Identification of genetic factors underpinning phenotypic heterogeneity in Huntington’s disease and other neurodegenerative disorders.

By Dr Davina J Hensman Moss

A thesis submitted to University College London for the degree of
Doctor of Philosophy

Department of Neurodegenerative Disease
Institute of Neurology
University College London (UCL)
2020
I, Davina Hensman Moss confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. Collaborative work is also indicated in this thesis.

Signature:

Date:
Abstract

Neurodegenerative diseases including Huntington’s disease (HD), the spinocerebellar ataxias and C9orf72 associated Amyotrophic Lateral Sclerosis / Frontotemporal dementia (ALS/FTD) do not present and progress in the same way in all patients. Instead there is phenotypic variability in age at onset, progression and symptoms. Understanding this variability is not only clinically valuable, but identification of the genetic factors underpinning this variability has the potential to highlight genes and pathways which may be amenable to therapeutic manipulation, hence help find drugs for these devastating and currently incurable diseases. Identification of genetic modifiers of neurodegenerative diseases is the overarching aim of this thesis.

To identify genetic variants which modify disease progression it is first necessary to have a detailed characterization of the disease and its trajectory over time. In this thesis clinical data from the TRACK-HD studies, for which I collected data as a clinical fellow, was used to study disease progression over time in HD, and give subjects a progression score for subsequent analysis. In this thesis I show blood transcriptomic signatures of HD status and stage which parallel HD brain and overlap with Alzheimer’s disease brain.

Using the Huntington’s disease progression score in a genome wide association study, both a locus on chromosome 5 tagging MSH3, and DNA handling pathways more broadly, are shown to modify HD progression: these results are explored. Transcriptomic signatures associated with HD progression rate are also investigated.

In this thesis I show that DNA repair variants also modify age at onset in spinocerebellar ataxias (1, 2, 3, 6, 7 and 17), which are, like HD, caused by triplet repeat expansions, suggesting a common mechanism. Extending this thesis’ examination of the relationship between phenotype and genotype I show that the C9orf72 expansion, normally associated with ALS/FTD, is also the commonest cause of HD phenocopy presentations.
**Impact Statement**

The work presented in this thesis has been disseminated and had an impact both within and beyond academia. Foremost among the positive impacts of this work has been the identification of MSH3 as a modifier of disease progression in Huntington’s disease, and, by extension of work presented in this thesis is likely to be a modifier of other polyglutamine repeat disorders. As a result of my work, ongoing study has been taking place about the likely mechanism of this modifier effect. It is likely that small molecule inhibitors of MSH3 may have the potential to slow progression of HD in patients, and as a consequence several pharmaceutical companies are working on this potential therapeutic avenue.

The following key papers have been published in scholarly journals based on material covered in this thesis, along with various other papers in which I had a smaller role or that were beyond the scope of this thesis (listed in Appendix 2). The relevant thesis chapter is given, and the number of papers citing each paper (on 25/03/2019) is given in parenthesis.


I have presented work from this thesis at various international meetings and conferences, and it has attracted the following prizes:
1. The Alzheimers Research UK Jean Corsan Prize which is awarded each year for the best scientific paper in neurodegeneration published by a PhD or MD/PhD student, London, March 2018.


I have also taken part in several activities to disseminate my work to a broader audience, including being interviewed for profile pieces in Lancet Neurology (https://doi.org/10.1016/S1474-4422(18)30329-6) and ARUK’s Demenia blog https://www.dementiablog.org/scientist-focus-davina-hensman-moss/) along with news features on the St George’s Hospital and UCL websites and talks to patient groups.
Acknowledgements

Firstly, I would like to thank the participants and those who support them who generously gave their time to be part of the research projects upon which this PhD Thesis is based, in particular those people who took part in the TRACK-HD and TrackOn-HD studies which were so critical to my work.

I would like to thank my supervisor Prof. Sarah Tabrizi for her unending support and guidance, in the clinic, in the lab and beyond. It has been a pleasure to work for such an inspirational role model. I would like to thank Prof. Lesley Jones, my secondary supervisor at Cardiff University for her wisdom, council and sound guidance, particularly for the genetics aspects of my work.

I would give thanks to the following: Prof. Simon Mead for the training in genetics, supervising the C9orf72 aspects of this thesis, and for ongoing advice; Prof. Douglas Langbehn for the extensive statistical input to my projects; Prof. Peter Holmans for his clear vision, and for statistical genetics input to my projects; Dr Antonio Pardiñas interesting discussions, and a huge amount of help with data analysis. The discussions between Sarah Tabrizi, Lesley Jones, Douglas Langbehn, Peter Holmans and myself about the nature of progression in Huntington’s disease and how best to examine it were an inspiration, and formed a valuable bedrock for the subsequent work in this thesis.

