake A.M., Stout J.C., Weaver M.R., and Foroud T.M. (2004). Oculomotor control in asymptomatic and recently diagnosed individuals with the genetic marker for Huntington's disease. Vision Research 44: 2729–36.

Boland C.R. and Goel A. (2010). Microsatellite instability in colorectal cancer. Gastroenterology 138: 2073-2087.e3.

Bomba L., Walter K., and Soranzo N. (2017). The impact of rare and low-frequency genetic variants in common disease. Genome Biology 18: 77.

Bourn R.L., de Biase I., Pinto R.M., Sandi C., Al-Mahdawi S., Pook M.A., and Bidichandani S.I. (2012). Pms2 suppresses large expansions of the (GAA·TTC)n sequence in neuronal tissues. PLOS ONE 7: e47085.

Bouwens J.A., van Duijn E., van der Mast R.C., Roos R.A.C., and Giltay E.J. (2015). Irritability in a Prospective Cohort of Huntington's Disease Mutation Carriers. The Journal of Neuropsychiatry and Clinical Neurosciences 27: 206–212.

Bradford J., Shin J.-Y., Roberts M., Wang C.-E., Li X.-J., and Li S. (2009). Expression of mutant huntingtin in mouse brain astrocytes causes age-dependent neurological symptoms. Proceedings of the National Academy of Sciences of the United States of America 106:

22480–5.

Braida C., Stefanatos R.K.A., Adam B., Mahajan N., Smeets H.J.M., Niel F., et al. (2010). Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. Human Molecular Genetics 19: 1399–412.

Brais B., Bouchard J.-P., Xie Y.-G., Rochefort D.L., Chrétien N., Tomé F.M.S., et al. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nature Genetics 18: 164–167.

Braisch U., Hay B., Muche R., Rothenbacher D., Landwehrmeyer B.G., Long J.D., et al. (2017). Identification of extreme motor phenotypes in Huntington's disease. American Journal of Medical Genetics, Part B 174: 283–294.

Braisch U., Muche R., Rothenbacher D., Landwehrmeyer G.B., Long J.D., Orth M., and REGISTRY Investigators of the European Huntington's Disease Network and COHORT Investigators of the Huntington Study Group (2019). Identification of symbol digit modality test score extremes in Huntington's disease. American Journal of Medical Genetics, Part B 180: 232–245.

Brandt J., Bylsma F.W., Gross R., Stine O.C., Ranen N., and Ross C.A. (1996). Trinucleotide repeat length and clinical progression in Huntington's disease. Neurology 46: 527–31.

Brinkman R.R., Mezei M.M., Theilmann J., Almqvist E., and Hayden M.R. (1997). The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. American Journal of Human Genetics 60: 1202–10.

Brocklebank D., Gayán J., Andresen J.M., Roberts S.A., Young A.B., Snodgrass S.R., et al. (2009). Repeat instability in the 27-39 CAG range of the HD gene in the Venezuelan kindreds: Counseling implications. American Journal of Medical Genetics, Part B 150B: 425–9.

van den Broek W.J.A.A., Nelen M.R., Wansink D.G., Coerwinkel M.M., te Riele H., Groenen P.J.T.A., and Wieringa B. (2002). Somatic expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. Human Molecular Genetics 11: 191–198.

Bromet E., Andrade L.H., Hwang I., Sampson N.A., Alonso J., de Girolamo G., et al. (2011). Cross-national epidemiology of DSM-IV major depressive episode. BMC Medicine 9: 90.

Bronner I.F., Quail M.A., Turner D.J., and Swerdlow H. (2014). Improved Protocols for

Illumina Sequencing. Current Protocols in Human Genetics 80: 18.2.1-42.

Brook J.D., McCurrach M.E., Harley H.G., Buckler A.J., Church D., Aburatani H., et al. (1992). Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell 68: 799–808.

Brotherton A., Campos L., Rowell A., Zoia V., Simpson S.A., and Rae D. (2012). Nutritional management of individuals with Huntington’s disease: nutritional guidelines. Neurodegenerative Disease Management 2: 33–43.

Brown J.S., O’Carrigan B., Jackson S.P., and Yap T.A. (2017). Targeting DNA Repair in Cancer: Beyond PARP Inhibitors. Cancer Discovery 7: 20–37.

Brown L.Y. and Brown S.A. (2004). Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends in Genetics 20: 51–8.

Bruse S., Moreau M., Bromberg Y., Jang J.-H., Wang N., Ha H., et al. (2016). Whole exome sequencing identifies novel candidate genes that modify chronic obstructive pulmonary disease susceptibility. Human Genomics 10: 1.

Budworth H., Harris F.R., Williams P., Lee D.Y., Holt A., Pahnke J., et al. (2015). Suppression of Somatic Expansion Delays the Onset of Pathophysiology in a Mouse Model of Huntington’s Disease. PLOS Genetics 11: 1–22.

Buniello A., MacArthur J.A.L., Cerezo M., Harris L.W., Hayhurst J., Malangone C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research 47: D1005–D1012.

Bunner K.D. and Rebec G. V (2016). Corticostriatal Dysfunction in Huntington’s Disease: The Basics. Frontiers in Human Neuroscience 10: 317.

Burdova K., Mihaljevic B., Sturzenegger A., Chappidi N., and Janscak P. (2015). The Mismatch-Binding Factor MutSβ Can Mediate ATR Activation in Response to DNA Double-Strand Breaks. Molecular Cell 59: 603–14.

Busse M.E., Wiles C.M., and Rosser A.E. (2009). Mobility and falls in people with Huntington’s disease. Journal of Neurology, Neurosurgery, and Psychiatry 80: 88–90.

Campuzano V., Montermini L., Molto M.D., Pianese L., Cossee M., Cavalcanti F., et al. (1996). Friedreich’s Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion. Science 271: 1423–1427.

Cannavo E., Gerrits B., Marra G., Schlapbach R., and Jiricny J. (2007). Characterization of the interactome of the human MutL homologues MLH1, PMS1, and PMS2. The Journal of Biological Chemistry 282: 2976–86.

Cannavo E., Marra G., Sabates-Bellver J., Menigatti M., Lipkin S.M., Fischer F., et al. (2005). Expression of the MutL homologue hMLH3 in human cells and its role in DNA mismatch repair. Cancer Research 65: 10759–66.

Cannella M., Gellera C., Maglione V., Giallonardo P., Cislaghi G., Muglia M., et al. (2004). The gender effect in juvenile Huntington disease patients of Italian origin. American Journal of Medical Genetics Part B 125B: 92–98.

Cannella M., Maglione V., Martino T., Ragona G., Frati L., Li G.-M., and Squitieri F. (2009). DNA instability in replicating Huntington's disease lymphoblasts. BMC Medical Genetics 10: 11.

Cardoso F. (2017). Nonmotor Symptoms in Huntington Disease. International Review of Neurobiology 134: 1397–1408.

de Carvalho-Siqueira G.Q., Ananina G., de Souza B.B., Borges M.G., Ito M.T., da Silva-Costa S.M., et al. (2019). Highlight article: Whole-exome sequencing indicates FLG2 variant associated with leg ulcers in Brazilian sickle cell anemia patients. Experimental Biology and Medicine 244: 932–939.

Caviston J.P., Ross J.L., Antony S.M., Tokito M., and Holzbaur E.L.F. (2007). Huntingtin facilitates dynein/dynactin-mediated vesicle transport. Proceedings of the National Academy of Sciences of the United States of America 104: 10045–10050.

Caviston J.P., Zajac A.L., Tokito M., and Holzbaur E.L.F. (2011). Huntingtin coordinates the dynein-mediated dynamic positioning of endosomes and lysosomes. Molecular Biology of the Cell 22: 478–492.

Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., and Lee J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4: 7.

Chao M.J., Kim K.-H., Shin J.W., Lucente D., Wheeler V.C., Li H., et al. (2018). Population-specific genetic modification of Huntington's disease in Venezuela. PLOS Genetics 14: e1007274.

Charles P., Camuzat A., Benammar N., Sellal F., Destée A., Bonnet A.-M., et al. (2007). Are interrupted SCA2 CAG repeat expansions responsible for parkinsonism? Neurology 69: 1970–5.

Chattopadhyay B., Baksi K., Mukhopadhyay S., and Bhattacharyya N.P. (2005). Modulation of age at onset of Huntington disease patients by variations in TP53 and human caspase activated DNase (hCAD) genes. Neuroscience Letters 374: 81–86.

Chaudhury I., Stroik D.R., and Sobeck A. (2014). FANCD2-controlled chromatin access of the Fanconi-associated nuclease FAN1 is crucial for the recovery of stalled replication forks. Molecular and Cellular Biology 34: 3939–54.

Chen H., Huffman J.E., Brody J.A., Wang C., Lee S., Li Z., et al. (2019). Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. American Journal of Human Genetics 104: 260–274.

Chen L., Hadd A., Sah S., Filipovic-Sadic S., Krosting J., Sekinger E., et al. (2010). An information-rich CGG repeat primed PCR that detects the full range of fragile X expanded alleles and minimizes the need for southern blot analysis. The Journal of Molecular Diagnostics 12: 589–600.

Chen Y.-Y. and Lai C.-H. (2010). Nationwide Population-Based Epidemiologic Study of Huntington’s Disease in Taiwan. Neuroepidemiology 35: 250–254.

Chiang M.-C., Chen H.-M., Lee Y.-H., Chang H.-H., Wu Y.-C., Soong B.-W., et al. (2007). Dysregulation of C/EBPα by mutant Huntingtin causes the urea cycle deficiency in Huntington’s disease. Human Molecular Genetics 16: 483–498.

Choi M., Scholl U.I., Ji W., Liu T., Tikhonova I.R., Zumbo P., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 106: 19096–101.

Chong S.S., McCall A.E., Cota J., Subramony S.H., Orr H.T., Hughes M.R., and Zoghbi H.Y. (1995). Gametic and somatic tissue-specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. Nature Genetics 10: 344–50.

Choudhry S., Mukerji M., Srivastava A.K., Jain S., and Brahmachari S.K. (2001). CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. Human Molecular Genetics 10: 2437–46.

Chung M.Y., Ranum L.P., Duvick L.A., Servadio A., Zoghbi H.Y., and Orr H.T. (1993). Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nature Genetics 5: 254–8.

Ciarmiello A., Cannella M., Lastoria S., Simonelli M., Frati L., Rubinsztein D.C., and Squitieri F. (2006). Brain white-matter volume loss and glucose hypometabolism precede the clinical symptoms of Huntington’s disease. Journal of Nuclear Medicine 47: 215–22.

Cingolani P., Platts A., Wang L.L., Coon M., Nguyen T., Wang L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6: 80–92.

Ciosi M., Cumming S.A., Mubarak A., Symeonidi E., Herzyk P., McGuinness D., et al.

(2018). Library preparation and MiSeq sequencing for the genotyping-by-sequencing of the Huntington disease HTT exon one trinucleotide repeat and the quantification of somatic mosaicism. Protocol Exchange.

Ciosi M., Maxwell A., Cumming S.A., Hensman Moss D., Alshammari A.M., Flower M., et al. (2019). A Genetic Association Study of Glutamine-Encoding DNA Sequence Structures, Somatic CAG Expansion, and DNA Repair Gene Variants, with Huntington Disease Clinical Outcomes. EBioMedicine [in press].

Cirulli E.T. and Goldstein D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics 11: 415–25.

Cleary J.D., Pattamatta A., and Ranum L.P.W. (2018). Repeat-associated non-ATG (RAN) translation. Journal of Biological Chemistry 293: 16127–16141.

Cloud L.J., Rosenblatt A., Margolis R.L., Ross C.A., Pillai J.A., Corey-Bloom J., et al. (2012). Seizures in juvenile Huntington's disease: Frequency and characterization in a multicenter cohort. Movement Disorders 27: 1797–1800.

Coffee B., Zhang F., Warren S.T., and Reines D. (1999). Acetylated histones are associated with FMR1 in normal but not fragile X-syndrome cells. Nature Genetics 22: 98–101.

Coleman A., Fountain J.W., Nobori T., Olopade O.I., Robertson G., Housman D.E., and Lugo T.G. (1994). Distinct deletions of chromosome 9p associated with melanoma versus glioma, lung cancer, and leukemia. Cancer Research 54: 344–8.

Colin E., Zala D., Liot G., Rangone H., Borrell-Pagès M., Li X.-J., et al. (2008). Huntingtin phosphorylation acts as a molecular switch for anterograde/retrograde transport in neurons. The EMBO Journal 27: 2124–2134.

Cortese A., Simone R., Sullivan R., Vandrovcova J., Tariq H., Yau W.Y., et al. (2019). Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nature Genetics 51: 649–658.

Craufurd D., Thompson J.C., and Snowden J.S. (2001). Behavioral changes in Huntington Disease. Neuropsychiatry, Neuropsychology, and Behavioral Neurology 14: 219–26.

Crespan E., Hübscher U., and Maga G. (2015). Expansion of CAG triplet repeats by human DNA polymerases λ and β in vitro, is regulated by flap endonuclease 1 and DNA ligase 1. DNA Repair 29: 101–11.

Critchley B.J., Isalan M., and Mielcarek M. (2018). Neuro-Cardio Mechanisms in Huntington's Disease and Other Neurodegenerative Disorders. Frontiers in Physiology 9: 559.

Croce K.R. and Yamamoto A. (2019). A role for autophagy in Huntington's disease. Neurobiology of Disease 122: 16–22.

Cui G., Jun S.B., Jin X., Pham M.D., Vogel S.S., Lovinger D.M., and Costa R.M. (2013). Concurrent activation of striatal direct and indirect pathways during action initiation. Nature 494: 238–42.

Cumming S.A., Hamilton M.J., Robb Y., Gregory H., McWilliam C., Cooper A., et al. (2018). De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. European Journal of Medical Genetics 26: 1635–1647.

Cunningham F., Achuthan P., Akanni W., Allen J., Amode M.R., Armean I.M., et al. (2019). Ensembl 2019. Nucleic Acids Research 47: D745–D751.

Dale M., Maltby J., Shimozaki S., Cramp R., Rickards H., and REGISTRY Investigators of the European Huntington's Disease Network (2016). Disease stage, but not sex, predicts depression and psychological distress in Huntington's disease: A European population study. Journal of Psychosomatic Research 80: 17–22.

Dashnow H., Lek M., Phipson B., Halman A., Sadedin S., Lonsdale A., et al. (2018). STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biology 19: 121.

David G., Dürr A., Stevanin G., Cancel G., Abbas N., Benomar A., et al. (1998). Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7). Human Molecular Genetics 7: 165–70.

Davies S.W., Turmaine M., Cozens B.A., DiFiglia M., Sharp A.H., Ross C.A., et al. (1997). Formation of Neuronal Intranuclear Inclusions Underlies the Neurological Dysfunction in Mice Transgenic for the HD Mutation. Cell 90: 537–548.

Debacker K., Winnepenninckx B., Longman C., Colgan J., Tolmie J., Murray R., et al. (2007). The molecular basis of the folate-sensitive fragile site FRA11A at 11q13. Cytogenetic and Genome Research 119: 9–14.

Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542: 433–438.

DeJesus-Hernandez M., Mackenzie I.R., Boeve B.F., Boxer A.L., Baker M., Rutherford N.J., et al. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. Neuron 72: 245–256.

Dekker A.M., Diekstra F.P., Pulit S.L., Tazelaar G.H.P., van der Spek R.A., van Rheenen

W., et al. (2019). Exome array analysis of rare and low frequency variants in amyotrophic lateral sclerosis. Scientific Reports 9: 5931.

Demetriou C.A., Heraclides A., Salafori C., Tanteles G.A., Christodoulou K., Christou Y., and Zamba-Papanicolaou E. (2018). Epidemiology of Huntington disease in Cyprus: A 20-year retrospective study. Clinical Genetics 93: 656–664.

Deng Y.P., Albin R.L., Penney J.B., Young A.B., Anderson K.D., and Reiner A. (2004). Differential loss of striatal projection systems in Huntington's disease: a quantitative immunohistochemical study. Journal of Chemical Neuroanatomy 27: 143–64.

DePristo M.A., Banks E., Poplin R., Garimella K. V, Maguire J.R., Hartl C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43: 491–8.

Desai A. and Gerson S. (2014). Exo1 independent DNA mismatch repair involves multiple compensatory nucleases. DNA Repair 21: 55–64.

Dewey F.E., Murray M.F., Overton J.D., Habegger L., Leader J.B., Fetterolf S.N., et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science 354: aaf6814.

Dherin C., Gueneau E., Francin M., Nunez M., Miron S., Liberti S.E., et al. (2009). Characterization of a highly conserved binding site of Mlh1 required for exonuclease I-dependent mismatch repair. Molecular and Cellular Biology 29: 907–18.

Dickey A.S. and La Spada A.R. (2018). Therapy development in Huntington disease: From current strategies to emerging opportunities. American Journal of Medical Genetics Part A 176: 842–861.

DiFiglia M., Sapp E., Chase K., Schwarz C., Meloni A., Young C., et al. (1995). Huntingtin is a cytoplasmic protein associated with vesicles in human and rat brain neurons. Neuron 14: 1075–81.

Djousse L., Knowlton B., Cupples L.A., Marder K., Shoulson I., and Myers R.H. (2002). Weight loss in early stage of Huntington's disease. Neurology 59: 1325–1330.

Djoussé L., Knowlton B., Hayden M., Almqvist E.W., Brinkman R., Ross C., et al. (2003). Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. American journal of medical genetics Part A 119A: 279–82.