My PhD has been highly collaborative and it has been a pleasure and a privilege to work with many brilliant collaborators during the course of these projects. I would like in particular to thank the following, in no particular order, for their invaluable input in various aspects of my work: Dr Kitty Lo, Dr Vincent Plagnol, Prof. Henry Houlden, Dr Conceicao Bettencourt, Dr Sarah Weithoff, Dr Michael Flower, Dr Willeke van Roon-Mom, Prof. Alexandra Durr, Dr Edward Wild, Dr Rob Goold, Dr Ralph Andre, Dr Alison Wood-Kaczmar, Dr Timothy Stone, Prof. Darren Monckton, Mark Poulter, Jon Beck, Gary Adamson, Tracy Campbell, Ruth Farmer, Dr Rachael Sahill, Dr Marina Papoutsi, Dr Peter McColgan, Nicci Robertson and all the members of the HD Research Centre.

I would also like to thank CHDI for its funding of TRACK-HD, TrackOn-HD and particularly for funding me as Clinical Fellow for TrackOn-HD; the European Commission for funding Neuromics and through that much of the genetic analysis; and the Guarantors of Brain who funded me through an exit fellowship and travel bursaries.

Finally I would like to thank my parents Barbara and Nigel Hensman who encouraged my curiosity and always had faith in me; my husband George Moss for his unfailing love, support and understanding, and my children Arthur, Freddie and Louisa who arrived during the course of this PhD and have made my world a richer place.
# Table of Contents

Abstract ........................................................................................................................................... 3  
Impact Statement ............................................................................................................................. 4  
Acknowledgements.......................................................................................................................... 6  
Table of Contents ............................................................................................................................ 7  
List of Figures .................................................................................................................................. 13  
List of Tables .................................................................................................................................... 16  
Chapter 1: Introduction .................................................................................................................... 20  
  1.1 Genes and disease ...................................................................................................................... 20  
  1.2 Huntington’s disease .................................................................................................................. 21  
    1.2.1 Clinical characteristics and prevalence .............................................................................. 21  
    1.2.2 Motor features .................................................................................................................... 22  
    1.2.3 Psychiatric features ............................................................................................................ 22  
    1.2.4 Cognitive features ............................................................................................................ 22  
    1.2.5 Disease onset ...................................................................................................................... 23  
    1.2.6 HD Genetics ...................................................................................................................... 24  
    1.2.7 Role of CAG repeat length in the phenotype of Huntington’s disease ......................... 24  
    1.2.8 Disease burden score and cumulative probability of disease onset ......................... 24  
    1.2.9 Intergenerational and somatic instability of the HTT CAG repeat .......................... 25  
  1.3 C9orf72 associated Amyotrophic Lateral Sclerosis / Frontotemporal dementia ............. 26  
    1.3.1 Clinical Features ............................................................................................................... 26  
    1.3.2 Genetics ............................................................................................................................ 26  
  1.4 The Spinocerebellar ataxias ..................................................................................................... 28  
  1.5 Genetic analysis ....................................................................................................................... 28  
  1.6 Previous work on Genetic Modifiers of Huntington’s disease .......................................... 33  
  1.7 DNA repair and Somatic Instability ...................................................................................... 35  
Aims of this Thesis ......................................................................................................................... 36  
Chapter 2: General Methods .......................................................................................................... 37  
  2.1 Consent and ethics .................................................................................................................... 37  
  2.2 Standard assessments commonly used to examine Huntington’s disease which are employed in this thesis .................................................................................................................. 37
3.2.3 Case ascertainment ........................................................................................................... 64
3.2.4 Relationship between progression scores used in TRACK-HD and REGISTRY .......... 64
3.2.5 Relationship between progression scores and other clinical measures ....................... 65
3.2.6 Genotyping ......................................................................................................................... 66
3.2.7 Relatedness and Population genetic analysis ................................................................. 66
3.2.8 Imputation ........................................................................................................................ 67
3.2.9 Mixed linear model GWAS ............................................................................................. 69
3.2.10 Co-localisation analyses ................................................................................................. 73
3.2.11 Gene based analyses ....................................................................................................... 74
3.2.12 Gene-set analyses ............................................................................................................ 74
3.2.13 Linking genetic variation to clinical measures ............................................................... 75
3.3 Results .................................................................................................................................. 76
3.3.1 Phenotypic clusters of Huntington’s disease were not observed ......................... 76
3.3.2 The progression scores are correlated with change in more widely used clinical measures of Huntington’s disease ................................................................. 80
3.3.3 Cross-sectional severity score used as the progression measure in REGISTRY ...... 81
3.3.4 The TRACK-HD and REGISTRY progression measures are correlated ...................... 83
3.3.5 Progression scores are associated with AAO ................................................................. 84
3.3.6 Genome wide association analysis highlights a locus associated with HD progression on chromosome 5 in TRACK-HD ................................................................. 86
3.3.7 The chromosome 5 signal is replicated in a genome wide association study in REGISTRY, and strengthened in meta-analysis ................................................................. 93
3.3.8 Variants associated with slower HD progression are associated with decreased MSH3 expression .................................................................................................................. 96
3.3.9 REGISTRY association analysis highlights locus on chromosome 15 .................... 99
3.3.10 The observed associations with progression are not all driven by age at onset ... 99
3.3.11 Effect of index MSH3 SNP on clinical measures ......................................................... 101
3.3.12 Pathway analysis shows association between HD progression and genes involved in DNA repair ............................................................................................................... 101
3.4 Discussion ............................................................................................................................ 109
Chapter 4: DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases

4.1 Introduction

4.2 Materials and Methods

4.2.1 Cohort

4.2.2 Selection of SNPs

4.2.3 Genotyping

4.2.4 Statistical analysis

4.3 Results

4.3.1 There is a combined effect of 22 DNA repair gene SNPs on Age at Onset

4.3.2 Individual SNPs were also significantly associated with onset

4.3.3 Looking at the combined effect of the SNPs in a polygenic score

4.4 Discussion

Chapter 5: Use of sequencing to look for rare variants of larger effect and identify sequence variants in loci highlighted by genetic analysis

5.1 Introduction

5.2 Materials and Methods

5.2.1 Whole Exome Sequencing

5.2.2 Pathway analysis of WES data

5.2.3 eQTL analysis of MSH3 variant

5.2.4 Sanger sequencing of MSH3 region of interest

5.2.6 Interrogation of RD-Connect database

5.2.7 MSH3 structural prediction

5.2.8 Phylogenetic analysis

5.3 Results

5.3.1 Whole Exome Sequencing

5.3.2 Several DNA repair pathways nominally associated with HD progression in the WES fast vs slow analysis

5.3.3 Sequence variants in FAN1 were identified from the exome sequence data

5.3.4 Two MSH3 variants were highlighted by the WES fast vs slow analysis
5.3.5 MSH3 coding variant rs557874766, the index SNP from TRACK-HD GWAS was not found in exome sequence data

5.3.6 SNP in high linkage disequilibrium with rs557874766 was identified

5.3.7 Sanger sequencing of TRACK-HD subjects provided further evidence for the presence of deletions in people expected to have rs557874766

5.3.8 rs557874766 was not found in sequence data of 1280 individuals

5.3.9 Structural predictions show that slow progressors have lost an alpha-helical region in the N-terminus of MSH3

5.3.10 Phylogenetic data suggest that the polyalanine can be viewed as a recent insertion

5.4 Discussion

Chapter 6 - C9orf72 repeat expansion disease: examination of intergenerational repeat stability and expansion of the known phenotype to encompass HD phenocopy presentations

6.1 Introduction

6.2 Materials and Methods

6.2.1 Standard Protocol Approvals, Registrations, and Patient Consents

6.2.2 Case ascertainment: Control samples for intergenerational stability analysis

6.2.3 Case ascertainment: HD phenocopy subjects

6.2.4 Clinical phenotyping

6.2.5 Repeat primed PCR

6.2.6 rs3849942 genotyping

6.2.7 Microsatellite genotyping

6.2.8 Southern hybridisation

6.3 Results

6.3.1 C9orf72 repeat intergenerational instability is seen in those with longer repeat lengths

6.3.2 Identification of C9orf72 expansion in HD phenocopy cases

6.3.3 Presence of risk haplotype in those with expansion mutations and with intergenerational repeat instability

6.3.4 Clinical data
6.3.5 Comparisons between C9orf72 positive cases and the rest of the HD phenocopy cohort ................................................................. 179
6.3.6 An illustrative case ..................................................................................................................................................... 180
6.3.7 An unusual case ..................................................................................................................................................... 181
6.3.8 A homozygous case ................................................................................................................................................... 182
6.4 Discussion .................................................................................................................................................................. 183

Chapter 7: Investigations of the effect of disease status, stage and rate of progression on the transcriptome in Huntington’s disease ........................................................................................................ 188
7.1 Introduction ............................................................................................................................................................... 188
7.2 Materials and methods .............................................................................................................................................. 190
7.2.1 Cohorts ......................................................................................................................................................... 190
7.2.2 Sample collection .............................................................................................................................................. 193
7.2.3 RNA preparation ............................................................................................................................................... 193
7.2.4 RNA Sequencing .............................................................................................................................................. 193
7.2.5 Quality control .................................................................................................................................................. 194
7.2.6 Gene expression analysis ................................................................................................................................... 194
7.2.7 Pathway analysis ................................................................................................................................................ 195
7.2.8 Gene co-expression networks ............................................................................................................................. 196
7.2.9 Concordance of fold change in gene expression between HD blood and cortex ................................................................................................................................. 197
7.3 Results: Effect of HD gene status and stage of disease on the transcriptome ......................................................... 198
7.3.1 No differential expression of individual transcripts in HD whole blood between disease stages or states ........................................................................................................................................... 198
7.3.2 Pathways are dysregulated in HD blood compared with controls ........................................................................ 200
7.3.3 Pathway dysregulation in HD whole blood overlaps with HD myeloid cells ...................................................... 211
7.3.4 Gene co-expression modules from HD striatum are significantly enriched for dysregulation in HD blood ........................................................................................................................................ 214
7.3.5 Expression changes in HD blood replicate those in HD prefrontal cortex ...................................................... 224
7.3.6 Pathways dysregulated in the blood of HD subjects are associated with motor score ........................................ 227
7.3.7 The Alzheimer’s disease brain transcriptional signature is significantly dysregulated in HD blood......................................................................................................................... 236

7.4 Results: Relationship between rate of HD progression and the transcriptome .......... 237

7.4.1 No differential expression of individual transcripts in HD whole blood with changing rate of disease progression............................................................................................................ 237

7.4.2 Pathways are dysregulated in HD subjects with faster vs slower rates of disease progression................................................................................................................................. 240

7.4.3 Gene co-expression modules and rate of HD Progression .................................... 246

7.4.4 Comparison of HD progression results to the HD vs control WGCNA results . 249

7.4.5 Attempted replication of TRACK-HD progression RNAseq results in the LUMC dataset........................................................................................................................................... 249

7.4.6 No individual transcripts are differentially expressed according to rate of HD progression in the LUMC cohort ........................................................................................................... 249

7.4.7 Pathway analysis of LUMC progression data ......................................................... 251

7.5 Discussion .................................................................................................................. 252

Chapter 8: Conclusion and future directions .................................................................. 258

References ....................................................................................................................... 266

Appendix 1: ...................................................................................................................... 302
General PCR and Sequencing protocol.............................................................................. 302

Appendix 2: ...................................................................................................................... 304
Published papers and book chapters .............................................................................. 304

List of Figures

Figure 1.1: Types of DNA within the human genome. ...................................................... 21

Figure 1.2: Longitudinal changes in cognitive measures from the Track-HD study over 24 months ............................................................................................................................................... 23

Figure 1.3: Cumulative probability of Huntington’s disease onset curves for various CAG lengths .................................................................................................................................................. 25

Figure 1.4: Clinical, genetic and pathological overlap of ALS and FTD. ......................... 27

Figure 1.5: Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). .............................................................................................................. 30
Figure 1.6: Ever-increasing sample sizes for genome-wide association studies (GWAS)........ 31
Figure 2.1: Study outline of TRACK-HD. Study sites, numbers of subjects in each disease
group at baseline, principle assessment modalities and years assessed are shown. ............. 40
Figure 2.2: Age-CAG severity function against clinical probability of onset (CPO) in REGISTRY.
................................................................................................................................. 50
Figure 2.3: Family history encoding........................................................... 52
Figure 3.1: Study Design....................................................................................... 63
Figure 3.2: Ancestry analysis of the TRACK-HD cohort........................................... 67
Figure 3.3: Genotype imputation in a sample of apparently unrelated individuals............. 68
Figure 3.4: QQ plots.............................................................................................. 73
Figure 3.5: Distribution of progression measure in 218 members of TRACK-HD cohort ....... 78
Figure 3.6: The first principal component accounts for a high proportion of the variance in the
TRACK-HD progression analysis. ................................................................................. 79
Figure 3.7: The first principal component accounts for a high proportion of the variance in the
REGISTRY progression analysis. ................................................................................. 81
Figure 3.8: Distribution of atypical severity (compared to predicted severity at final visit) in
1835 members of the REGISTRY cohort ................................................................. 82
Figure 3.9: Assessing progression in Huntington’s disease ........................................... 83
Figure 3.10: TRACK-HD and REGISTRY progression scores are correlated.................... 84
Figure 3.11: Observed versus Expected Age of Onset.................................................. 85
Figure 3.12: REGISTRY progression measure (Residual severity score) and atypical onset age
(Standardised onset) are modestly correlated in REGISTRY........................................... 86
Figure 3.13: Genome-wide Association Analysis of Progression Score......................... 88
Figure 3.14: Locus zoom plot of the TRACK-HD (top), REGISTRY (middle) and meta-analysis
(bottom) data............................................................................................................... 89
Figure 3.15: Regional plot of TRACK-HD and REGISTRY meta-analysis GWAS signal in the
MSH3-DHFR region before (top) and after (bottom) conditioning on the most significant SNP
in TRACK-HD (rs557874766). ......................................................................................... 93
Figure 3.16: Regional plot of REGISTRY GWAS signal in the MSH3-DHFR region before (top)
and after (bottom) conditioning on the most significant SNP in TRACK-HD ................... 94
Figure 3.17: Conditional analysis.................................................................................. 96
Figure 3.18: Expression analysis.................................................................................... 97
Figure 3.19: Significant genes are functionally linked and may cause somatic expansion of the
HTT CAG repeat tract ................................................................................................. 106
Figure 3.20: Schematic of DNA damage recognized by the MMR pathway................... 113
Figure 3.21: A Schematic diagram showing how DNA mismatch repair proteins may be involved in somatic expansion of the CAG tract................................................................................................. 116