Dolzhenko E., van Vugt J.J.F.A., Shaw R.J., Bekritsky M.A., van Blitterswijk M., Narzisi G., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Research 27: 1895–1903.

Douglas I., Evans S., Rawlins M.D., Smeeth L., Tabrizi S.J., and Wexler N.S. (2013). Juvenile Huntington's disease: a population-based study using the General Practice Research Database. BMJ Open 3: e002085.

Dragileva E., Hendricks A., Teed A., Gillis T., Lopez E.T., Friedberg E.C., et al. (2009). Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. Neurobiology of Disease 33: 37–47.

Drummond J., Li G., Longley M., and Modrich P. (1995). Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. Science 268: 1909–1912.

Du J., Campau E., Soragni E., Ku S., Puckett J.W., Dervan P.B., and Gottesfeld J.M. (2012). Role of Mismatch Repair Enzymes in GAA·TTC Triplet-repeat Expansion in Friedreich Ataxia Induced Pluripotent Stem Cells. Journal of Biological Chemistry 287: 29861–29872.

Duff K., Paulsen J., Mills J., Beglinger L.J., Moser D.J., Smith M.M., et al. (2010). Mild cognitive impairment in prediagnosed Huntington disease. Neurology 75: 500–7.

van Duijn E., Craufurd D., Hubers A.A.M., Giltay E.J., Bonelli R., Rickards H., et al. (2014). Neuropsychiatric symptoms in a European Huntington's disease cohort (REGISTRY). Journal of Neurology, Neurosurgery, and Psychiatry 85: 1411–8.

van Duijn E., Kingma E.M., and van der Mast R.C. (2007). Psychopathology in verified Huntington's disease gene carriers. The Journal of Neuropsychiatry and Clinical Neurosciences 19: 441–8.

van Duijn E., Vrijmoeth E.M., Giltay E.J., Bernhard Landwehrmeyer G., and REGISTRY investigators of the European Huntington's Disease Network (2018). Suicidal ideation and suicidal behavior according to the C-SSRS in a European cohort of Huntington's disease gene expansion carriers. Journal of Affective Disorders 228: 194–204.

Dumas E.M., van den Bogaard S.J.A., Middelkoop H.A.M., and Roos R.A.C. (2013). A review of cognition in Huntington's disease. Frontiers in Bioscience 5: 1–18.

Dunah A.W., Jeong H., Griffin A., Kim Y.-M., Standaert D.G., Hersch S.M., et al. (2002). Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. Science 296: 2238–43.

Dutta D., Scott L., Boehnke M., and Lee S. (2019). Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. Genetic Epidemiology 43: 4–23.

Duyao M., Ambrose C., Myers R., Novelletto A., Persichetti F., Frontali M., et al. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. Nature genetics 4: 387–92.

Duyao M., Auerbach A., Ryan A., Persichetti F., Barnes G., McNeil S., et al. (1995). Inactivation of the mouse Huntington's disease gene homolog Hdh. Science 269: 407–410.

Eddy C.M., Parkinson E.G., and Rickards H.E. (2016). Changes in mental state and behaviour in Huntington's disease. The Lancet Psychiatry 3: 1079–1086.

Ehrnhoefer D.E., Sutton L., and Hayden M.R. (2011). Small changes, big impact: posttranslational modifications and function of huntingtin in Huntington disease. The Neuroscientist 17: 475–92.

Eichler E.E., Holden J.J., Popovich B.W., Reiss A.L., Snow K., Thibodeau S.N., et al. (1994). Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Nature Genetics 8: 88–94.

Elden A.C., Kim H.-J., Hart M.P., Chen-Plotkin A.S., Johnson B.S., Fang X., et al. (2010). Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature 466: 1069–75.

Elias S., Thion M.S., Yu H., Sousa C.M., Lasgi C., Morin X., and Humbert S. (2014). Huntingtin Regulates Mammary Stem Cell Division and Differentiation. Stem Cell Reports 2: 491–506.

Ellis N., Tee A., McAllister B., Massey T., McLauchlan D., Stone T., et al. (2019). Genetic risk underlying psychiatric and cognitive symptoms in Huntington's Disease. [in review, bioarchive: https://www.biorxiv.org/content/10.1101/639658v3].

Emond M.J., Louie T., Emerson J., Zhao W., Mathias R.A., Knowles M.R., et al. (2012). Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nature Genetics 44: 886–9.

Epping E.A., Mills J.A., Beglinger L.J., Fiedorowicz J.G., Craufurd D., Smith M.M., et al. (2013). Characterization of depression in prodromal Huntington disease in the neurobiological predictors of HD (PREDICT-HD) study. Journal of Psychiatric Research 47: 1423–31.

Escott-Price V., Bracher-Smith M., Menzies G., Walters J., Kirov G., Owen M.J., and O'Donovan M.C. (2019). Genetic liability to schizophrenia is negatively associated with educational attainment in UK Biobank. Molecular Psychiatry [ahead of print].

Evans S.J.W., Douglas I., Rawlins M.D., Wexler N.S., Tabrizi S.J., and Smeeth L. (2013). Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. Journal of Neurology, Neurosurgery, and Psychiatry 84: 1156–60.

Ewing B. and Green P. (1998). Base-calling of automated sequencer traces using phred. II.

Error probabilities. Genome Research 8: 186–94.

Ewing B., Hillier L., Wendl M.C., and Green P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8: 175–85.

Faideau M., Kim J., Cormier K., Gilmore R., Welch M., Auregan G., et al. (2010). In vivo expression of polyglutamine-expanded huntingtin by mouse striatal astrocytes impairs glutamate transport: a correlation with Huntington's disease subjects. Human Molecular Genetics 19: 3053–67.

Falik-Zaccai T.C., Shachak E., Yalon M., Lis Z., Borochowitz Z., Macpherson J.N., et al. (1997). Predisposition to the fragile X syndrome in Jews of Tunisian descent is due to the absence of AGG interruptions on a rare Mediterranean haplotype. American Journal of Human Genetics 60: 103–12.

Fardaei M., Rogers M.T., Thorpe H.M., Larkin K., Hamshere M.G., Harper P.S., and Brook J.D. (2002). Three proteins, MBNL, MBLL and MBXL, co-localize in vivo with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells. Human Molecular Genetics 11: 805–14.

Farrer L.A., Cupples L.A., Wiater P., Conneally P.M., Gusella J.F., and Myers R.H. (1993). The Normal Huntington Disease (HD) Allele, or a Closely Linked Gene, Influences Age at Onset of HD. American Journal of Human Genetics 53: 125–130.

Favaro F.P., Alvizi L., Zechi-Ceide R.M., Bertola D., Felix T.M., de Souza J., et al. (2014). A Noncoding Expansion in EIF4A3 Causes Richieri-Costa-Pereira Syndrome, a Craniofacial Disorder Associated with Limb Defects. American Journal of Human Genetics 94: 120–128.

Ferrari A.J., Somerville A.J., Baxter A.J., Norman R., Patten S.B., Vos T., and Whiteford H.A. (2013). Global variation in the prevalence and incidence of major depressive disorder: a systematic review of the epidemiological literature. Psychological Medicine 43: 471–81.

Fishel R. (2015). Mismatch Repair. Journal of Biological Chemistry 290: 26395–26403.

Fishel R. and Lee J.-B. (2016). Mismatch Repair. In, *DNA Replication, Recombination, and Repair*. Springer Japan, Tokyo, pp. 305–339.

Fisher E.R. and Hayden M.R. (2014). Multisource ascertainment of Huntington disease in Canada: Prevalence and population at risk. Movement Disorders 29: 105–114.

Flannick J., Mercader J.M., Fuchsberger C., Udler M.S., Mahajan A., Wessel J., et al. (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature 570: 71–76.

Flower M., Lomeikaite V., Ciosi M., Cumming S., Morales F., Lo K., et al. (2019). MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. Brain 142: 1876–1886.

Foiry L., Dong L., Savouret C., Hubert L., te Riele H., Junien C., and Gourdon G. (2006). Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. Human Genetics 119: 520–6.

Folstein S., Abbott M.H., Chase G.A., Jensen B.A., and Folstein M.F. (1983a). The association of affective disorder with Huntington's disease in a case series and in families. Psychological Medicine 13: 537–42.

Folstein S.E., Franz M.L., Jensen B.A., Chase G.A., and Folstein M.F. (1983b). Conduct disorder and affective disorder among the offspring of patients with Huntington's disease. Psychological Medicine 13: 45–52.

Franich N.R., Hickey M.A., Zhu C., Osborne G.F., Ali N., Chu T., et al. (2019). Phenotype onset in Huntington's disease knock-in mice is correlated with the incomplete splicing of the mutant huntingtin gene. Journal of Neuroscience Research [ahead of print].

Frankish A., Diekhans M., Ferreira A.-M., Johnson R., Jungreis I., Loveland J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research 47: D766–D773.

Freudenreich C.H. (2018). R-loops: targets for nuclease cleavage and repeat instability. Current Genetics 64: 789–794.

Fritz N.E., Boileau N.R., Stout J.C., Ready R., Perlmutter J.S., Paulsen J.S., et al. (2018). Relationships Among Apathy, Health-Related Quality of Life, and Function in Huntington's Disease. The Journal of Neuropsychiatry and Clinical Neurosciences 30: 194–201.

Fromer M., Pocklington A.J., Kavanagh D.H., Williams H.J., Dwyer S., Gormley P., et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature 506: 179–84.

Fu Y.-H., Kuhl D.P.A., Pizzuti A., Pieretti M., Sutcliffe J.S., Richards S., et al. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: Resolution of the Sherman paradox. Cell 67: 1047–1058.

Fuller C.W., Middendorf L.R., Benner S.A., Church G.M., Harris T., Huang X., et al. (2009). The challenges of sequencing by synthesis. Nature Biotechnology 27: 1013–23.

Fusilli C., Migliore S., Mazza T., Consoli F., De Luca A., Barbagallo G., et al. (2018). Biological and clinical manifestations of juvenile Huntington's disease: a retrospective analysis. The Lancet Neurology 17: 986–993.

Gaba A.M., Zhang K., Marder K., Moskowitz C.B., Werner P., and Boozer C.N. (2005).

Energy balance in early-stage Huntington disease. The American Journal of Clinical Nutrition 81: 1335–1341.

Gacy A.M., Goellner G., Juranić N., Macura S., and McMurray C.T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell 81: 533–40.

Gafni J. and Ellerby L.M. (2002). Calpain activation in Huntington's disease. The Journal of Neuroscience 22: 4842–9.

Gafni J., Hermel E., Young J.E., Wellington C.L., Hayden M.R., and Ellerby L.M. (2004). Inhibition of calpain cleavage of huntingtin reduces toxicity: accumulation of calpain/caspase fragments in the nucleus. The Journal of Biological Chemistry 279: 20211–20.

Gao R., Matsuura T., Coolbaugh M., Zühlke C., Nakamura K., Rasmussen A., et al. (2008). Instability of expanded CAG/CAA repeats in spinocerebellar ataxia type 17. European Journal of Medical Genetics 16: 215–22.

Gardiner S.L., Boogaard M.W., Trompet S., de Mutsert R., Rosendaal F.R., Gussekloo J., et al. (2019). Prevalence of Carriers of Intermediate and Pathological Polyglutamine Disease–Associated Alleles Among Large Population-Based Cohorts. JAMA Neurology 76: 650.

Garrison E. and Marth G. (2012). Haplotype-based variant detection from short-read sequencing. [pre-print, arXiv: https://arxiv.org/abs/1207.3907].

Gauthier L.R., Charrin B.C., Borrell-Pagès M., Dompierre J.P., Rangone H., Cordelières F.P., et al. (2004). Huntingtin Controls Neurotrophic Support and Survival of Neurons by Enhancing BDNF Vesicular Transport along Microtubules. Cell 118: 127–138.

GeM-HD Consortium (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell 178: 887–900.

GeM-HD Consortium (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell 162: 516–526.

Genovese G., Fromer M., Stahl E.A., Ruderfer D.M., Chambert K., Landén M., et al. (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nature Neuroscience 19: 1433–1441.

Genschel J., Bazemore L.R., and Modrich P. (2002). Human exonuclease I is required for 5' and 3' mismatch repair. The Journal of Biological Chemistry 277: 13302–11.

Gerfen C.R. (1992). The neostriatal mosaic: multiple levels of compartmental organization. Trends in Neurosciences 15: 133–9.

Gertler T.S., Chan C.S., and Surmeier D.J. (2008). Dichotomous Anatomical Properties of Adult Striatal Medium Spiny Neurons. Journal of Neuroscience 28: 10814–10824.

Glass M., Dragunow M., and Faull R.L. (2000). The pattern of neurodegeneration in Huntington's disease: a comparative study of cannabinoid, dopamine, adenosine and GABA(A) receptor alterations in the human basal ganglia in Huntington's disease. Neuroscience 97: 505–19.

Gloss B.S. and Dinger M.E. (2018). Realizing the significance of noncoding functionality in clinical genomics. Experimental & Molecular Medicine 50: 97.

Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E.M., Brockman W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature biotechnology 27: 182–9.

Goldberg Y.P., McMurray C.T., Zeisler J., Almqvist E., Sillence D., Richards F., et al. (1995). Increased instability of intermediate alleles in families with sporadic Huntington disease compared to similar sized intermediate alleles in the general population. Human Molecular Genetics 4: 1911–8.

Goldberg Y.P., Nicholson D.W., Rasper D.M., Kalchman M.A., Koide H.B., Graham R.K., et al. (1996). Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. Nature Genetics 13: 442–449.

Goldstrohm A.C., Albrecht T.R., Suñé C., Bedford M.T., and Garcia-Blanco M.A. (2001). The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. Molecular and Cellular Biology 21: 7617–28.

Gomes-Pereira M., Fortune M.T., Ingram L., McAbney J.P., and Monckton D.G. (2004). Pms2 is a genetic enhancer of trinucleotide CAG.CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. Human Molecular Genetics 13: 1815–25.

Gonitel R., Moffitt H., Sathasivam K., Woodman B., Detloff P.J., Faull R.L.M., and Bates G.P. (2008). DNA instability in postmitotic neurons. Proceedings of the National Academy of Sciences of the United States of America 105: 3467–72.

Gonzalez-Alegre P. and Afifi A.K. (2006). Clinical characteristics of childhood-onset (juvenile) Huntington disease: report of 12 patients and review of the literature. Journal of Child Neurology 21: 223–9.

Goold R., Flower M., Moss D.H., Medway C., Wood-Kaczmar A., Andre R., et al. (2019). FAN1 modifies Huntington's disease progression by stabilizing the expanded HTT CAG repeat. Human Molecular Genetics 28: 650–661.

Grabczyk E. and Usdin K. (2000). The GAA*TTC triplet repeat expanded in Friedreich's ataxia impedes transcription elongation by T7 RNA polymerase in a length and supercoil dependent manner. Nucleic Acids Research 28: 2815–22.

Gray M., Shirasaki D.I., Cepeda C., André V.M., Wilburn B., Lu X.-H., et al. (2008). Full-length human mutant huntingtin with a stable polyglutamine repeat can elicit progressive and selective neuropathogenesis in BACHD mice. The Journal of Neuroscience 28: 6182–95.

Graybiel A.M. (1995). The basal ganglia. Trends in Neurosciences 18: 60–2.

Gréen H., Hasmats J., Kupershmidt I., Edsgärd D., de Petris L., Lewensohn R., et al. (2016). Using Whole-Exome Sequencing to Identify Genetic Markers for Carboplatin and Gemcitabine-Induced Toxicities. Clinical Cancer Research 22: 366–73.

Grimbergen Y.A.M., Knol M.J., Bloem B.R., Kremer B.P.H., Roos R.A.C., and Munneke M. (2008). Falls and gait disturbances in Huntington's disease. Movement Disorders 23: 970–976.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature 550: 204–213.

Guo J., Gu L., Leffak M., and Li G.-M. (2016). MutSβ promotes trinucleotide repeat expansion by recruiting DNA polymerase β to nascent (CAG)n or (CTG)n hairpins for error-prone DNA synthesis. Cell Research 26: 775–786.

Guo Q., Bin Huang, Cheng J., Seefelder M., Engler T., Pfeifer G., et al. (2018). The cryo-electron microscopy structure of huntingtin. Nature 555: 117–120.

Gusella J.F. and MacDonald M.E. (2009). Huntington's disease: the case for genetic modifiers. Genome Medicine 1: 80.

Gusella J.F., MacDonald M.E., and Lee J.-M. (2014). Genetic modifiers of Huntington's disease. Movement Disorders 29: 1359–65.

Gusella J.F., Wexler N.S., Conneally P.M., Naylor S.L., Anderson M.A., Tanzi R.E., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234–8.

Gutekunst C.A., Li S.H., Yi H., Mulroy J.S., Kuemmerle S., Jones R., et al. (1999). Nuclear and neuropil aggregates in Huntington's disease: relationship to neuropathology. The Journal of Neuroscience 19: 2522–34.

Gwon G.H., Kim Y., Liu Y., Watson A.T., Jo A., Etheridge T.J., et al. (2014). Crystal structure of a Fanconi anemia-associated nuclease homolog bound to 5' flap DNA: basis of interstrand cross-link repair by FAN1. Genes & Development 28: 2276–90.

Haber S.N., Fudge J.L., and McFarland N.R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. The Journal of Neuroscience 20: 2369–82.

Habraken Y., Sung P., Prakash L., and Prakash S. (1996). Binding of insertion/deletion DNA mismatches by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3. Current Biology 6: 1185–1187.

Halabi A., Fuselier K.T.B., and Grabczyk E. (2018). GAA•TTC repeat expansion in human cells is mediated by mismatch repair complex MutLγ and depends upon the endonuclease domain in MLH3 isoform one. Nucleic Acids Research 46: 4022–4032.

Hard G.C. (2018). Critical review of renal tubule karyomegaly in non-clinical safety evaluation studies and its significance for human risk assessment. Critical Reviews in Toxicology 48: 575–595.

Harrington D.L., Smith M.M., Zhang Y., Carlozzi N.E., Paulsen J.S., and PREDICT-HD Investigators of the Huntington Study Group (2012). Cognitive domains that predict time to diagnosis in prodromal Huntington disease. Journal of Neurology, Neurosurgery, and Psychiatry 83: 612–9.

Hayden M.R., MacGregor J.M., and Beighton P.H. (1980). The prevalence of Huntington's chorea in South Africa. South African Medical Journal 58: 193–6.

Hayward B.E. and Usdin K. (2017). Improved Assays for AGG Interruptions in Fragile X Premutation Carriers. The Journal of Molecular Diagnostics 19: 828–835.

Hayward B.E., Zhou Y., Kumari D., and Usdin K. (2016). A Set of Assays for the Comprehensive Analysis of FMR1 Alleles in the Fragile X-Related Disorders. The Journal of Molecular Diagnostics 18: 762–774.

Hazeki N., Nakamura K., Goto J., and Kanazawa I. (1999). Rapid aggregate formation of the huntingtin N-terminal fragment carrying an expanded polyglutamine tract. Biochemical and Biophysical Research Communications 256: 361–6.

HD iPSC Consortium (2012). Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-expansion-associated phenotypes. Cell Stem Cell 11: 264–78.

Hendricks A.E., Latourelle J.C., Lunetta K.L., Cupples L.A., Wheeler V., MacDonald M.E., et al. (2009). Estimating the probability of de novo HD cases from transmissions of expanded penetrant CAG alleles in the Huntington disease gene from male carriers of high normal alleles (27-35 CAG). American Journal of Medical Genetics Part A 149A: 1375–1381.

Hensman Moss D.J., Pardiñas A.F., Langbehn D., Lo K., Leavitt B.R., Roos R., et al. (2017). Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. The Lancet Neurology 16: 701–711.

Hodges E., Xuan Z., Balija V., Kramer M., Molla M.N., Smith S.W., et al. (2007). Genome-wide in situ exon capture for selective resequencing. Nature Genetics 39: 1522–7.

Hoffman-Andrews L. (2017). The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. Journal of Law and the Biosciences 4: 648–657.

Hoffmann R., Stüwe S.H., Goetze O., Banasch M., Klotz P., Lukas C., et al. (2014). Progressive hepatic mitochondrial dysfunction in premanifest Huntington's disease. Movement Disorders 29: 831–834.

Holbert S., Denghien I., Kiechle T., Rosenblatt A., Wellington C., Hayden M.R., et al. (2001). The Gln-Ala repeat transcriptional activator CA150 interacts with huntingtin: neuropathologic and genetic evidence for a role in Huntington's disease pathogenesis. Proceedings of the National Academy of Sciences of the United States of America 98: 1811–6.

Holmes S.E., O'Hearn E., Rosenblatt A., Callahan C., Hwang H.S., Ingersoll-Ashworth R.G., et al. (2001). A repeat expansion in the gene encoding junctophilin-3 is associated with Huntington disease-like 2. Nature Genetics 29: 377–8.

Holmes S.E., O'Hearn E.E., McInnis M.G., Gorelick-Feldman D.A., Kleiderlein J.J., Callahan C., et al. (1999). Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12. Nature Genetics 23: 391–392.

Hornbeck P. V, Zhang B., Murray B., Kornhauser J.M., Latham V., and Skrzypek E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Research 43: D512-20.

Hsieh P. and Zhang Y. (2017). The Devil is in the details for DNA mismatch repair. Proceedings of the National Academy of Sciences of the United States of America 114: 3552–3554.

Hsu R.-J., Hsiao K.-M., Lin M.-J., Li C.-Y., Wang L.-C., Chen L.-K., and Pan H. (2011). Long tract of untranslated CAG repeats is deleterious in transgenic mice. PLOS ONE 6: e16417.

Huang Y.-F., Gulko B., and Siepel A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nature Genetics 49: 618–624.

Huntington G. (1872). On Chorea. The Medical and Surgical Reporter 26: 317–321.

Huntington Study Group (1996). Unified Huntington's Disease Rating Scale: reliability and consistency. Movement Disorders 11: 136–42.

Huntington Study Group PHAROS Investigators, Biglan K.M., Shoulson I., Kieburtz K., Oakes D., Kayson E., et al. (2016). Clinical-Genetic Associations in the Prospective Huntington at Risk Observational Study (PHAROS): Implications for Clinical Trials. JAMA Neurology 73: 102–10.

Illarioshkin S.N., Igarashi S., Onodera O., Markova E.D., Nikolskaya N.N., Tanaka H., et al. (1994). Trinucleotide repeat length and rate of progression of Huntington's disease. Annals of Neurology 36: 630–5.

Ionita-Laza I., McCallum K., Xu B., and Buxbaum J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics 48: 214–20.

Ishiura H., Doi K., Mitsui J., Yoshimura J., Matsukawa M.K., Fujiyama A., et al. (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nature Genetics 50: 581–590.

Isnard P., Rabant M., Labaye J., Antignac C., Knebelmann B., and Zaidan M. (2016). Karyomegalic Interstitial Nephritis: A Case Report and Review of the Literature. Medicine 95: e3349.

Itan Y., Shang L., Boisson B., Ciancanelli M.J., Markle J.G., Martinez-Barricarte R., et al. (2016). The mutation significance cutoff: gene-level thresholds for variant predictions. Nature Methods 13: 109–10.

Jain A. and Vale R.D. (2017). RNA phase transitions in repeat expansion disorders. Nature 546: 243–247.

Jalali Sefid Dashti M. and Gamieldien J. (2017). A practical guide to filtering and prioritizing genetic variants. BioTechniques 62: 18–30.

Jamali Z., Dianatpour M., Miryounesi M., and Modarressi M.H. (2018). A study of CAG repeat instability of HTT gene following spermatogenesis, by single sperm analysis. Gene Reports 12: 294–298.

Jarem D.A., Huckaby L. V., and Delaney S. (2010). AGG Interruptions in (CGG) n DNA Repeat Tracts Modulate the Structure and Thermodynamics of Non-B Conformations in Vitro. Biochemistry 49: 6826–6837.

Jensen P., Fenger K., Bolwig T.G., and Sorensen S.A. (1998). Crime in Huntington's disease: a study of registered offences among patients, relatives, and controls. Journal of Neurology, Neurosurgery, and Psychiatry 65: 467–471.

Jeon J.-P., Shim S.-M., Nam H.-Y., Baik S.-Y., Kim J.-W., and Han B.-G. (2007). Copy number increase of 1p36.33 and mitochondrial genome amplification in Epstein-Barr virus-transformed lymphoblastoid cell lines. Cancer Genetics and Cytogenetics 173: 122–30.

Jimenez-Sanchez M., Licitra F., Underwood B.R., and Rubinsztein D.C. (2017). Huntington's Disease: Mechanisms of Pathogenesis and Therapeutic Strategies. Cold Spring Harbor Perspectives in Medicine 7: a024240.

Jin H. and Cho Y. (2017). Structural and functional relationships of FAN1. DNA Repair 56: 135–143.

Jin H., Roy U., Lee G., Schärer O.D., and Cho Y. (2018). Structural mechanism of DNA interstrand cross-link unhooking by the bacterial FAN1 nuclease. The Journal of Biological Chemistry 293: 6482–6496.

Jiricny J. (2006). The multifaceted mismatch-repair system. Nature Reviews Molecular Cell Biology 7: 335–346.

Jodeiri Farshbaf M. and Ghaedi K. (2017). Huntington's Disease and Mitochondria. Neurotoxicity Research 32: 518–529.

Jodice C., Mantuano E., Veneziano L., Trettel F., Sabbadini G., Calandriello L., et al. (1997). Episodic ataxia type 2 (EA2) and spinocerebellar ataxia type 6 (SCA6) due to CAG repeat expansion in the CACNA1A gene on chromosome 19p. Human Molecular Genetics 6: 1973–8.

Joesch-Cohen L.M. and Glusman G. (2017). Differences between the genomes of lymphoblastoid cell lines and blood-derived samples. Advances in Genomics and Genetics 7: 1–9.

Josefsen K., Nielsen S.M.B., Campos A., Seifert T., Hasholt L., Nielsen J.E., et al. (2010). Reduced gluconeogenesis and lactate clearance in Huntington's disease. Neurobiology of Disease 40: 656–62.

Julien C.L., Thompson J.C., Wild S., Yardumian P., Snowden J.S., Turner G., and Craufurd D. (2007). Psychiatric disorders in preclinical Huntington's disease. Journal of Neurology, Neurosurgery, and Psychiatry 78: 939–43.

Jun G., Flickinger M., Hetrick K.N., Romm J.M., Doheny K.F., Abecasis G.R., et al. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. American Journal of Human Genetics 91: 839–48.

Kantartzis A., Williams G.M., Balakrishnan L., Roberts R.L., Surtees J.A., and Bambara R.A. (2012). Msh2-Msh3 Interferes with Okazaki Fragment Processing to Promote Trinucleotide Repeat Expansions. Cell Reports 2: 216–222.

Karczewski K.J., Francioli L.C., Tiao G., Cummings B.B., Alföldi J., Wang Q., et al. (2019).

Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. [pre-print, bioarchive: https://www.biorxiv.org/content/early/2019/01/30/531210].

Kato N., Kawasoe Y., Williams H., Coates E., Roy U., Shi Y., et al. (2017). Sensing and Processing of DNA Interstrand Crosslinks by the Mismatch Repair Pathway. Cell Reports 21: 1375–1385.

Kay C., Collins J.A., Miedzybrodzka Z., Madore S.J., Gordon E.S., Gerry N., et al. (2016). Huntington disease reduced penetrance alleles occur at high frequency in the general population. Neurology 87: 282–288.

Kay C., Collins J.A., Wright G.E.B., Baine F., Miedzybrodzka Z., Aminkeng F., et al. (2018). The molecular epidemiology of Huntington disease is related to intermediate allele frequency and haplotype in the general population. American Journal of Medical Genetics Part B 177: 346–357.

Kay C., Hayden M.R., and Leavitt B.R. (2017). Epidemiology of Huntington disease. Handbook of Clinical Neurology 144: 31–46.

Kehoe P., Krawczak M., Harper P.S., Owen M.J., and Jones A.L. (1999). Age of onset in Huntington disease: sex specific influence of apolipoprotein E genotype and normal CAG repeat length. Journal of Medical Genetics 36: 108–11.

Kennedy L., Evans E., Chen C.-M., Craven L., Detloff P.J., Ennis M., and Shelbourne P.F. (2003). Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. Human Molecular Genetics 12: 3359–67.

Kessler R.C., Berglund P., Demler O., Jin R., Merikangas K.R., and Walters E.E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry 62: 593–602.

Kieburtz K., MacDonald M., Shih C., Feigin A., Steinberg K., Bordwell K., et al. (1994). Trinucleotide repeat length and progression of illness in Huntington’s disease. Journal of Medical Genetics 31: 872–874.

Kiliszek A., Kierzek R., Krzyzosiak W.J., and Rypniewski W. (2010). Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. Nucleic Acids Research 38: 8370–6.

Kim H.S., Lyoo C.H., Lee P.H., Kim S.J., Park M.Y., Ma H.-I., et al. (2015). Current Status of Huntington’s Disease in Korea: A Nationwide Survey and National Registry Analysis. Journal of Movement Disorders 8: 14–20.

Kim Y.J., Yi Y., Sapp E., Wang Y., Cuiffo B., Kegel K.B., et al. (2001). Caspase 3-cleaved N-terminal fragments of wild-type and mutant huntingtin are present in normal and Huntington's disease brains, associate with membranes, and undergo calpain-dependent proteolysis. Proceedings of the National Academy of Sciences of the United States of America 98: 12784–12789.

Kingma E.M., van Duijn E., Timman R., van der Mast R.C., and Roos R.A.C. (2008). Behavioural problems in Huntington's disease using the Problem Behaviours Assessment. General Hospital Psychiatry 30: 155–61.

Kircher M., Witten D.M., Jain P., O'Roak B.J., Cooper G.M., and Shendure J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics 46: 310–5.

Kita H. and Kitai S.T. (1988). Glutamate decarboxylase immunoreactive neurons in rat neostriatum: their morphological types and populations. Brain Research 447: 346–52.

Klawonn A.M. and Malenka R.C. (2018). Nucleus Accumbens Modulation in Reward and Aversion. Cold Spring Harbor Symposia on Quantitative Biology 83: 119–129.

Kleinstein S.E., Rein M., Abdelmalek M.F., Guy C.D., Goldstein D.B., Mae Diehl A., and Moylan C.A. (2018). Whole-Exome Sequencing Study of Extreme Phenotypes of NAFLD. Hepatology Communications 2: 1021–1029.

Klempíř J., Zidovská J., Stochl J., Ing V.K., Uhrová T., and Roth J. (2011). The number of CAG repeats within the normal allele does not influence the age of onset in Huntington's disease. Movement Disorders 26: 125–9.

Klöppel S., Gregory S., Scheller E., Minkova L., Razi A., Durr A., et al. (2015). Compensation in Preclinical Huntington's Disease: Evidence From the Track-On HD Study. EBioMedicine 2: 1420–1429.

Klöppel S., Stonnington C.M., Petrovic P., Mobbs D., Tüscher O., Craufurd D., et al. (2010). Irritability in pre-clinical Huntington's disease. Neuropsychologia 48: 549–57.

Knight S.J.L., Flannery A.V., Hirst M.C., Campbell L., Christodoulou Z., Phelps S.R., et al. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. Cell 74: 127–134.

Kobayashi H., Abe K., Matsuura T., Ikeda Y., Hitomi T., Akechi Y., et al. (2011). Expansion of Intronic GGCCTG Hexanucleotide Repeat in NOP56 Causes SCA36, a Type of Spinocerebellar Ataxia Accompanied by Motor Neuron Involvement. American Journal of Human Genetics 89: 121–130.

Koide R., Ikeuchi T., Onodera O., Tanaka H., Igarashi S., Endo K., et al. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nature Genetics 6: 9–13.

Koob M.D., Moseley M.L., Schut L.J., Benzow K.A., Bird T.D., Day J.W., and Ranum L.P.W. (1999). An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nature Genetics 21: 379–384.

Koressaar T. and Remm M. (2007). Enhancements and modifications of primer design program Primer3. Bioinformatics 23: 1289–91.

Kovalenko M., Dragileva E., St Claire J., Gillis T., Guide J.R., New J., et al. (2012). Msh2 acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. PLOS ONE 7: e44273.

Kovtun I. V, Liu Y., Bjoras M., Klungland A., Wilson S.H., and Mcmurray C.T. (2009). OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. Nature 447: 447–452.

Kramara J., Osia B., and Malkova A. (2018). Break-Induced Replication: The Where, The Why, and The How. Trends in Genetics 34: 518–531.

Kratz K., Schöpf B., Kaden S., Sendoel A., Eberhard R., Lademann C., et al. (2010). Deficiency of FANCD2-associated nuclease KIAA1018/FAN1 sensitizes cells to interstrand crosslinking agents. Cell 142: 77–88.

Ku C.S., Polychronakos C., Tan E.K., Naidoo N., Pawitan Y., Roukos D.H., et al. (2013). A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. Molecular Psychiatry 18: 141–53.

Kuemmerle S., Gutekunst C.A., Klein A.M., Li X.J., Li S.H., Beal M.F., et al. (1999). Huntington aggregates may not predict neuronal death in Huntington's disease. Annals of Neurology 46: 842–9.

van Kuilenburg A.B.P., Tarailo-Graovac M., Richmond P.A., Drögemöller B.I., Pouladi M.A., Leen R., et al. (2019). Glutaminase Deficiency Caused by Short Tandem Repeat Expansion in GLS. New England Journal of Medicine 380: 1433–1441.

Kumar C., Williams G.M., Havens B., Dinicola M.K., and Surtees J.A. (2013). Distinct requirements within the Msh3 nucleotide binding pocket for mismatch and double-strand break repair. Journal of Molecular Biology 425: 1881–1898.

Kumar H. and Jog M. (2011). Missing Huntington's disease for tardive dyskinesia: a preventable error. The Canadian Journal of Neurological Sciences 38: 762–4.

de la Monte S.M., Vonsattel J.P., and Richardson E.P. (1988). Morphometric demonstration of atrophic changes in the cerebral cortex, white matter, and neostriatum in Huntington's disease. Journal of Neuropathology and Experimental Neurology 47: 516–25.

Lachaud C., Slean M., Marchesi F., Lock C., Odell E., Castor D., et al. (2016). Karyomegalic interstitial nephritis and DNA damage-induced polyploidy in Fan1 nuclease-defective knock-in mice. v 30: 639–44.

LaCroix A.J., Stabley D., Sahraoui R., Adam M.P., Mehaffey M., Kernan K., et al. (2019). GGC Repeat Expansion and Exon 1 Methylation of XYLT1 Is a Common Pathogenic Variant in Baratela-Scott Syndrome. American Journal of Human Genetics 104: 35–44.

Lai Y., Budworth H., Beaver J.M., Chan N.L.S., Zhang Z., McMurray C.T., and Liu Y. (2016). Crosstalk between MSH2–MSH3 and polβ promotes trinucleotide repeat expansion during base excision repair. Nature Communications 7: 12465.