Figure 4.1: Boxplot of residual AAO (across all samples) by quartiles of polygenic age at onset score...................................................................................................................................................... 138

Figure 4.2: String diagram illustrating the functional connection between the proteins included in this study.................................................................................................................................................. 139

Figure 4.3: Potential mechanism by which variants in DNA repair could influence somatic expansion of CAG repeats........................................................................................................................................ 141

Figure 5.1: Distribution of the Progression scores in the TRACK-HD cohort..................................... 141

Figure 5.2: Influence of rs184967 allele status on brain expression of MSH3...................................... 156

Figure 5.3: Secondary structure predictions for MSH3........................................................................ 159

Figure 5.4: Predicted solvent exposure for MSH3................................................................................ 160

Figure 5.5: Tertiary structure predictions of MSH3.............................................................................. 161

Figure 5.6: Cladogram of the apes showing the MSH3 protein sequence at the repetitive region of interest in six different ape species......................................................................................... 162

Figure 5.7: Excerpt of the MSH3 exon 1 sequence................................................................................ 165

Figure 5.8: Alternative MSH3 deletions achieve the same protein sequence result......................... 165

Figure 5.9: Figure of MSH3 showing the putative binding domains for proteins with which it interacts, and the ATP binding site........................................................................................................ 167

Figure 6.1: Fragment analysis of CEPH families with inter-generational repeat slippage...... 176

Figure 6.2: Southern Blot of eight HD phenocopy patient DNAs, blot produced by Mark Poulter...................................................................................................................................................... 177

Figure 6.3: Algorithm for the investigation of HD phenocopy cases................................................. 185

Figure 7.1: Upregulated pathways in HD versus control blood................................................................. 206

Figure 7.2: Downregulated pathways in HD versus control blood......................................................... 208

Figure 7.3: Network diagram of the relationship between significantly (q<0.05) upregulated gene modules (Table 7.10) and generic biological pathways (Table 7.5) based on shared gene membership............................................................................................................................. 222

Figure 7.4: Network diagram of the relationship between significantly (q<0.05) downregulated gene modules (Table 7.10) and generic biological pathways (Table 6) based on shared gene membership............................................................................................................................. 223

Figure 7.5: Cell cycle pathways expression is associated with rate of HD progression ............ 244

Figure 7.6: Pathways related to progression in the LUMC cohort ....................................................... 252

Figure 8.1: The main DNA damage response (DDR) pathways with the proteins suspected to be involved in each. ................................................................................................................................. 260
Figure 8.2: Innate immune pathways in neurodegenerative diseases. A maladaptive innate immune response has emerged as a critical driving force in the pathogenesis of many neurodegenerative diseases. .......................................................................................................................... 264