Lakra P., Aditi K., and Agrawal N. (2019). Peripheral Expression of Mutant Huntingtin is a Critical Determinant of Weight Loss and Metabolic Disturbances in Huntington's Disease. Scientific Reports 9: 10127.

Lalić N.M., Marić J., Svetel M., Jotić A., Stefanova E., Lalić K., et al. (2008). Glucose Homeostasis in Huntington Disease. Archives of Neurology 65: 476.

Lalioti M.D., Mirotsou M., Buresi C., Peitsch M.C., Rossier C., Ouazzani R., et al. (1997). Identification of mutations in cystatin B, the gene responsible for the Unverricht-Lundborg type of progressive myoclonus epilepsy (EPM1). American Journal of Human Genetics 60: 342–51.

Landwehrmeyer G.B., Fitzer-Attas C.J., Giuliano J.D., Gonçalves N., Anderson K.E., Cardoso F., et al. (2016). Data Analytics from Enroll-HD, a Global Clinical Research Platform for Huntington's Disease. Movement Disorders Clinical Practice 4: 212–224.

Langbehn D.R., Brinkman R.R., Falush D., Paulsen J.S., and Hayden M.R. (2004). A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clinical Genetics 65: 267–277.

Langbehn D.R., Hayden M.R., Paulsen J.S., Johnson H., Aylward E., Biglan K., et al. (2010). CAG-repeat length and the age of onset in Huntington Disease (HD): A review and validation study of statistical approaches. American Journal of Medical Genetics, Part B 153: 397–408.

Langbehn D.R., Stout J.C., Gregory S., Mills J.A., Durr A., Leavitt B.R., et al. (2019). Association of CAG Repeats With Long-term Progression in Huntington Disease. JAMA Neurology [ahead of print].

Larsson M.U., Luszcz M.A., Bui T.-H., and Wahlin T.-B.R. (2006). Depression and suicidal ideation after predictive testing for Huntington's disease: a two-year follow-up study. Journal of genetic counseling 15: 361–74.

Lee B.I. and Wilson D.M. (1999). The RAD2 domain of human exonuclease 1 exhibits 5' to 3' exonuclease and flap structure-specific endonuclease activities. The Journal of Biological Chemistry 274: 37763–9.

Lee J.-H., Lee J.-M., Ramos E.M., Gillis T., Mysore J.S., Kishikawa S., et al. (2012a). TAA repeat variation in the GRIK2 gene does not influence age at onset in Huntington's disease. Biochemical and Biophysical Research Communications 424: 404–8.

Lee J.-M., Chao M.J., Harold D., Abu Elneel K., Gillis T., Holmans P., et al. (2017). A modifier of Huntington's disease onset at the MLH1 locus. Human Molecular Genetics 26: 3859–3867.

Lee J.-M., Gillis T., Mysore J.S., Ramos E.M., Myers R.H., Hayden M.R., et al. (2012b). Common SNP-based haplotype analysis of the 4p16.3 Huntington disease gene region. American Journal of Human Genetics 90: 434–44.

Lee J.-M., Pinto R.M., Gillis T., St. Claire J.C., and Wheeler V.C. (2011). Quantification of Age-Dependent Somatic CAG Repeat Instability in Hdh CAG Knock-In Mice Reveals Different Expansion Dynamics in Striatum and Liver. PLOS ONE 6: e23647.

Lee J.-M., Zhang J., Su A.I., Walker J.R., Wiltshire T., Kang K., et al. (2010). A novel approach to investigate tissue-specific trinucleotide repeat instability. BMC Systems Biology 4: 29.

Lee J.M., Ramos E.M., Lee J.H., Gillis T., Mysore J.S., Hayden M.R., et al. (2012c). CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology 78: 690–695.

Lee S., Abecasis G.R., Boehnke M., and Lin X. (2014). Rare-variant association analysis: study designs and statistical tests. American Journal of Human Genetics 95: 5–23.

Lee S., Emond M.J., Bamshad M.J., Barnes K.C., Rieder M.J., Nickerson D.A., et al. (2012d). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American Journal of Human Genetics 91: 224–37.

Lee S., Wu M.C., and Lin X. (2012e). Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13: 762–75.

Lek M., Karczewski K.J., Minikel E. V, Samocha K.E., Banks E., Fennell T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285–91.

Lemay M., Fimbel E., Beuter A., Chouinard S., and Richer F. (2005). Sensorimotor mapping affects movement correction deficits in early Huntington’s disease. Experimental Brain Research 165: 454–460.

Levin D.S., McKenna A.E., Motycka T.A., Matsumoto Y., and Tomkinson A.E. (2000). Interaction between PCNA and DNA ligase I is critical for joining of Okazaki fragments and long-patch base-excision repair. Current Biology 10: 919–22.

Li D., Lewinger J.P., Gauderman W.J., Murcray C.E., and Conti D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genetic Epidemiology 35: 790–9.

Li H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–93.

Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–60.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–9.

Li L.-B., Yu Z., Teng X., and Bonini N.M. (2008). RNA toxicity is a component of ataxin-3 degeneration in Drosophila. Nature 453: 1107–1111.

Li Y., Shen J., and Niu H. (2019). DNA duplex recognition activates Exo1 nuclease activity. Journal of Biological Chemistry 294: 11559–11567.

Liddelow S.A., Guttenplan K.A., Clarke L.E., Bennett F.C., Bohlen C.J., Schirmer L., et al. (2017). Neurotoxic reactive astrocytes are induced by activated microglia. Nature 541: 481–487.

Lieberman A.P., Shakkottai V.G., and Albin R.L. (2019). Polyglutamine Repeats in Neurodegenerative Diseases. Annual review of Pathology 14: 1–27.

Lim G.Y., Tam W.W., Lu Y., Ho C.S., Zhang M.W., and Ho R.C. (2018). Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. Scientific Reports 8: 2861.

Lin Y., Dent S.Y.R., Wilson J.H., Wells R.D., and Napierala M. (2010). R loops stimulate genetic instability of CTG.CAG repeats. Proceedings of the National Academy of Sciences of the United States of America 107: 692–7.

Liot G., Zala D., Pla P., Mottet G., Piel M., and Saudou F. (2013). Mutant Huntingtin Alters Retrograde Transport of TrkB Receptors in Striatal Dendrites. Journal of Neuroscience 33:

6298–6309.

Liquori C.L., Ricker K., Moseley M.L., Jacobsen J.F., Kress W., Naylor S.L., et al. (2001). Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science 293: 864–7.

Liu B., Hu J., Wang J., and Kong D. (2017). Direct Visualization of RNA-DNA Primer Removal from Okazaki Fragments Provides Support for Flap Cleavage and Exonucleolytic Pathways in Eukaryotic Cells. The Journal of Biological Chemistry 292: 4777–4788.

Liu G., Chen X., Bissler J.J., Sinden R.R., and Leffak M. (2010a). Replication-dependent instability at (CTG) x (CAG) repeat hairpins in human cells. Nature Chemical Biology 6: 652–9.

Liu G., Chen X., and Leffak M. (2013a). Oligodeoxynucleotide binding to (CTG) · (CAG) microsatellite repeats inhibits replication fork stalling, hairpin formation, and genome instability. Molecular and Cellular Biology 33: 571–81.

Liu J., Fan S., Lee C.-J., Greenleaf A.L., and Zhou P. (2013b). Specific interaction of the transcription elongation regulator TCERG1 with RNA polymerase II requires simultaneous phosphorylation at Ser2, Ser5, and Ser7 within the carboxyl-terminal domain repeat. The Journal of Biological Chemistry 288: 10890–901.

Liu K.-Y., Shyu Y.-C., Barbaro B.A., Lin Y.-T., Chern Y., Thompson L.M., et al. (2015). Disruption of the nuclear membrane by perinuclear inclusions of mutant huntingtin causes cell-cycle re-entry and striatal cell death in mouse and cell models of Huntington's disease. Human Molecular Genetics 24: 1602–16.

Liu T., Ghosal G., Yuan J., Chen J., and Huang J. (2010b). FAN1 acts with FANCI-FANCD2 to promote DNA interstrand cross-link repair. Science 329: 693–6.

Liu X., Jian X., and Boerwinkle E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human Mutation 32: 894–9.

Liu X., Wu C., Li C., and Boerwinkle E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Human Mutation 37: 235–41.

Liu Y. and Bambara R.A. (2003). Analysis of Human Flap Endonuclease 1 Mutants Reveals a Mechanism to Prevent Triplet Repeat Expansion. Journal of Biological Chemistry 278: 13728–13739.

Liu Y., Prasad R., Beard W.A., Hou E.W., Horton J.K., McMurray C.T., and Wilson S.H. (2009). Coordination between Polymerase β and FEN1 Can Modulate CAG Repeat

Expansion. Journal of Biological Chemistry 284: 28352–28366.

Londin E.R., Keller M.A., D’Andrea M.R., Delgrosso K., Ertel A., Surrey S., and Fortina P. (2011). Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. BMC Genomics 12: 464.

Longley M.J., Pierce A.J., and Modrich P. (1997). DNA polymerase delta is required for human mismatch repair in vitro. The Journal of Biological Chemistry 272: 10917–21.

Lopes C., Aubert S., Bourgois-Rocha F., Barnat M., Rego A.C., Déglon N., et al. (2016). Dominant-Negative Effects of Adult-Onset Huntingtin Mutations Alter the Division of Human Embryonic Stem Cells-Derived Neural Cells. PLOS ONE 11: e0148680.

Lovestone S., Hodgson S., Sham P., Differ A.M., and Levy R. (1996). Familial psychiatric presentation of Huntington’s disease. Journal of Medical Genetics 33: 128–131.

Luo Y., Maity A., Wu M.C., Smith C., Duan Q., Li Y., and Tzeng J.-Y. (2018). On the substructure controls in rare variant analysis: Principal components or variance components? Genetic Epidemiology 42: 276–287.

Macaulay I.C., Haerty W., Kumar P., Li Y.I., Hu T.X., Teng M.J., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nature Methods 12: 519–22.

Macaulay I.C., Teng M.J., Haerty W., Kumar P., Ponting C.P., and Voet T. (2016). Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&amp;T-seq. Nature Protocols 11: 2081–2103.

MacDonald M.E., Barnes G., Srinidhi J., Duyao M.P., Ambrose C.M., Myers R.H., et al. (1993). Gametic but not somatic instability of CAG repeat length in Huntington’s disease. Journal of Medical Genetics 30: 982–6.

MacDonald M.E., Vonsattel J.P., Shrinidhi J., Couropmitree N.N., Cupples L.A., Bird E.D., et al. (1999). Evidence for the GluR6 gene associated with younger onset age of Huntington’s disease. Neurology 53: 1330–2.

MacKay C., Déclais A.-C., Lundin C., Agostinho A., Deans A.J., MacArtney T.J., et al. (2010). Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. Cell 142: 65–76.

Mahadevan M., Tsilfidis C., Sabourin L., Shutler G., Amemiya C., Jansen G., et al. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3’ untranslated region of the gene. Science 255: 1253–1255.

Maiuri T., Mocle A.J., Hung C.L., Xia J., van Roon-Mom W.M.C., and Truant R. (2017). Huntingtin is a scaffolding protein in the ATM oxidative DNA damage response complex.

Human Molecular Genetics 26: 395–406.

Maiuri T., Suart C.E., Hung C.L.K., Graham K.J., Barba Bazan C.A., and Truant R. (2019). DNA Damage Repair in Huntington’s Disease and Other Neurodegenerative Diseases. Neurotherapeutics [ahead of print].

Maltecca F., Filla A., Castaldo I., Coppola G., Fragassi N.A., Carella M., et al. (2003). Intergenerational instability and marked anticipation in SCA-17. Neurology 61: 1441–3.

Mangiarini L., Sathasivam K., Seller M., Cozens B., Harper A., Hetherington C., et al. (1996). Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. Cell 87: 493–506.

Manley K., Shirley T.L., Flaherty L., and Messer A. (1999). Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nature Genetics 23: 471–3.

Mantere T., Kersten S., and Hoischen A. (2019). Long-Read Sequencing Emerging in Medical Genetics. Frontiers in Genetics 10: 426.

Margolis R.L., McInnis M.G., Rosenblatt A., and Ross C.A. (1999). Trinucleotide repeat expansion and neuropsychiatric disease. Archives of General Psychiatry 56: 1019–31.

Mariotti C., Castellotti B., Pareyson D., Testa D., Eoli M., Antozzi C., et al. (2000). Phenotypic manifestations associated with CAG-repeat expansion in the androgen receptor gene in male patients and heterozygous females: a clinical and molecular study of 30 families. Neuromuscular Disorders 10: 391–7.

Markianos M., Panas M., Kalfakis N., and Vassilopoulos D. (2005). Plasma testosterone in male patients with Huntington’s disease: Relations to severity of illness and dementia. Annals of Neurology 57: 520–525.

Martin D.D.O., Heit R.J., Yap M.C., Davidson M.W., Hayden M.R., and Berthiaume L.G. (2014). Identification of a post-translationally myristoylated autophagy-inducing domain released by caspase cleavage of Huntingtin. Human Molecular Genetics 23: 3166–3179.

Martin D.D.O., Ladha S., Ehrnhoefer D.E., and Hayden M.R. (2015). Autophagy in Huntington disease and huntingtin in autophagy. Trends in Neurosciences 38: 26–35.

Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10.

Martinez-Horta S., Perez-Perez J., van Duijn E., Fernandez-Bobadilla R., Carceller M., Pagonabarraga J., et al. (2016). Neuropsychiatric symptoms are very common in premanifest and early stage Huntington’s Disease. Parkinsonism & Related Disorders 25:

58–64.

Martino D., Stamelou M., and Bhatia K.P. (2013). The differential diagnosis of Huntington's disease-like syndromes: "red flags" for the clinician. Journal of Neurology, Neurosurgery, and Psychiatry 84: 650–6.

Massey T., McAllister B., and Jones L. (2018). Methods for Assessing DNA Repair and Repeat Expansion in Huntington's Disease. Methods in Molecular biology 1780: 483–495.

Massey T.H. and Jones L. (2018). The central role of DNA damage and repair in CAG repeat diseases. Disease Models & Mechanisms 11: dmm031930.

Matsumoto Y. and Kim K. (1995). Excision of deoxyribose phosphate residues by DNA polymerase beta during DNA repair. Science 269: 699–702.

Matsuura T., Fang P., Pearson C.E., Jayakar P., Ashizawa T., Roa B.B., and Nelson D.L. (2006). Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: repeat purity as a disease modifier? American Journal of Human Genetics 78: 125–9.

Matsuura T., Yamagata T., Burgess D.L., Rasmussen A., Grewal R.P., Watase K., et al. (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. Nature Genetics 26: 191–194.

Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E., Wang H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science 337: 1190–5.

McCauley M.J., Furman L., Dietrich C.A., Rouzina I., Núñez M.E., and Williams M.C. (2018). Quantifying the stability of oxidatively damaged DNA by single-molecule DNA stretching. Nucleic Acids Research 46: 4033–4043.

McCourt A.C., O'Donovan K.L., Ekblad E., Sand E., Craufurd D., Rosser A., et al. (2015). Characterization of Gastric Mucosa Biopsies Reveals Alterations in Huntington's Disease. PLOS Currents.

McCusker E. and Loy C.T. (2014). The many facets of unawareness in huntington disease. Tremor and Other Hyperkinetic Movements 4: 257.

McCusker E.A., Gunn D.G., Epping E.A., Loy C.T., Radford K., Griffith J., et al. (2013). Unawareness of motor phenoconversion in Huntington disease. Neurology 81: 1141–7.

McFarland K.N., Liu J., Landrian I., Godiska R., Shanker S., Yu F., et al. (2015). SMRT Sequencing of Long Tandem Nucleotide Repeats in SCA10 Reveals Unique Insight of Repeat Expansion Structure. PLOS ONE 10: e0135906.

McFarland K.N., Liu J., Landrian I., Zeng D., Raskin S., Moscovich M., et al. (2014). Repeat interruptions in spinocerebellar ataxia type 10 expansions are strongly associated with epileptic seizures. Neurogenetics 15: 59–64.

McGinty R.J. and Mirkin S.M. (2018). Cis- and Trans-Modifiers of Repeat Expansions: Blending Model Systems with Human Genetics. Trends in Genetics 34: 448–465.

McInnis M.G. (1996). Anticipation: an old idea in new genes. American Journal of Human Genetics 59: 973–9.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297–303.

McLaren W., Gil L., Hunt S.E., Riat H.S., Ritchie G.R.S., Thormann A., et al. (2016). The Ensembl Variant Effect Predictor. Genome Biology 17: 122.

McNally J.R. and O'Brien P.J. (2017). Kinetic analyses of single-stranded break repair by human DNA ligase III isoforms reveal biochemical differences from DNA ligase I. Journal of Biological Chemistry 292: 15870–15879.

Mehrabi N.F., Singh-Bains M., Waldvogel H., and Faull R. (2016). Cortico-Basal Ganglia Interactions in Huntington's Disease. Annals of Neurodegenerative Disorders 1: 1007.

Mende-Mueller L.M., Toneff T., Hwang S.-R., Chesselet M.-F., and Hook V.Y.H. (2001). Tissue-Specific Proteolysis of Huntingtin (htt) in Human Brain: Evidence of Enhanced Levels of N- and C-Terminal htt Fragments in Huntington's Disease Striatum. The Journal of Neuroscience 21: 1830–1837.

Menon R.P., Nethisinghe S., Faggiano S., Vannocci T., Rezaei H., Pemble S., et al. (2013). The role of interruptions in polyQ in the pathology of SCA1. PLOS Genetics 9: e1003648.

Messias E.L., Chen C.-Y., and Eaton W.W. (2007). Epidemiology of schizophrenia: review of findings and myths. The Psychiatric Clinics of North America 30: 323–38.

Mestre T.A., van Duijn E., Davis A.M., Bachoud-Lévi A.-C., Busse M., Anderson K.E., et al. (2016). Rating scales for behavioral symptoms in Huntington's disease: Critique and recommendations. Movement Disorders 31: 1466–1478.

Metsu S., Rainger J.K., Debacker K., Bernhard B., Rooms L., Grafodatskaya D., et al. (2014a). A CGG-repeat expansion mutation in ZNF713 causes FRA7A: association with autistic spectrum disorder in two families. Human Mutation 35: 1295–300.

Metsu S., Rooms L., Rainger J., Taylor M.S., Bengani H., Wilson D.I., et al. (2014b). FRA2A is a CGG repeat expansion associated with silencing of AFF3. PLOS Genetics 10: e1004242.

Metzger S., Rong J., Nguyen H.-P., Cape A., Tomiuk J., Soehn A.S., et al. (2008). Huntingtin-associated protein-1 is a modifier of the age-at-onset of Huntington's disease. Human Molecular Genetics 17: 1137–46.

de Mezer M., Wojciechowska M., Napierala M., Sobczak K., and Krzyzosiak W.J. (2011). Mutant CAG repeats of Huntingtin transcript fold into hairpins, form nuclear foci and are targets for RNA interference. Nucleic Acids Research 39: 3852–63.

Migliore S., Jankovic J., and Squitieri F. (2019). Genetic Counseling in Huntington's Disease: Potential New Challenges on Horizon? Frontiers in Neurology 10: 453.

Miller J.P., Holcomb J., Al-Ramahi I., de Haro M., Gafni J., Zhang N., et al. (2010). Matrix metalloproteinases are modifiers of huntingtin proteolysis and toxicity in Huntington's disease. Neuron 67: 199–212.

Miller N.J., Schick K., Timchenko N., Harrison E., and Roesler W.J. (2016). The Glutamine-Alanine Repeat Domain of TCERG1 is Required for the Inhibition of the Growth Arrest Activity of C/EBPα. Journal of Cellular Biochemistry 117: 612–20.

Milne I., Stephen G., Bayer M., Cock P.J.A., Pritchard L., Cardle L., et al. (2013). Using Tablet for visual exploration of second-generation sequencing data. Briefings in BIoinformatics 14: 193–202.

Misiura M.B., Ciarochi J., Vaidya J., Bockholt J., Johnson H.J., Calhoun V.D., et al. (2019). Apathy Is Related to Cognitive Control and Striatum Volumes in Prodromal Huntington's Disease. Journal of the International Neuropsychological Society 25: 462–469.

Misiura M.B., Lourens S., Calhoun V.D., Long J., Bockholt J., Johnson H., et al. (2017). Cognitive Control, Learning, and Clinical Motor Ratings Are Most Highly Associated with Basal Ganglia Brain Volumes in the Premanifest Huntington's Disease Phenotype. Journal of the International Neuropsychological Society 23: 159–170.

Mistry V., Bockett N.A., Levine A.P., Mirza M.M., Hunt K.A., Ciclitira P.J., et al. (2015). Exome sequencing of 75 individuals from multiply affected coeliac families and large scale resequencing follow up. PLOS ONE 10: e0116845.

Mitas M., Yu A., Dill J., Kamp T.J., Chambers E.J., and Haworth I.S. (1995). Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: (CTG)15. Nucleic Acids Research 23: 1050–9.

Mitchell A.L., Attwood T.K., Babbitt P.C., Blum M., Bork P., Bridge A., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research 47: D351–D360.

Mizuguchi T., Toyota T., Adachi H., Miyake N., Matsumoto N., and Miyatake S. (2019). Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. Journal of Human Genetics 64: 191–197.

Mochel F., Charles P., Seguin F., Barritault J., Coussieu C., Perin L., et al. (2007). Early Energy Deficit in Huntington Disease: Identification of a Plasma Biomarker Traceable during Disease Progression. PLOS ONE 2: e647.

Mohyuddin A., Ayub Q., Siddiqi S., Carvalho-Silva D.R., Mazhar K., Rehman S., et al. (2004). Genetic instability in EBV-transformed lymphoblastoid cell lines. Biochimica et Biophysica Acta 1670: 81–3.

Molina-Calavita M., Barnat M., Elias S., Aparicio E., Piel M., and Humbert S. (2014). Mutant Huntingtin Affects Cortical Progenitor Cell Division and Development of the Mouse Neocortex. Journal of Neuroscience 34: 10034–10040.

Møllersen L., Rowe A.D., Illuzzi J.L., Hildrestrand G.A., Gerhold K.J., Tveterås L., et al. (2012). Neil1 is a genetic modifier of somatic and germline CAG trinucleotide repeat instability in R6/1 mice. Human Molecular Genetics 21: 4939–47.

Montecucco A., Rossi R., Levin D.S., Gary R., Park M.S., Motycka T.A., et al. (1998). DNA ligase I is recruited to sites of DNA replication by an interaction with proliferating cell nuclear antigen: identification of a common targeting mechanism for the assembly of replication factories. The EMBO Journal 17: 3786–95.

Montermini L., Andermann E., Labuda M., Richter A., Pandolfo M., Cavalcanti F., et al. (1997). The Friedreich ataxia GAA triplet repeat: premutation and normal alleles. Human Molecular Genetics 6: 1261–6.

Mootha V.V., Gong X., Ku H.-C., and Xing C. (2014). Association and familial segregation of CTG18.1 trinucleotide repeat expansion of TCF4 gene in Fuchs' endothelial corneal dystrophy. Investigative Ophthalmology & Visual Science 55: 33–42.

Morales F., Vásquez M., Santamaría C., Cuenca P., Corrales E., and Monckton D.G. (2016). A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. DNA Repair 40: 57–66.

Moreno-Küstner B., Martín C., and Pastor L. (2018). Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. PLOS ONE 13: e0195687.

Mori K., Weng S.-M., Arzberger T., May S., Rentzsch K., Kremmer E., et al. (2013). The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. Science 339: 1335–8.

Mu W., Lu H.-M., Chen J., Li S., and Elliott A.M. (2016). Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. The Journal of Molecular Diagnostics 18: 923–932.

Muona M., Berkovic S.F., Dibbens L.M., Oliver K.L., Maljevic S., Bayly M.A., et al. (2015). A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. Nature Genetics 47: 39–46.

Musova Z., Mazanec R., Krepelova A., Ehler E., Vales J., Jaklova R., et al. (2009). Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. American journal of medical genetics Part A 149A: 1365–74.

Myers R.H., MacDonald M.E., Koroshetz W.J., Duyao M.P., Ambrose C.M., Taylor S.A., et al. (1993). De novo expansion of a (CAG)n repeat in sporadic Huntington’s disease. Nature Genetics 5: 168–73.

Mykowska A., Sobczak K., Wojciechowska M., Kozlowski P., and Krzyzosiak W.J. (2011). CAG repeats mimic CUG repeats in the misregulation of alternative splicing. Nucleic Acids Research 39: 8938–51.

Nagasaki M., Yasuda J., Katsuoka F., Nariai N., Kojima K., Kawai Y., et al. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nature Communications 6: 8018.

Nahhas F.A., Garbern J., Krajewski K.M., Roa B.B., and Feldman G.L. (2005). Juvenile onset Huntington disease resulting from a very large maternal expansion. American Journal of Medical Genetics Part A 137A: 328–31.

Nakajima E., Orimo H., Ikejima M., and Shimada T. (1995). Nine-bp repeat polymorphism in exon 1 of the hMSH3 gene. The Japanese Journal of Human Genetics 40: 343–5.

Nakamura K., Jeong S.Y., Uchihara T., Anno M., Nagashima K., Nagashima T., et al. (2001). SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Human Molecular Genetics 10: 1441–8.

Nakatani R., Nakamori M., Fujimura H., Mochizuki H., and Takahashi M.P. (2015). Large expansion of CTG•CAG repeats is exacerbated by MutSβ in human cells. Scientific Reports 5: 11020.

Napierala M., Bacolla A., and Wells R.D. (2005). Increased negative superhelical density in vivo enhances the genetic instability of triplet repeat sequences. The Journal of Biological Chemistry 280: 37366–76.

Nasir J., Floresco S.B., O'Kusky J.R., Diewert V.M., Richman J.M., Zeisler J., et al. (1995). Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. Cell 81: 811–823.

Neale B.M., Kou Y., Liu L., Ma'ayan A., Samocha K.E., Sabo A., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242–5.

Negrette A. (1955). Corea de Huntington: Estudio de Una Sola Familia a Traves de Varias Genereaciones (Universidad de Zulia, Maracaibo, Venezuela).

Neil A.J., Kim J.C., and Mirkin S.M. (2017). Precarious maintenance of simple DNA repeats in eukaryotes. BioEssays 39: 1700077.

Neitzel H. (1986). A routine method for the establishment of permanent growing lymphoblastoid cell lines. Human Genetics 73: 320–6.

Neto J.L., Lee J.-M., Afridi A., Gillis T., Guide J.R., Dempsey S., et al. (2017). Genetic Contributors to Intergenerational CAG Repeat Instability in Huntington's Disease Knock-In Mice. Genetics 205: 503–516.

Neueder A., Dumas A.A., Benjamin A.C., and Bates G.P. (2018). Regulatory mechanisms of incomplete huntingtin mRNA splicing. Nature Communications 9: 3955.

Neueder A., Landles C., Ghosh R., Howland D., Myers R.H., Faull R.L.M., et al. (2017). The pathogenic exon 1 HTT protein is produced by incomplete splicing in Huntington's disease patients. Scientific reports 7: 1307.

Ng P.C. and Henikoff S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research 31: 3812–4.

Ng S.B., Bigham A.W., Buckingham K.J., Hannibal M.C., McMillin M.J., Gildersleeve H.I., et al. (2010a). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature Genetics 42: 790–3.

Ng S.B., Buckingham K.J., Lee C., Bigham A.W., Tabor H.K., Dent K.M., et al. (2010b). Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics 42: 30–5.

Ng S.B., Turner E.H., Robertson P.D., Flygare S.D., Bigham A.W., Lee C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272–6.

Niccolini F., Pagano G., Fusar-Poli P., Wood A., Mrzljak L., Sampaio C., and Politis M. (2018). Striatal molecular alterations in HD gene carriers: a systematic review and meta-analysis of PET studies. Journal of Neurology, Neurosurgery, and Psychiatry 89: 185–196.

Nickles D., Madireddy L., Yang S., Khankhanian P., Lincoln S., Hauser S.L., et al. (2012). In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing. BMC Genomics 13: 477.

Nicolas G., Devys D., Goldenberg A., Maltête D., Hervé C., Hannequin D., and Guyant-Maréchal L. (2011). Juvenile Huntington disease in an 18-month-old boy revealed by global developmental delay and reduced cerebellar volume. American Journal of Medical Genetics Part A 155: 815–818.

Nissen S.K., Christiansen M., Helleberg M., Kjær K., Jørgensen S.E., Gerstoft J., et al. (2018). Whole Exome Sequencing of HIV-1 long-term non-progressors identifies rare variants in genes encoding innate immune sensors and signaling molecules. Scientific Reports 8: 15253.

Nolin S.L., Glicksman A., Ersalesi N., Dobkin C., Brown W.T., Cao R., et al. (2015). Fragile X full mutation expansions are inhibited by one or more AGG interruptions in premutation carriers. Genetics in Medicine 17: 358–64.

Nolin S.L., Sah S., Glicksman A., Sherman S.L., Allen E., Berry-Kravis E., et al. (2013). Fragile X AGG analysis provides new risk predictions for 45-69 repeat alleles. American Journal of Medical Genetics Part A 161A: 771–8.

Obmolova G., Ban C., Hsieh P., and Yang W. (2000). Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. Nature 407: 703–10.

Oepen G., Clarenbach P., and Thoden U. (1981). Disturbance of eye movements in Huntington's chorea. Archiv fur Psychiatrie und Nervenkrankheiten 229: 205–13.

Okbay A., Beauchamp J.P., Fontana M.A., Lee J.J., Pers T.H., Rietveld C.A., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533: 539–42.

Okubo M., Doi H., Fukai R., Fujita A., Mitsuhashi S., Hashiguchi S., et al. (2019). GGC repeat expansion of NOTCH2NLC in adult patients with leukoencephalopathy. Annals of Neurology ana.25586.

Okun M.S. and Thommi N. (2004). Americo Negrette (1924 to 2003): Diagnosing Huntington disease in Venezuela. Neurology 63: 340–343.

Omi N., Tokuda Y., Ikeda Y., Ueno M., Mori K., Sotozono C., et al. (2017). Efficient and reliable establishment of lymphoblastoid cell lines by Epstein-Barr virus transformation from a limited amount of peripheral blood. Scientific Reports 7: 43833.

Ooi J., Langley S.R., Xu X., Utami K.H., Sim B., Huang Y., et al. (2019). Unbiased Profiling of Isogenic Huntington Disease hPSC-Derived CNS and Peripheral Cells Reveals Strong Cell-Type Specificity of CAG Length Effects. Cell Reports 26: 2494-2508.e7.

Orr H.T., Chung M., Banfi S., Kwiatkowski T.J., Servadio A., Beaudet A.L., et al. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nature Genetics 4: 221–226.

Ortega Z. and Lucas J.J. (2014). Ubiquitin-proteasome system involvement in Huntington's disease. Frontiers in Molecular Neuroscience 7: 77.

Orth M., Handley O.J., Schwenke C., Dunnett S.B., Craufurd D., Ho A.K., et al. (2010). Observing Huntington's Disease: the European Huntington's Disease Network's REGISTRY. PLOS Currents 2: RRN1184.

Orth M. and Schwenke C. (2011). Age-at-onset in Huntington disease. PLOS Currents 3: RRN1258.

Owen B.A.L., Yang Z., Lai M., Gajec M., Gajek M., Badger J.D., et al. (2005). (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. Nature Structural & Molecular Biology 12: 663–70.

Palombo F., Gallinari P., Iaccarino I., Lettieri T., Hughes M., D'Arrigo A., et al. (1995). GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. Science 268: 1912–1914.

Palombo F., Iaccarino I., Nakajima E., Ikejima M., Shimada T., and Jiricny J. (1996). hMutSβ, a heterodimer of hMSH2 and hMSH3, binds to insertion/deletion loops in DNA. Current Biology 6: 1181–1184.

Panarella M. and Burkett K.M. (2019). A Cautionary Note on the Effects of Population Stratification Under an Extreme Phenotype Sampling Design. Frontiers in Genetics 10: 398.

Panigrahi G.B., Slean M.M., Simard J.P., Gileadi O., and Pearson C.E. (2010). Isolated short CTG/CAG DNA slip-outs are repaired efficiently by hMutSβ, but clustered slip-outs are poorly repaired. Proceedings of the National Academy of Sciences of the United States of America 107: 12593–12598.

Panov A. V, Gutekunst C.-A., Leavitt B.R., Hayden M.R., Burke J.R., Strittmatter W.J., and Greenamyre J.T. (2002). Early mitochondrial calcium defects in Huntington's disease are a direct effect of polyglutamines. Nature Neuroscience 5: 731–6.

Papoutsi M., Labuschagne I., Tabrizi S.J., and Stout J.C. (2014). The cognitive burden in Huntington's disease: pathology, phenotype, and mechanisms of compensation. Movement

Disorders 29: 673–83.

Papp K. V., Snyder P.J., Mills J.A., Duff K., Westervelt H.J., Long J.D., et al. (2013). Measuring Executive Dysfunction Longitudinally and in Relation to Genetic Burden, Brain Volumetrics, and Depression in Prodromal Huntington Disease. Archives of Clinical Neuropsychology 28: 156–168.

Parent A. and Hazrati L.-N. (1993). Anatomical aspects of information processing in primate basal ganglia. Trends in Neurosciences 16: 111–116.

Parent A. and Hazrati L.N. (1995). Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. Brain Research Reviews 20: 91–127.

Parker J.G., Marshall J.D., Ahanonu B., Wu Y.-W., Kim T.H., Grewe B.F., et al. (2018). Diametric neural ensemble dynamics in parkinsonian and dyskinetic states. Nature 557: 177–182.

Parrish J.E., Oostra B.A., Verkerk A.J.M.H., Richards C.S., Reynolds J., Spikes A.S., et al. (1994). Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE. Nature Genetics 8: 229–235.

Pascal J.M., O'Brien P.J., Tomkinson A.E., and Ellenberger T. (2004). Human DNA ligase I completely encircles and partially unwinds nicked DNA. Nature 432: 473–8.

Pascu A.M., Ifteni P., Teodorescu A., Burtea V., and Correll C.U. (2015). Delayed identification and diagnosis of Huntington's disease due to psychiatric symptoms. International Journal of Mental Health Systems 9: 33.

Patterson N., Price A.L., and Reich D. (2006). Population structure and eigenanalysis. PLOS Genetics 2: e190.

Paulsen J.S., Hoth K.F., Nehl C., and Stierman L. (2005a). Critical periods of suicide risk in Huntington's disease. The American Journal of Psychiatry 162: 725–31.

Paulsen J.S., Langbehn D.R., Stout J.C., Aylward E., Ross C.A., Nance M., et al. (2008). Detection of Huntington's disease decades before diagnosis: the Predict-HD study. Journal of Neurology, Neurosurgery, and Psychiatry 79: 874–80.

Paulsen J.S. and Long J.D. (2014). Onset of Huntington's disease: Can it be purely cognitive? Movement Disorders 29: 1342–1350.