List of Tables

Table 1.1: Relationship between size of CAG repeat expansion and clinical outcome.............. 24
Table 2.1: Total Functional Capacity Scale. HD: Huntington’s disease................................. 37
Table 2.2: Outline of TrackOn-HD assessment day. ................................................................. 41
Table 2.3: List of Variables to be used in TRACK-HD progression analyses.......................... 44
Table 2.4: Parameter estimates of variables in the model used to generate the REGISTRY cross sectional severity score........................................................................................................ 48
Table 2.5: Format for family history encoding ..................................................................... 52
Table 3.1: Proportion of variance among variables present in TRACK-HD and REGISTRY which are accounted for by the first PC in the combined analysis......................................................... 65
Table 3.2: Relationship between change in progression score and rate of change in Total Motor Score (TMS) and Total Functional Capacity (TFC)................................................................. 75
Table 3.3: Correlations among Domain-Specific Residual Principal Components in the TRACK-HD analysis.............................................................................................................................. 77
Table 3.4: PCA of Residual Longitudinal Change Among Variables from All 3 Domains in the TRACK-HD analysis showing that the variables that correlated with the domain specific analyses also correlated with the common principal component analysis. Dom- dominant; nondom- nondominant; std dev- standard deviation. ...................................................................................... 80
Table 3.5: Factor pattern of the first two principal component analysis of the REGISTRY severity score which was used as a progression score for the Registry data........................... 82
Table 3.6: Independent association signals from the TRACK-HD Progression GWAS (at p-value < 10^{-5}). ................................................................................................................................. 90
Table 3.7: Gene-wide p-values for top genes in TRACK-HD, REGISTRY, the TRACK-REGISTRY meta analysis (p(META)), and GeM from the MAGMA analysis........................................ 92
Table 3.8: Independent association signals from the meta-analysis of TRACK-HD and REGISTRY Progression GWAS (at p-value < 10^{-5}). ............................................................................. 95
Table 3.9: Significant (p<0.001) SNPs from TRACK-HD GWAS chromosome 5 region showing direction of effect (beta) on progression (GWAS) and expression (eQTL). ............................. 98
Table 3.10: Gene-wide p-values for all genes in TRACK-HD, REGISTRY and the TRACK-REGISTRY meta-analysis after conditioning on AAO [p(TRACKcond); p(REGcond), p(METAcond) respectively], compared to their values without conditioning................................................................. 100

Table 3.11: Setscreen enrichment p-values for the 14 pathways highlighted in GeM-HD (8). 102

Table 3.12: Setscreen enrichment p-values for the Pearl et al. (2015) pathways in TRACK-HD, REGISTRY, the TRACK-HD meta-analysis and GeM................................................................. 105

Table 3.13: Gene-wide p-values for the most significant genes in the two Pearl et al. pathways showing significant enrichment in TRACK (Pearl et al., 2015) ..................................................................................................................... 105

Table 3.14: Effect of removing MSH3 on the Setscreen enrichment p-values for the top 14 GeM pathways in TRACK-HD, REGISTRY and the TRACK-REGISTRY meta-analysis. .................. 107

Table 3.15: Setscreen enrichment p-values for the large set of GeM pathways in TRACK-HD and REGISTRY ............................................................................................................................................. 109

Table 3.16: Summary of missing data in REGISTRY ......................................................................................................................... 112

Table 4.1: Characteristics of the polyglutamine diseases showing epidemiology, clinical features, and CAG repeat ranges of polyglutamine diseases ......................................................................................... 122

Table 4.2: Cohort characteristics: HD – Huntington’s disease; SCA – spinocerebellar ataxia; AAO – age at onset; SD – standard deviation...................................................................................................................... 127

Table 4.3: Characteristics of single nucleotide polymorphisms (SNPs) used in this study ...... 130

Table 4.4: Seed sense sequences for SNP KASP assay design .................................................. 133

Table 4.5: Effects of repeat length of the expanded allele on the age at onset. ..................... 134

Table 4.6: Results of combined analysis of SNPs ................................................................................................................................................................................................. 136

Table 5.1: Pathways with an association to age of onset in the GeM GWAS (p<0.05) that also are associated with HD progression (p<0.05) in the TRACK-HD WES analysis .................. 151

Table 5.2: Number of variants identified in cases showing an excess of rare variants in FAN1 compared to other genes in the Ch15 region of interest highlighted by the GeM-GWAS ...... 152

Table 5.3: FAN1 variants identified in fast (n=5) and slow (n=3) progressing subjects from the TRACK-HD cohort ........................................................................................................................................... 154

Table 5.4: Frequency of MSH3 variant rs184967 alleles in fast and slow progressors .......... 155

Table 5.5: Frequency of rs201874762 in TRACK-HD fast and slow progressors .................... 156

Table 5.6: Results from the Sanger sequencing of TRACK-HD cohort subjects, showing the expected genotypes based on the GWAS, and whether deletions were found ...................... 158

Table 5.7: Allelic sizes and frequencies at exon 1 of the hMSH3 gene in 58 unrelated Japanese individuals, from Nakajima et al (Nakajima et al., 1995) ............................................................................. 166

Table 6.1: Modified Goldman scoring system. FHx: Family History. AAO: Age At Onset of symptoms .................................................................................................................................................... 173

Table 6.2: Age at onset and genetic results of C9orf72 expansion positive cases ............. 176
Table 6.3: Summary of the clinical features of ten C9orf72 expansion-positive cases. UMN = upper motor neuron. .............................................................................................................................. 179

Table 6.4: Phenotypic features of C9orf72 negative & positive cases within HD phenocopy cohort, and outcome of Fisher’s exact test to test for association between clinical feature and genetic test outcome. ........................................................................................................................................ 180

Table 7.1: Track-HD and Leiden cohorts for RNA-Seq analysis. ................................................................. 192