Paulsen J.S., Long J.D., Johnson H.J., Aylward E.H., Ross C.A., Williams J.K., et al. (2014). Clinical and Biomarker Changes in Premanifest Huntington Disease Show Trial Feasibility: A Decade of the PREDICT-HD Study. Frontiers in Aging Neuroscience 6: 78.

Paulsen J.S., Miller A.C., Hayes T., and Shaw E. (2017). Cognitive and behavioral changes in Huntington disease before diagnosis. Handbook of Clinical Neurology 144: 69–91.

Paulsen J.S., Nehl C., Hoth K.F., Kanz J.E., Benjamin M., Conybeare R., et al. (2005b). Depression and Stages of Huntington’s Disease. The Journal of Neuropsychiatry and Clinical Neurosciences 17: 496–502.

Paulsen J.S., Ready R.E., Hamilton J.M., Mega M.S., and Cummings J.L. (2001). Neuropsychiatric aspects of Huntington’s disease. Journal of Neurology, Neurosurgery, and Psychiatry 71: 310–4.

Paulsen J.S., Wang C., Duff K., Barker R., Nance M., Beglinger L., et al. (2010). Challenges assessing clinical endpoints in early Huntington disease. Movement Disorders 25: 2595–603.

Pearson C.E., Eichler E.E., Lorenzetti D., Kramer S.F., Zoghbi H.Y., Nelson D.L., and Sinden R.R. (1998). Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. Biochemistry 37: 2701–8.

Pearson C.E., Tam M., Wang Y.-H., Montgomery S.E., Dar A.C., Cleary J.D., and Nichol K. (2002). Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. Nucleic Acids Research 30: 4534–47.

Peavy G.M., Jacobson M.W., Goldstein J.L., Hamilton J.M., Kane A., Gamst A.C., et al. (2010). Cognitive and functional decline in Huntington’s disease: Dementia criteria revisited. Movement Disorders 25: 1163–1169.

Pêcheux C., Mouret J.F., Dürr A., Agid Y., Feingold J., Brice A., et al. (1995). Sequence analysis of the CCG polymorphic region adjacent to the CAG triplet repeat of the HD gene in normal and HD chromosomes. Journal of Medical Genetics 32: 399–400.

Pedersen B.S. and Quinlan A.R. (2017). Who’s Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. American Journal of Human Genetics 100: 406–413.

Pekmezovic T., Svetel M., Maric J., Dujmovic-Basuroski I., Dragasevic N., Keckarevic M., et al. (2007). Survival of Huntington’s disease patients in Serbia: longer survival in female patients. European Journal of Epidemiology 22: 523–6.

Peloso G.M., Rader D.J., Gabriel S., Kathiresan S., Daly M.J., and Neale B.M. (2016). Phenotypic extremes in rare variant study designs. European Journal of Medical Genetics 24: 924–30.

Pennell S., Déclais A.-C., Li J., Haire L.F., Berg W., Saldanha J.W., et al. (2014). FAN1 activity on asymmetric repair intermediates is mediated by an atypical monomeric virus-type replication-repair nuclease domain. Cell Reports 8: 84–93.

Perez-Riba A. and Itzhaki L.S. (2019). The tetratricopeptide-repeat motif is a versatile platform that enables diverse modes of molecular recognition. Current Opinion in Structural Biology 54: 43–49.

Pešović J., Perić S., Brkušanin M., Brajušković G., Rakočević-Stojanović V., and Savić-Pavićević D. (2018). Repeat Interruptions Modify Age at Onset in Myotonic Dystrophy Type 1 by Stabilizing DMPK Expansions in Somatic Cells. Frontiers in Genetics 9: 601.

Petersen B.-S., Fredrich B., Hoeppner M.P., Ellinghaus D., and Franke A. (2017). Opportunities and challenges of whole-genome and -exome sequencing. BMC Genetics 18: 14.

Pflanz S., Besson J.A., Ebmeier K.P., and Simpson S. (1991). The clinical manifestation of mental disorder in Huntington's disease: a retrospective case record study of disease progression. Acta Psychiatrica Scandinavica 83: 53–60.

Piccinelli M. and Wilkinson G. (2000). Gender differences in depression. Critical review. The British Journal of Psychiatry 177: 486–92.

Pinto R.M., Dragileva E., Kirby A., Lloret A., Lopez E., St Claire J., et al. (2013). Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. PLOS Genetics 9: e1003930.

Pizzolato J., Mukherjee S., Schärer O.D., and Jiricny J. (2015). FANCD2-associated nuclease 1, but not exonuclease 1 or flap endonuclease 1, is able to unhook DNA interstrand cross-links in vitro. The Journal of Biological Chemistry 290: 22602–11.

Pliner H.A., Mann D.M., and Traynor B.J. (2014). Searching for Grendel: origin and global spread of the C9ORF72 repeat expansion. Acta Neuropathologica 127: 391–396.

Podlutsky A.J., Dianova I.I., Podust V.N., Bohr V.A., and Dianov G.L. (2001). Human DNA polymerase beta initiates DNA synthesis during long-patch repair of reduced AP sites in DNA. The EMBO Journal 20: 1477–82.

Podolsky S., Leopold N., and Sax D. (1972). Increased frequency of diabetes mellitus in patients with Huntington's chorea. The Lancet 299: 1356–1359.

Podolsky S. and Leopold N.A. (1977). Abnormal Glucose Tolerance and Arginine Tolerance Tests in Huntington's Disease. Gerontology 23: 55–63.

Polyzos A.A. and McMurray C.T. (2017). Close encounters: Moving along bumps, breaks, and bubbles on expanded trinucleotide tracts. DNA Repair 56: 144–155.

Porro A., Berti M., Pizzolato J., Bologna S., Kaden S., Saxer A., et al. (2017). FAN1 interaction with ubiquitylated PCNA alleviates replication stress and preserves genomic integrity independently of BRCA2. Nature Communications 8: 1073.

Potter N.T., Spector E.B., and Prior T.W. (2004). Technical Standards and Guidelines for Huntington Disease Testing. Genetics in Medicine 6: 61–65.

Pouladi M.A., Stanek L.M., Xie Y., Franciosi S., Southwell A.L., Deng Y., et al. (2012). Marked differences in neurochemistry and aggregates despite similar behavioural and neuropathological features of Huntington disease in the full-length BACHD and YAC128 mice. Human Molecular Genetics 21: 2219–32.

Price A.L., Zaitlen N.A., Reich D., and Patterson N. (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics 11: 459–63.

Pridmore S.A. and Adams G.C. (1991). The Fertility of HD-Affected Individuals in Tasmania. Australian & New Zealand Journal of Psychiatry 25: 262–264.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559–75.

Purcell S.M., Moran J.L., Fromer M., Ruderfer D., Solovieff N., Roussos P., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506: 185–90.

Qiu J., Qian Y., Chen V., Guan M.X., and Shen B. (1999). Human exonuclease 1 functionally complements its yeast homologues in DNA recombination, RNA primer removal, and mutation avoidance. The Journal of Biological Chemistry 274: 17893–900.

Quan F., Janas J., and Popovich B.W. (1995). A novel CAG repeat configuration in the SCA1 gene: implications for the molecular diagnostics of spinocerebellar ataxia type 1. Human Molecular Genetics 4: 2411–2413.

Quang D., Chen Y., and Xie X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 31: 761–3.

Quarrell O., O'Donovan K.L., Bandmann O., and Strong M. (2012). The Prevalence of Juvenile Huntington's Disease: A Review of the Literature and Meta-Analysis. PLOS Currents 4: e4f8606b742ef3.

Quinlan A.R. and Hall I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–2.

Van Raamsdonk J.M., Murphy Z., Selva D.M., Hamidizadeh R., Pearson J., Petersén Ǻsa, et al. (2007). Testicular degeneration in Huntington disease. Neurobiology of Disease 26: 512–

520.

Rafehi H., Szmulewicz D.J., Bennett M.F., Sobreira N.L.M., Pope K., Smith K.R., et al. (2019). Bioinformatics-Based Identification of Expanded Repeats: A Non-reference Intronic Pentamer Expansion in RFC1 Causes CANVAS. American Journal of Human Genetics 105: 151–165.

Raghavan N.S., Brickman A.M., Andrews H., Manly J.J., Schupf N., Lantigua R., et al. (2018). Whole-exome sequencing in 20,197 persons for rare variants in Alzheimer's disease. Annals of Clinical and Translational Neurology 5: 832–842.

Ramos E.M., Cerqueira J., Lemos C., Pinto-Basto J., Alonso I., and Sequeiros J. (2012a). Intergenerational instability in Huntington disease: Extreme repeat changes among 134 transmissions. Movement Disorders 27: 583–585.

Ramos E.M., Latourelle J.C., Lee J.-M.J.-H., Gillis T., Mysore J.S., Squitieri F., et al. (2012b). Population stratification may bias analysis of PGC-1α as a modifier of age at Huntington disease motor onset. Human Genetics 131: 1833–1840.

Ranen N.G., Stine O.C., Abbott M.H., Sherr M., Codori A.M., Franz M.L., et al. (1995). Anticipation and instability of IT-15 (CAG)n repeats in parent-offspring pairs with Huntington disease. American Journal of Human Genetics 57: 593–602.

Rao A.K., Mazzoni P., Wasserman P., and Marder K. (2011). Longitudinal Change in Gait and Motor Function in Pre-manifest Huntington's Disease. PLOS Currents 3: RRN1268.

Räschle M., Marra G., Nyström-Lahti M., Schär P., and Jiricny J. (1999). Identification of hMutLβ, a Heterodimer of hMLH1 and hPMS1. Journal of Biological Chemistry 274: 32368–32375.

Rawlins M.D., Wexler N.S., Wexler A.R., Tabrizi S.J., Douglas I., Evans S.J.W., and Smeeth L. (2016). The Prevalence of Huntington's Disease. Neuroepidemiology 46: 144–153.

Raymond L.A. (2017). Striatal synaptic dysfunction and altered calcium regulation in Huntington disease. Biochemical and Biophysical Research Communications 483: 1051–1062.

Reading S.A.J., Dziorny A.C., Peroutka L.A., Schreiber M., Gourley L.M., Yallapragada V., et al. (2004). Functional brain changes in presymptomatic Huntington's disease. Annals of Neurology 55: 879–83.

Ready R.E., Mathews M., Leserman A., and Paulsen J.S. (2008). Patient and caregiver quality of life in Huntington's disease. Movement Disorders 23: 721–726.

Reddy K., Schmidt M.H.M., Geist J.M., Thakkar N.P., Panigrahi G.B., Wang Y.-H., and

Pearson C.E. (2014). Processing of double-R-loops in (CAG)·(CTG) and C9orf72 (GGGGCC)·(GGCCCC) repeats causes instability. Nucleic Acids Research 42: 10473–87.

Reddy K., Tam M., Bowater R.P., Barber M., Tomlinson M., Nichol Edamura K., et al. (2011). Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. Nucleic Acids Research 39: 1749–62.

Reedeker N., Bouwens J.A., Giltay E.J., Le Mair S.E., Roos R.A.C., van der Mast R.C., and van Duijn E. (2012). Irritability in Huntington's disease. Psychiatry Research 200: 813–8.

Rees E., Carrera N., Morgan J., Hambridge K., Escott-Price V., Pocklington A.J., et al. (2019). Targeted Sequencing of 10,198 Samples Confirms Abnormalities in Neuronal Activity and Implicates Voltage-Gated Sodium Channels in Schizophrenia Pathogenesis. Biological Psychiatry 85: 554–562.

Reilmann R., Leavitt B.R., and Ross C.A. (2014). Diagnostic criteria for Huntington's disease based on natural history. Movement Disorders 29: 1335–41.

Reilmann R. and Schubert R. (2017). Motor outcome measures in Huntington disease clinical trials. Handbook of Clinical Neurology 144: 209–225.

Reiner A., Albin R.L., Anderson K.D., D'Amato C.J., Penney J.B., and Young A.B. (1988). Differential loss of striatal projection neurons in Huntington disease. Proceedings of the National Academy of Sciences of the United States of America 85: 5733–5737.

Renton A.E., Majounie E., Waite A., Simón-Sánchez J., Rollinson S., Gibbs J.R., et al. (2011). A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. Neuron 72: 257–268.

Rentzsch P., Witten D., Cooper G.M., Shendure J., and Kircher M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research 47: D886–D894.

Retshabile G., Mlotshwa B.C., Williams L., Mwesigwa S., Mboowa G., Huang Z., et al. (2018). Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana. American Journal of Human Genetics 102: 731–743.

Ribaï P., Nguyen K., Hahn-Barma V., Gourfinkel-An I., Vidailhet M., Legout A., et al. (2007). Psychiatric and Cognitive Difficulties as Indicators of Juvenile Huntington Disease Onset in 29 Patients. Archives of Neurology 64: 813.

Ribeiro D.T., Madzak C., Sarasin A., Di Mascio P., Sies H., and Menck C.F. (1992). Singlet oxygen induced DNA damage and mutagenicity in a single-stranded SV40-based shuttle vector. Photochemistry and Photobiology 55: 39–45.

Richfield E.K., Maguire-Zeiss K.A., Vonkeman H.E., and Voorn P. (1995). Preferential loss of preproenkephalin versus preprotachykinin neurons from the striatum of Huntington's disease patients. Annals of Neurology 38: 852–861.

Ridley R.M., Frith C.D., Crow T.J., and Conneally P.M. (1988). Anticipation in Huntington's disease is inherited through the male line but may originate in the female. Journal of Medical Genetics 25: 589–95.

Rinaldi C., Salvatore E., Giordano I., De Matteis S., Tucci T., Cinzia V.R., et al. (2012). Predictors of survival in a Huntington's disease population from southern Italy. The Canadian Journal of Neurological Sciences 39: 48–51.

Rinne J.O., Rummukainen J., Paljärvi L., and Rinne U.K. (1989). Dementia in Parkinson's disease is related to neuronal loss in the medial substantia nigra. Annals of Neurology 26: 47–50.

Roach J.C., Glusman G., Smit A.F.A., Huff C.D., Hubley R., Shannon P.T., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636–9.

Robins Wahlin T.-B. (2007). To know or not to know: a review of behaviour and suicidal ideation in preclinical Huntington's disease. Patient Education and Counseling 65: 279–87.

Rocha N.P., Mwangi B., Gutierrez Candano C.A., Sampaio C., Furr Stimming E., and Teixeira A.L. (2018). The Clinical Picture of Psychosis in Manifest Huntington's Disease: A Comprehensive Analysis of the Enroll-HD Database. Frontiers in Neurology 9: 930.

Rolfsmeier M.L. and Lahue R.S. (2000). Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. Molecular and Cellular Biology 20: 173–80.

Rolls E.T. (2015). Limbic systems for emotion and for memory, but no single limbic system. Cortex 62: 119–57.

Roos R.A.C. (2010). Huntington's disease: a clinical review. Orphanet Journal of Rare Diseases 5: 40.

Rosas H.D., Koroshetz W.J., Chen Y.I., Skeuse C., Vangel M., Cudkowicz M.E., et al. (2003). Evidence for more widespread cerebral pathology in early HD: an MRI-based morphometric analysis. Neurology 60: 1615–20.

Rosas H.D., Liu A.K., Hersch S., Glessner M., Ferrante R.J., Salat D.H., et al. (2002). Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology 58: 695–701.

Rosenblatt A., Brinkman R.R., Liang K.Y., Almqvist E.W., Margolis R.L., Huang C.Y., et al. (2001). Familial influence on age of onset among siblings with Huntington disease. American Journal of Human Genetics 105: 399–403.

Ross C.A., Aylward E.H., Wild E.J., Langbehn D.R., Long J.D., Warner J.H., et al. (2014). Huntington disease: natural history, biomarkers and prospects for therapeutics. Nature Reviews Neurology 10: 204–216.

Rowe K.C., Paulsen J.S., Langbehn D.R., Duff K., Beglinger L.J., Wang C., et al. (2010). Self-paced timing detects and tracks change in prodromal Huntington disease. Neuropsychology 24: 435–442.

Ruano L., Melo C., Silva M.C., and Coutinho P. (2014). The Global Epidemiology of Hereditary Ataxia and Spastic Paraplegia: A Systematic Review of Prevalence Studies. Neuroepidemiology 42: 174–183.

De Rubeis S., He X., Goldberg A.P., Poultney C.S., Samocha K., Cicek A.E., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515: 209–15.

Rubinsztein D.C., Leggo J., Chiano M., Dodge A., Norbury G., Rosser E., and Craufurd D. (1997). Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. Proceedings of the National Academy of Sciences of the United States of America 94: 3872–6.

Rubinsztein D.C., Leggo J., Coles R., Almqvist E., Biancalana V., Cassiman J.J., et al. (1996). Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. American Journal of Human Genetics 59: 16–22.

Rué L., Bañez-Coronel M., Creus-Muncunill J., Giralt A., Alcalá-Vida R., Mentxaka G., et al. (2016). Targeting CAG repeat RNAs reduces Huntington's disease phenotype independently of huntingtin levels. The Journal of Clinical Investigation 126: 4319–4330.

Ruocco H.H., Bonilha L., Li L.M., Lopes-Cendes I., and Cendes F. (2008). Longitudinal analysis of regional grey matter loss in Huntington disease: effects of the length of the expanded CAG repeat. Journal of Neurology, Neurosurgery, and Psychiatry 79: 130–5.

Ruzo A., Croft G.F., Metzger J.J., Galgoczi S., Gerber L.J., Pellegrini C., et al. (2018). Chromosomal instability during neurogenesis in Huntington's disease. Development 145:.