Table 7.2: Differential expression of transcripts for the TRACK-HD manifest HD vs premanifest HD samples showing that there are no individually significant differentially expressed transcripts. ...................................................................................................................................... 199

Table 7.3: Differential expression analysis in HD (premanifest and manifest combined) versus controls for the combined Track-HD and Leiden cohorts................................................................................................................................. 200

Table 7.4: Overlap analysis of Track-HD and LUMC cohorts shows that a significant excess of pathways are associated with HD (p < 0.05) in both datasets. .............................................................. 201

Table 7.5: 53 ‘generic’ pathways which are significantly upregulated in HD versus control blood GSEA......................................................................................................................................................... 205

Table 7.6: 14 ‘generic’ pathways which are significantly downregulated in HD versus control blood GSEA......................................................................................................................................................... 207

Table 7.7: The 10 most significantly dysregulated genes (p<0.01) in up or downregulated generic pathways (q<0.05)......................................................................................................................... 210

Table 7.8: Number of pathways nominally significantly enriched (uncorrected p<0.05) in both the combined Track-HD/Leiden blood dataset and the unstimulated myeloid data of Miller et al. (2016a) ........................................................................................................................................... 211

Table 7.9: Pathways significantly (p<0.05) upregulated in both the combined Track-HD and Leiden whole blood data and the unstimulated myeloid cell dataset of Miller et al. (2016a) 213

Table 7.10: All WGCNA brain expression modules significantly dysregulated (p < 0.05) in both Track-HD and Leiden datasets in HD versus control blood. ..................................................................................................................... 218

Table 7.11: Ten most significantly dysregulated genes (p<0.05) from the WGCNA brain expression modules that were dysregulated (up or down) in HD blood................................................................................ 219

Table 7.12: Brain expression modules significantly dysregulated both in HD brain and HD blood .................................................................................................................................................................. 220

Table 7.13: Ten most significantly upregulated and downregulated generic pathways in both HD blood and prefrontal cortex.......................................................................................................................... 226

Table 7.14: Correlation between gene expression and TMS in gene positive Track-HD subjects. .............................................................................................................................................................................. 227
Table 7.15: Enrichment of up or downregulated pathways from HD vs. control blood with TMS in the combined Track-HD and Leiden cohort. p(combined-diffexp) – enrichment p-value for upregulated genes in the combined Track-HD and Leiden sample .......................................................... 230

Table 7.16: Enrichment of modules from HD vs control blood (Table S9) with TMS in the combined Track- HD and Leiden cohort. ......................................................................................... 234

Table 7.17: Correlation between genes differentially expressed in HD from Mastrokolias et al (Mastrokolias et al., 2015) and TMS in the Track-HD gene positive subjects. ....................... 235

Table 7.18: WGCNA co-expression modules from the Gibbs et al. (2010) control brain expression dataset significantly associated with late-onset Alzheimer’s disease (LOAD) in the IGAP GWAS are upregulated in HD blood .................................................................................................. 236

Table 7.19: Differential expression analysis with rate of HD progression in gene positive members of the TRACK-HD cohort. ............................................................................................................. 239

Table 7.20: Relationship between generic pathways and rate of HD progression showing that while there are multiple pathways significantly downregulated with faster progression, but there are no pathways significantly upregulated with faster progression ................................................. 242

Table 7.21: Cell cycle pathways are enriched in GOrilla analysis of ranked transcripts from the TRACK-HD progression differential progression analysis ........................................................................ 245

Table 7.22: Correlation enrichment between HD modules from Neueder & Bates (Neueder and Bates, 2014) and differential transcription according to progression .................................................. 247

Table 7.23: Correlation enrichment between Gibbs modules(Gibbs et al., 2010) and differential transcription according to progression ........................................................................................................ 249

Table 7.24: Differential transcription of transcripts according to atypical severity score from the LUMC cohort ................................................................................................................................. 251

Table 7.25: Ten pathways most enriched in a GOrilla pathway analysis of the differential transcription in the LUMC samples according to cross-sectional severity score ......................... 251
**Chapter 1: Introduction**

**1.1 Genes and disease**

To better understand the pivotal relationship between genotype and phenotype is core to modern biology, and study of factors that control the form of organisms has transformed over the past 200 years from Lamarckian views on inheritance of acquired characteristics, to the role of the gene in the neodarwinian synthesis, to the use of large scale multi-omic studies. Susceptibility to disease is a phenotypic attribute which may be influenced by genes, and understanding these genetic influences on disease has the potential to illuminate pathogenesis. By better understanding the molecular cellular processes underpinning disease we may be able to define treatment targets. This, combined by increased affordability of large scale genetic studies, has led to a burgeoning interest in disease genetics.