Sackley C., Hoppitt T.J., Calvert M., Gill P., Eaton B., Yao G., and Pall H. (2011). Huntington's disease: current epidemiology and pharmacological management in UK primary care. Neuroepidemiology 37: 216–21.

Saft C., Andrich J.E., Brune N., Gencik M., Kraus P.H., Przuntek H., and Epplen J.T. (2004). Apolipoprotein E genotypes do not influence the age of onset in Huntington’s disease. Journal of Neurology, Neurosurgery, and Psychiatry 75: 1692–6.

Salk R.H., Hyde J.S., and Abramson L.Y. (2017). Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. Psychological Bulletin 143: 783–822.

Sánchez-Hernández N., Ruiz L., Sánchez-Álvarez M., Montes M., Macias M.J., Hernández-Munain C., and Suñé C. (2012). The FF4 and FF5 domains of transcription elongation regulator 1 (TCERG1) target proteins to the periphery of speckles. The Journal of Biological Chemistry 287: 17789–800.

Sandri T.L., Andrade F.A., Lidani K.C.F., Einig E., Boldt A.B.W., Mordmüller B., et al. (2019). Human collectin-11 (COLEC11) and its synergic genetic interaction with MASP2 are associated with the pathophysiology of Chagas Disease. PLOS Neglected Tropical Diseases 13: e0007324.

Sanpei K., Takano H., Igarashi S., Sato T., Oyake M., Sasaki H., et al. (1996). Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. Nature Genetics 14: 277–284.

Sathasivam K., Neueder A., Gipson T.A., Landles C., Benjamin A.C., Bondulich M.K., et al. (2013). Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. Proceedings of the National Academy of Sciences of the United States of America 110: 2366–70.

Sato N., Amino T., Kobayashi K., Asakawa S., Ishiguro T., Tsunemi T., et al. (2009). Spinocerebellar Ataxia Type 31 Is Associated with “Inserted” Penta-Nucleotide Repeats Containing (TGGAA)n. American Journal of Human Genetics 85: 544–557.

Satterstrom F.K., Kosmicki J.A., Wang J., Breen M.S., De Rubeis S., An J.-Y., et al. (2019). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. [pre-print, bioarchive: https://www.biorxiv.org/content/10.1101/484113v3].

Saudou F. and Humbert S. (2016). The Biology of Huntingtin. Neuron 89: 910–926.

Say M.J., Jones R., Scahill R.I., Dumas E.M., Coleman A., Santos R.C.D., et al. (2011). Visuomotor integration deficits precede clinical onset in Huntington’s disease. Neuropsychologia 49: 264–270.

Schafer C.M., Campbell N.G., Cai G., Yu F., Makarov V., Yoon S., et al. (2013). Whole exome sequencing reveals minimal differences between cell line and whole blood derived DNA. Genomics 102: 270–277.

Scherzinger E., Lurz R., Turmaine M., Mangiarini L., Hollenbach B., Hasenbank R., et al. (1997). Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. Cell 90: 549–58.

Schilling G., Becher M.W., Sharp A.H., Jinnah H.A., Duan K., Kotzuk J.A., et al. (1999). Intranuclear inclusions and neuritic aggregates in transgenic mice expressing a mutant N-terminal fragment of huntingtin. Human Molecular Genetics 8: 397–407.

Schilling J., Broemer M., Atanassov I., Duernberger Y., Vorberg I., Dieterich C., et al. (2019). Deregulated Splicing Is a Major Mechanism of RNA-Induced Toxicity in Huntington's Disease. Journal of Molecular Biology 431: 1869–1877.

Schmutte C., Marinescu R.C., Sadoff M.M., Guerrette S., Overhauser J., and Fishel R. (1998). Human exonuclease I interacts with the mismatch repair protein hMSH2. Cancer Research 58: 4537–42.

Schmutte C., Sadoff M.M., Shim K.-S.S., Acharya S., and Fishel R. (2001). The Interaction of DNA Mismatch Repair Proteins with Human Exonuclease I. The Journal of Biological Chemistry 276: 33011–33018.

Schwarze K., Buchanan J., Taylor J.C., and Wordsworth S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genetics in Medicine 20: 1122–1130.

Sciacca S., Favellato M., Madonna M., Metro D., Marano M., and Squitieri F. (2017). Early enteric neuron dysfunction in mouse and human Huntington disease. Parkinsonism & Related Disorders 34: 73–74.

Scott S., Collet J.-P., Baber U., Yang Y., Peter I., Linderman M., et al. (2016). Exome sequencing of extreme clopidogrel response phenotypes identifies B4GALT2 as a determinant of on-treatment platelet reactivity. Clinical Pharmacology & Therapeutics 100: 287–294.

Scrimgeour E.M. and Pfumojena J.W. (1992). Huntington disease in black Zimbabwean families living near the Mozambique border. American Journal of Medical Genetics 44: 762–766.

Seixas A.I., Loureiro J.R., Costa C., Ordóñez-Ugalde A., Marcelino H., Oliveira C.L., et al. (2017). A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1 , Mapping to SCA37 , Causes Spinocerebellar Ataxia. American Journal of Human Genetics

101: 87–103.

Semaka A., Collins J.A., and Hayden M.R. (2010). Unstable familial transmissions of Huntington disease alleles with 27-35 CAG repeats (intermediate alleles). American Journal of Medical Genetics Part B 153B: 314–320.

Semaka A. and Hayden M.R. (2014). Evidence-based genetic counselling implications for Huntington disease intermediate allele predictive test results. Clinical Genetics 85: 303–311.

Semaka A., Kay C., Doty C., Collins J.A., Bijlsma E.K., Richards F., et al. (2013). CAG size-specific risk estimates for intermediate allele repeat instability in Huntington disease. Journal of Medical Genetics 50: 696–703.

Sepers M.D. and Raymond L.A. (2014). Mechanisms of synaptic dysfunction and excitotoxicity in Huntington's disease. Drug Discovery Today 19: 990–6.

Seriola A., Spits C., Simard J.P., Hilven P., Haentjens P., Pearson C.E., and Sermon K. (2011). Huntington's and myotonic dystrophy hESCs: down-regulated trinucleotide repeat instability and mismatch repair machinery expression upon differentiation. Human Molecular Genetics 20: 176–85.

Sgroi S. and Tonini R. (2018). Opioidergic Modulation of Striatal Circuits, Implications in Parkinson's Disease and Levodopa Induced Dyskinesia. Frontiers in Neurology 9: 524.

Shelbourne P.F., Keller-McGandy C., Bi W.L., Yoon S.-R., Dubeau L., Veitch N.J., et al. (2007). Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Human Molecular Genetics 16: 1133–42.

Shereda R.D., Machida Y., and Machida Y.J. (2010). Human KIAA1018/FAN1 localizes to stalled replication forks via its ubiquitin-binding domain. Cell cycle 9: 3977–83.

Shirley M.D., Baugher J.D., Stevens E.L., Tang Z., Gerry N., Beiswanger C.M., et al. (2012). Chromosomal variation in lymphoblastoid cell lines. Human Mutation 33: 1075–86.

Shiwach R.S. and Patel V. (1993). Aggressive behaviour in Huntington's disease: a cross-sectional study in a nursing home population. Behavioural Neurology 6: 43–7.

Sie L., Loong S., and Tan E.K. (2009). Utility of lymphoblastoid cell lines. Journal of Neuroscience Research 87: 1953–9.

Sim N.-L., Kumar P., Hu J., Henikoff S., Schneider G., and Ng P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research 40: W452-7.

Simard O., Grégoire M.-C., Arguin M., Brazeau M.-A., Leduc F., Marois I., et al. (2014). Instability of trinucleotidic repeats during chromatin remodeling in spermatids. Human

Mutation 35: 1280–4.

Sipilä J.O.T., Hietala M., Siitonen A., Päivärinta M., and Majamaa K. (2015). Epidemiology of Huntington's disease in Finland. Parkinsonism & Related Disorders 21: 46–49.

Sipilä J.O.T., Soilu-Hänninen M., and Majamaa K. (2016). Comorbid epilepsy in Finnish patients with adult-onset Huntington's disease. BMC Neurology 16: 24.

Sitek E.J., Thompson J.C., Craufurd D., and Snowden J.S. (2014). Unawareness of deficits in Huntington's disease. Journal of Huntington's disease 3: 125–35.

Smogorzewska A., Desetty R., Saito T.T., Schlabach M., Lach F.P., Sowa M.E., et al. (2010). A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. Molecular Cell 39: 36–47.

Snaith R.P., Constantopoulos A.A., Jardine M.Y., and McGuffin P. (1978). A clinical scale for the self-assessment of irritability. The British Journal of Psychiatry 132: 164–71.

Snell R., MacMillan J., Cheadle J., Fenton I., Lazarou L., Davies P., et al. (1993). Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nature Genetics 4: 393–397.

Sobczak K. and Krzyzosiak W.J. (2005). CAG repeats containing CAA interruptions form branched hairpin structures in spinocerebellar ataxia type 2 transcripts. The Journal of Biological Chemistry 280: 3898–910.

Sobczak K., de Mezer M., Michlewski G., Krol J., and Krzyzosiak W.J. (2003). RNA structure of trinucleotide repeats associated with human neurological diseases. Nucleic Acids Research 31: 5469–82.

Sone J., Mitsuhashi S., Fujita A., Mizuguchi T., Hamanaka K., Mori K., et al. (2019). Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nature Genetics 51: 1215–1221.

Sørensen S.A. and Fenger K. (1992). Causes of death in patients with Huntington's disease and in unaffected first degree relatives. Journal of Medical Genetics 29: 911–4.

De Souza J., Jones L.A., and Rickards H. (2010). Validation of self-report depression rating scales in Huntington's disease. Movement Disorders 25: 91–6.

La Spada A.R., Wilson E.M., Lubahn D.B., Harding A.E., and Fischbeck K.H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352: 77–9.

Spoendlin M., Moch H., Brunner F., Brunner W., Burger H.R., Kiss D., et al. (1995). Karyomegalic interstitial nephritis: further support for a distinct entity and evidence for a genetic defect. American Journal of Kidney Diseases 25: 242–52.

Squitieri F., Frati L., Ciarmiello A., Lastoria S., and Quarrell O. (2006). Juvenile Huntington's disease: Does a dosage-effect pathogenic mechanism differ from the classical adult disease? Mechanisms of Ageing and Development 127: 208–212.

Squitieri F., Gellera C., Cannella M., Mariotti C., Cislaghi G., Rubinsztein D.C., et al. (2003). Homozygosity for CAG mutation in Huntington disease is associated with a more severe clinical course. Brain 126: 946–55.

Squitieri F., Griguoli A., Capelli G., Porcellini A., and D'Alessio B. (2016). Epidemiology of Huntington disease: first post- HTT gene analysis of prevalence in Italy. Clinical Genetics 89: 367–370.

Starr P.A., Kang G.A., Heath S., Shimamoto S., and Turner R.S. (2008). Pallidal neuronal discharge in Huntington's disease: support for selective loss of striatal cells originating the indirect pathway. Experimental Neurology 211: 227–33.

Steffan J.S. (2010). Does Huntingtin play a role in selective macroautophagy? Cell Cycle 9: 3401–3413.

Steffan J.S., Kazantsev A., Spasic-Boskovic O., Greenwald M., Zhu Y.Z., Gohler H., et al. (2000). The Huntington's disease protein interacts with p53 and CREB-binding protein and represses transcription. Proceedings of the National Academy of Sciences of the United States of America 97: 6763–8.

Stevanin G., Le Guern E., Ravisé N., Chneiweiss H., Dürr A., Cancel G., et al. (1994). A third locus for autosomal dominant cerebellar ataxia type I maps to chromosome 14q24.3-qter: evidence for the existence of a fourth locus. American Journal of Human Genetics 54: 11–20.

Stout J.C., Jones R., Labuschagne I., O'Regan A.M., Say M.J., Dumas E.M., et al. (2012). Evaluation of longitudinal 12 and 24 month cognitive outcomes in premanifest and early Huntington's disease. Journal of Neurology, Neurosurgery, and Psychiatry 83: 687–694.

Stout J.C., Paulsen J.S., Queller S., Solomon A.C., Whitlock K.B., Campbell J.C., et al. (2011). Neurocognitive signs in prodromal Huntington disease. Neuropsychology 25: 1–14.

Su X.A. and Freudenreich C.H. (2017). Cytosine deamination and base excision repair cause R-loop–induced CAG repeat fragility and instability in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 114: E8392–E8401.

Subramanian J., Vijayakumar S., Tomkinson A.E., and Arnheim N. (2005). Genetic instability induced by overexpression of DNA ligase I in budding yeast. Genetics 171: 427–41.

Suñé C., Hayashi T., Liu Y., Lane W.S., Young R.A., and Garcia-Blanco M.A. (1997). CA150, a nuclear protein associated with the RNA polymerase II holoenzyme, is involved in Tat-activated human immunodeficiency virus type 1 transcription. Molecular and Cellular Biology 17: 6029–39.

Surmeier D.J., Ding J., Day M., Wang Z., and Shen W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends in Neurosciences 30: 228–35.

Suwinski P., Ong C., Ling M.H.T., Poh Y.M., Khan A.M., and Ong H.S. (2019). Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics. Frontiers in Genetics 10: 49.

Swami M., Hendricks A.E., Gillis T., Massood T., Mysore J., Myers R.H., and Wheeler V.C. (2009). Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. Human Molecular Genetics 18: 3039–47.

Sznajder Ł.J. and Swanson M.S. (2019). Short Tandem Repeat Expansions and RNA-Mediated Pathogenesis in Myotonic Dystrophy. International Journal of Molecular Sciences 20: 3365.

Tabrizi S.J., Langbehn D.R., Leavitt B.R., Roos R.A., Durr A., Craufurd D., et al. (2009). Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. The Lancet Neurology 8: 791–801.

Tabrizi S.J., Reilmann R., Roos R.A.C., Durr A., Leavitt B., Owen G., et al. (2012). Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. The Lancet Neurology 11: 42–53.

Tabrizi S.J., Scahill R.I., Durr A., Roos R.A., Leavitt B.R., Jones R., et al. (2011). Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. The Lancet Neurology 10: 31–42.

Tabrizi S.J., Scahill R.I., Owen G., Durr A., Leavitt B.R., Roos R.A., et al. (2013). Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. The Lancet Neurology 12: 637–49.

Taherzadeh-Fard E., Saft C., Andrich J., Wieczorek S., and Arning L. (2009). PGC-1alpha as modifier of onset age in Huntington disease. Molecular Neurodegeneration 4: 10.

Takano H. and Gusella J.F. (2002). The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor. BMC Neuroscience 3: 15.

Takata A. (2019). Estimating contribution of rare non-coding variants to neuropsychiatric disorders. Psychiatry and Clinical Neurosciences 73: 2–10.

Tallaksen-Greene S.J., Ordway J.M., Crouse A.B., Jackson W.S., Detloff P.J., and Albin R.L. (2003). Hprt(CAG)146 mice: Age of onset of behavioral abnormalities, time course of neuronal intranuclear inclusion accumulation, neurotransmitter marker alterations, mitochondrial function markers, and susceptibility to 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridi. The Journal of Comparative Neurology 465: 205–219.

Tang H., Kirkness E.F., Lippert C., Biggs W.H., Fabani M., Guzman E., et al. (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. American Journal of Human Genetics 101: 700–715.

Tartier L., Michalik V., Spotheim-Maurizot M., Rahmouni A.R., Sabattier R., and Charlier M. (1994). Radiolytic signature of Z-DNA. Nucleic Acids Research 22: 5565–70.

Tawani A. and Kumar A. (2015). Structural Insights Reveal the Dynamics of the Repeating r(CAG) Transcript Found in Huntington's Disease (HD) and Spinocerebellar Ataxias (SCAs). PLOS ONE 10: e0131788.

Teisberg P. (1995). The genetic background of anticipation. Journal of the Royal Society of Medicine 88: 185–187.

Telenius H., Kremer B., Goldberg Y.P., Theilmann J., Andrew S.E., Zeisler J., et al. (1994). Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. Nature Genetics 6: 409–14.

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research 47: D330–D338.

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72: 971–983.

Thompson J.C., Harris J., Sollom A.C., Stopford C.L., Howard E., Snowden J.S., and Craufurd D. (2012). Longitudinal evaluation of neuropsychiatric symptoms in Huntington's disease. The Journal of Neuropsychiatry and Clinical Neurosciences 24: 53–60.

Thongthip S., Bellani M., Gregg S.Q., Sridhar S., Conti B.A., Chen Y., et al. (2016). Fan1 deficiency results in DNA interstrand cross-link repair defects, enhanced tissue karyomegaly, and organ dysfunction. Genes & Development 30: 645–59.

Thu D.C. V, Oorschot D.E., Tippett L.J., Nana A.L., Hogg V.M., Synek B.J., et al. (2010). Cell loss in the motor and cingulate cortex correlates with symptomatology in Huntington’s disease. Brain 133: 1094–110.

Tian Y., Wang J.-P.J.-L., Huang W., Zeng S., Jiao B., Liu Z., et al. (2019). Expansion of Human-Specific GGC Repeat in Neuronal Intranuclear Inclusion Disease-Related Disorders. American Journal of Human Genetics 105: 166–176.

Timchenko N.A., Cai Z.J., Welm A.L., Reddy S., Ashizawa T., and Timchenko L.T. (2001). RNA CUG repeats sequester CUGBP1 and alter protein levels and activity of CUGBP1. The Journal of Biological Chemistry 276: 7820–6.

Tin A., Li Y., Brody J.A., Nutile T., Chu A.Y., Huffman J.E., et al. (2018). Large-scale whole-exome sequencing association studies identify rare functional variants influencing serum urate levels. Nature Communications 9: 4228.

Tishkoff D.X., Boerger A.L., Bertrand P., Filosi N., Gaida G.M., Kane M.F., and Kolodner R.D. (1997). Identification and characterization of Saccharomyces cerevisiae EXO1, a gene encoding an exonuclease that interacts with MSH2. Proceedings of the National Academy of Sciences of the United States of America 94: 7487–7492.

Tomé S., Dandelot E., Dogan C., Bertrand A., Geneviève D., Péréon Y., et al. (2018). Unusual association of a unique CAG interruption in 5’ of DM1 CTG repeats with intergenerational contractions and low somatic mosaicism. Human Mutation 39: 970–982.

Tomé S., Manley K., Simard J.P., Clark G.W., Slean M.M., Swami M., et al. (2013). MSH3 Polymorphisms and Protein Levels Affect CAG Repeat Instability in Huntington’s Disease Mice. PLOS Genetics 9: e1003280.

Tomé S., Panigrahi G.B., López Castel A., Foiry L., Melton D.W., Gourdon G., and Pearson C.E. (2011). Maternal germline-specific effect of DNA ligase I on CTG/CAG instability. Human molecular genetics 20: 2131–43.

Tran P.T., Erdeniz N., Symington L.S., and Liskay R.M. (2004). EXO1-A multi-tasking eukaryotic nuclease. DNA Repair 3: 1549–59.

Trejo A., Tarrats R.M., Alonso M.E., Boll M.-C., Ochoa A., and Velásquez L. (2004). Assessment of the nutrition status of patients with Huntington’s disease. Nutrition 20: 192–196.

Tsuang D., Almqvist E., Lipe H., Strgar F., DiGiacomo L., Hoff D., et al. (2000). Familial aggregation of psychotic symptoms in Huntington’s disease. American Journal of Psychiatry 157: 1955–1959.

Tsuang D.W., Greenwood T.A., Jayadev S., Davis M., Shutes-David A., and Bird T.D. (2018). A Genetic Study of Psychosis in Huntington's Disease: Evidence for the Involvement of Glutamate Signaling Pathways. Journal of Huntington's disease 7: 51–59.

Twelvetrees A.E., Yuen E.Y., Arancibia-Carcamo I.L., MacAskill A.F., Rostaing P., Lumb M.J., et al. (2010). Delivery of GABAARs to Synapses Is Mediated by HAP1-KIF5 and Disrupted by Mutant Huntingtin. Neuron 65: 53–65.

Tyler A., Harper P.S., Davies K., and Newcome R.G. (1983). Family break-down and stress in Huntington's chorea. Journal of Biosocial Science 15: 127–38.

Underwood M., Bonas S., and Dale M. (2017). Huntington's Disease: Prevalence and Psychological Indicators of Pain. Movement Disorders Clinical Practice 4: 198–204.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 47: D506–D515.

Untergasser A., Cutcutache I., Koressaar T., Ye J., Faircloth B.C., Remm M., and Rozen S.G. (2012). Primer3--new capabilities and interfaces. Nucleic Acids Research 40: e115.

Urbanek M.O., Jazurek M., Switonski P.M., Figura G., and Krzyzosiak W.J. (2016). Nuclear speckles are detention centers for transcripts containing expanded CAG repeats. Biochimica et Biophysica Acta 1862: 1513–20.

Vamos M., Hambridge J., Edwards M., and Conaghan J. (2007). The impact of Huntington's disease on family life. Psychosomatics 48: 400–4.

Vaser R., Adusumalli S., Leng S.N., Sikic M., and Ng P.C. (2016). SIFT missense predictions for genomes. Nature Protocols 11: 1–9.

Vassos E., Panas M., Kladi A., and Vassilopoulos D. (2008). Effect of CAG repeat length on psychiatric disorders in Huntington's disease. Journal of Psychiatric Research 42: 544–9.

Verkerk A.J.M.H., Pieretti M., Sutcliffe J.S., Fu Y.-H., Kuhl D.P.A., Pizzuti A., et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65: 905–914.

Videnovic A., Leurgans S., Fan W., Jaglin J., and Shannon K.M. (2009). Daytime somnolence and nocturnal sleep disturbances in Huntington disease. Parkinsonism & Related Disorders 15: 471–474.

Visscher P.M., Wray N.R., Zhang Q., Sklar P., McCarthy M.I., Brown M.A., and Yang J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics 101: 5–22.

Volle C.B., Jarem D.A., and Delaney S. (2012). Trinucleotide repeat DNA alters structure to minimize the thermodynamic impact of 8-oxo-7,8-dihydroguanine. Biochemistry 51: 52–62.

Vonsattel J.P. and DiFiglia M. (1998). Huntington disease. Journal of Neuropathology and Experimental Neurology 57: 369–84.

Vonsattel J.P., Myers R.H., Stevens T.J., Ferrante R.J., Bird E.D., and Richardson E.P. (1985). Neuropathological classification of Huntington's disease. Journal of Neuropathology and Experimental Neurology 44: 559–77.

Waldvogel H.J., Kim E.H., Tippett L.J., Vonsattel J.-P.G., and Faull R.L.M. (2015). The Neuropathology of Huntington's Disease. Current Topics in Behavioral Neurosciences 22: 33–80.

Wang C.-E., Tydlacka S., Orr A.L., Yang S.-H., Graham R.K., Hayden M.R., et al. (2008). Accumulation of N-terminal mutant huntingtin in mouse and monkey models implicated as a pathogenic mechanism in Huntington's disease. Human Molecular Genetics 17: 2738–51.

Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F.A., et al. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research 17: 1665–1674.

Wang Q., Lu Q., and Zhao H. (2015). A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Frontiers in Genetics 6: 149.

Wang Q., Shashikant C.S., Jensen M., Altman N.S., and Girirajan S. (2017). Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. Scientific Reports 7: 885.

Wang R., Persky N.S., Yoo B., Ouerfelli O., Smogorzewska A., Elledge S.J., and Pavletich N.P. (2014). Mechanism of DNA interstrand cross-link processing by repair nuclease FAN1. Science 346: 1127–30.

Wang Z., Song Y., Li S., Kurian S., Xiang R., Chiba T., and Wu X. (2019). DNA polymerase θ (POLQ) is important for repair of DNA double-strand breaks caused by fork collapse. Journal of Biological Chemistry 294: 3909–3919.

Warby S.C., Visscher H., Collins J.A., Doty C.N., Carter C., Butland S.L., et al. (2011). HTT haplotypes contribute to differences in Huntington disease prevalence between Europe and East Asia. European Journal of Human Genetics 19: 561–566.

Warner J.P., Barron L.H., and Brock D.J. (1993). A new polymerase chain reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded on Huntington's disease chromosomes. Molecular and Cellular Probes 7: 235–9.

Weeke P., Mosley J.D., Hanna D., Delaney J.T., Shaffer C., Wells Q.S., et al. (2014). Exome sequencing implicates an increased burden of rare potassium channel variants in the risk of drug-induced long QT interval syndrome. Journal of the American College of Cardiology 63: 1430–7.

Wellington C.L., Ellerby L.M., Hackam A.S., Margolis R.L., Trifiro M.A., Singaraja R., et al. (1998). Caspase Cleavage of Gene Products Associated with Triplet Expansion Disorders Generates Truncated Fragments Containing the Polyglutamine Tract. Journal of Biological Chemistry 273: 9158–9167.

Wexler N.S. (2013). Three decades of caring for the Venezuelan Huntington's disease families. The Lancet Neurology 12: 738.

Wexler N.S., Lorimer J., Porter J., Gomez F., Moskowitz C., Shackell E., et al. (2004). Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proceedings of the National Academy of Sciences of the United States of America 101: 3498–503.

Weydt P., Soyal S.M., Gellera C., Didonato S., Weidinger C., Oberkofler H., et al. (2009). The gene coding for PGC-1alpha modifies age at onset in Huntington's Disease. Molecular Neurodegeneration 4: 3.

Wheeler V.C., Lebel L.A., Vrbanac V., Teed A., te Riele H.T., and MacDonald M.E. (2003). Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum. Human Molecular Genetics 12: 273–281.

Wheeler V.C., Persichetti F., McNeil S.M., Mysore J.S., Mysore S.S., MacDonald M.E., et al. (2007). Factors associated with HD CAG repeat instability in Huntington disease. Journal of Medical Genetics 44: 695–701.

Wieben E.D., Aleff R.A., Tosakulwong N., Butz M.L., Highsmith W.E., Edwards A.O., and Baratz K.H. (2012). A Common Trinucleotide Repeat Expansion within the Transcription Factor 4 (TCF4, E2-2) Gene Predicts Fuchs Corneal Dystrophy. PLOS ONE 7: e49083.

Willems T., Zielinski D., Yuan J., Gordon A., Gymrek M., and Erlich Y. (2017). Genome-wide profiling of heritable and de novo STR variations. Nature Methods 14: 590–592.

Williams S.A., Wilson J.B., Clark A.P., Mitson-Salazar A., Tomashevski A., Ananth S., et al. (2011). Functional and physical interaction between the mismatch repair and FA-BRCA pathways. Human Molecular Genetics 20: 4395–410.

Wilson H., De Micco R., Niccolini F., and Politis M. (2017). Molecular Imaging Markers to Track Huntington's Disease Pathology. Frontiers in Neurology 8: 11.

Winder J.Y. and Roos R.A.C. (2018). Premanifest Huntington’s disease: Examination of oculomotor abnormalities in clinical practice. PLOS ONE 13: e0193866.

Winnepenninckx B., Debacker K., Ramsay J., Smeets D., Smits A., FitzPatrick D.R., and Kooy R.F. (2007). CGG-Repeat Expansion in the DIP2B Gene Is Associated with the Fragile Site FRA12A on Chromosome 12q13.1. American Journal of Human Genetics 80: 221–231.

Woerner A.C., Frottin F., Hornburg D., Feng L.R., Meissner F., Patra M., et al. (2016). Cytoplasmic protein aggregates interfere with nucleocytoplasmic transport of protein and RNA. Science 351: 173–6.

Wong Y.C. and Holzbaur E.L.F. (2014). The Regulation of Autophagosome Dynamics by Huntingtin and HAP1 Is Disrupted by Expression of Mutant Huntingtin, Leading to Defective Cargo Degradation. Journal of Neuroscience 34: 1293–1305.

Wood T.E., Barry J., Yang Z., Cepeda C., Levine M.S., and Gray M. (2018). Mutant huntingtin reduction in astrocytes slows disease progression in the bachd conditional huntington’s disease mouse model. Human Molecular Genetics 28: 487–500.

Wright G.E.B., Collins J.A., Kay C., McDonald C., Dolzhenko E., Xia Q., et al. (2019). Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. American Journal of Human Genetics 104: 1116–1126.

Wu B. and Pankow J.S. (2016). On Sample Size and Power Calculation for Variant Set-Based Association Tests. Annals of Human Genetics 80: 136–143.

Wu D., Faria A. V, Younes L., Mori S., Brown T., Johnson H., et al. (2017). Mapping the order and pattern of brain structural MRI changes using change-point analysis in premanifest Huntington’s disease. Human Brain Mapping 38: 5035–5050.

Wu M.C., Lee S., Cai T., Li Y., Boehnke M., and Lin X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics 89: 82–93.

Xu B., Roos J.L., Dexheimer P., Boone B., Plummer B., Levy S., et al. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nature Genetics 43: 864–8.

Xu J., Mashimo T., and Südhof T.C. (2007). Synaptotagmin-1, -2, and -9: Ca(2+) sensors for fast release that specify distinct presynaptic properties in subsets of neurons. Neuron 54: 567–81.

Yamada M., O’Regan E., Brown R., and Karran P. (1997). Selective recognition of a cisplatin-DNA adduct by human mismatch repair proteins. Nucleic Acids Research 25: 491–6.

Yan P.-X., Huo Y.-G., and Jiang T. (2015). Crystal structure of human Fanconi-associated nuclease 1. Protein & Cell 6: 225–8.

Yang Y., Muzny D.M., Reid J.G., Bainbridge M.N., Willis A., Ward P.A., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. The New England Journal of Medicine 369: 1502–11.

Yildirim I., Park H., Disney M.D., and Schatz G.C. (2013). A Dynamic Structural Model of Expanded RNA CAG Repeats: A Refined X-ray Structure and Computational Investigations Using Molecular Dynamics and Umbrella Sampling Simulations. Journal of the American Chemical Society 135: 3528–3538.

Yoon G., Kramer J., Zanko A., Guzijan M., Lin S., Foster-Barber A., and Boxer A.L. (2006). Speech and language delay are early manifestations of juvenile-onset Huntington disease. Neurology 67: 1265–1267.

Yoon S.-R., Dubeau L., de Young M., Wexler N.S., and Arnheim N. (2003). Huntington disease expansion mutations in humans can occur before meiosis is completed. Proceedings of the National Academy of Sciences of the United States of America 100: 8834–8.

Yoshikiyo K., Kratz K., Hirota K., Nishihara K., Takata M., Kurumizaka H., et al. (2010). KIAA1018/FAN1 nuclease protects cells against genomic instability induced by interstrand cross-linking agents. Proceedings of the National Academy of Sciences of the United States of America 107: 21553–7.

Yrigollen C.M., Durbin-Johnson B., Gane L., Nelson D.L., Hagerman R., Hagerman P.J., and Tassone F. (2012). AGG interruptions within the maternal FMR1 gene reduce the risk of offspring with fragile X syndrome. Genetics in Medicine 29: 997–1003.

Yu S., Fimmel A., Fung D., and Trent R.J. (2000). Polymorphisms in the CAG repeat--a source of error in Huntington disease DNA testing. Clinical Genetics 58: 469–72.

Yu Z., Zhu Y., Chen-Plotkin A.S., Clay-Falcone D., McCluskey L., Elman L., et al. (2011). PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. PLOS ONE 6: e17951.

Zaharia M., Franklin M.J., Ghodsi A., Gonzalez J., Shenker S., Stoica I., et al. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM 59: 56–65.

Zeitlin S., Liu J.-P., Chapman D.L., Papaioannou V.E., and Efstratiadis A. (1995). Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. Nature Genetics 11: 155–163.

Zhang T., Huang J., Gu L., and Li G.-M. (2012). In vitro repair of DNA hairpins containing various numbers of CAG/CTG trinucleotide repeats. DNA Repair 11: 201–209.

Zhao Q., Xue X., Longerich S., Sung P., and Xiong Y. (2014). Structural insights into 5' flap DNA unwinding and incision by the human FAN1 dimer. Nature Communications 5: 5726.

Zhao X.-N., Lokanga R., Allette K., Gazy I., Wu D., and Usdin K. (2016). A MutSβ-Dependent Contribution of MutSα to Repeat Expansions in Fragile X Premutation Mice? PLOS Genetics 12: e1006190.

Zhao X.-N. and Usdin K. (2018). FAN1 protects against repeat expansions in a Fragile X mouse model. DNA Repair 69: 1–5.

Zhao X., Zhang Y., Wilkins K., Edelmann W., and Usdin K. (2018). MutLγ promotes repeat expansion in a Fragile X mouse model while EXO1 is protective. PLOS Genetics 14: e1007719.

Zhou L. and Zhao F. (2018). Prioritization and functional assessment of noncoding variants associated with complex diseases. Genome Medicine 10: 53.

Zhou W., Otto E.A., Cluckey A., Airik R., Hurd T.W., Chaki M., et al. (2012). FAN1 mutations cause karyomegalic interstitial nephritis, linking chronic kidney failure to defective DNA damage repair. Nature Genetics 44: 910–5.

Zigmond A.S. and Snaith R.P. (1983). The hospital anxiety and depression scale. Acta Psychiatrica Scandinavica 67: 361–70.

Zu T., Cleary J.D., Liu Y., Bañez-Coronel M., Bubenik J.L., Ayhan F., et al. (2017). RAN Translation Regulated by Muscleblind Proteins in Myotonic Dystrophy Type 2. Neuron 95: 1292-1305.e5.

Zu T., Gibbens B., Doty N.S., Gomes-Pereira M., Huguet A., Stone M.D., et al. (2011). Non-ATG-initiated translation directed by microsatellite expansions. Proceedings of the National Academy of Sciences of the United States of America 108: 260–5.

Zuccato C., Ciammola A., Rigamonti D., Leavitt B.R., Goffredo D., Conti L., et al. (2001). Loss of huntingtin-mediated BDNF gene transcription in Huntington's disease. Science 293: 493–8.

Zuccato C., Tartari M., Crotti A., Goffredo D., Valenza M., Conti L., et al. (2003). Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes.

Nature Genetics 35: 76–83.

Zuccato C., Valenza M., and Cattaneo E. (2010). Molecular Mechanisms and Potential Therapeutical Targets in Huntington's Disease. Physiological Reviews 90: 905–981.

Zühlke C., Hellenbroich Y., Dalski A., Kononowa N., Hagenah J., Vieregge P., et al. (2001). Different types of repeat expansion in the TATA-binding protein gene are associated with a new form of inherited ataxia. European Journal of Medical Genetics 9: 160–4.

Zühlke C., Riess O., Bockel B., Lange H., and Thies U. (1993). Mitotic stability and meiotic variability of the (CAG)n repeat in the Huntington disease gene. Human Molecular Genetics 2: 2063–7.

Zuk O., Schaffner S.F., Samocha K., Do R., Hechter E., Kathiresan S., et al. (2014). Searching for missing heritability: designing rare variant association studies. Proceedings of the National Academy of Sciences of the United States of America 111: E455-64.