Despite their immense public health burden neurodegenerative diseases remain poorly understood in terms of basic biology, and we lack treatments to prevent or slow them. In this thesis I focus on a set of neurodegenerative diseases caused by repeat expansion mutations: primarily Huntington’s disease (HD), but also the polyglutamine spinocerebellar ataxias (SCAs) and C9orf72 associated Amyotrophic Lateral Sclerosis / Frontotemporal dementia (ALS/FTD). While the disease-causing mutations have been identified for these conditions, there is variability in how the symptoms developed and how they progress. It is hoped that characterizing and understanding this phenotypic variability will be clinically valuable, and assist in finding drugs for these devastating and currently incurable diseases through the identification of genes and pathways amenable to therapeutic manipulation.

A common theme among the diseases discussed in this thesis is that they are associated with expansions in tracts of repetitive DNA: Huntington’s disease and the polyglutamine SCAs are associated with CAG repeat expansions and C9orf72 associated ALS/FTD is associated with a GGGGCC repeat expansion. Tandemly repeated DNA is a common feature of eukaryotic genomes and is also seen in prokaryotes (Bichara et al., 2006), and are thought to have arisen by expansion of a progenitor sequence. Repetitive DNA elements make up a substantive portion of the genome in many organisms, including humans where estimates suggest that this represents >65% of the genome (de Koning et al., 2011). There are various types of repetitive elements, ranging from microsatellites up to whole genes (Figure 1). Microsatellites are the shortest type of tandem repeats, they are usually <150 base pairs, and the repeat unit is usually 4bp or less but can be up to 6bp, typically repeated 10-20 times. Microsatellites
with a CA repeat make up 0.5% of the genome. Over 30 human developmental and neurodegenerative diseases are caused by expansion of unstable microsatellite sequences (McMurray, 2010): HD, polyglutamine SCAs and C9orf72 associated ALS/FTD among them.

**Figure 1.1:** Types of DNA within the human genome. Types of DNA within the human genome. Tandem repeats tend to be located in blocks at one or more locations on chromosomes. Interspersed repetitive sequences may be widespread over the genome, located over broad regions of one or more chromosomes. Bp: base pair. LINE: long interspersed element; SINE: short interspersed element.

### 1.2 Huntington’s disease

#### 1.2.1 Clinical characteristics and prevalence

Huntington’s disease is the most common genetically determined neurodegenerative disease with a prevalence of at least 12.4 per 100,000 people in the UK (Rawlins, 2010a). It is an autosomal dominant neurodegenerative condition caused by a CAG repeat (translated to polyglutamine) expansion in exon 1 of the gene encoding huntingtin (HTT), and is typically characterised by a triad of psychiatric, movement and cognitive impairment. HD can produce a wide range of phenotypic presentations, and as the disease progresses, the signs and symptoms change. Symptoms usually develop between 35-45 years of age, but onset has been described between 2-87 years. The disease progresses inexorably and, with the exception of late-onset cases, is uniformly fatal a median of 18 years from motor onset (Ross et al., 2014). The highest prevalence in the world is in Venezuela near Lake Maracaibo: 700
per 100 000, and it is the collaboration of people in this region and an international group of researchers that was crucial in the identification of the HD gene.

1.2.2 Motor features
The cardinal motor symptoms of HD are chorea and dystonia which are present in 90% and 95% of symptomatic patients respectively (Wild and Tabrizi, 2012, Louis et al., 1999). Gait is impaired, not only due to the chorea and dystonia, but also due to impairment in motor control and postural reflexes, making patients prone to falling. Hypophonia, dysarthria and dysphagia all cause significant morbidity. Dysphagia with choking episodes is reported even in early disease. Eye movement abnormalities occur early. As the disease progresses head thrusting is used to initiate gaze shifts, pursuit is impaired with saccadic instructions and there is gaze impersistence.

1.2.3 Psychiatric features
Psychiatric problems, particularly anxiety and depression, are a common and major cause of morbidity in HD and may occur many years before symptom onset (Paulsen et al., 2005). Psychosis is relatively rare, additional familial factors may predispose to schizophrenia-like symptoms in HD (Lovestone et al., 1996). Hypomania, and more rarely mania is seen (Craufurd and Snowden, 2002).

Irritability is common (65.4%) (Paulsen et al., 2001a) and some patients become aggressive. Apathy is prevalent in both symptomatic HD (55.8%) (Reedeker N, 2011, Paulsen et al., 2001a), and prior to motor onset (Duff et al., 2010). Obsessions and compulsions can be features of the disease.

1.2.4 Cognitive features
The severity of cognitive involvement in HD is variable, and becomes more prevalent and marked as the disease progresses. Cognitive deficits are particularly apparent in executive functioning, and also attention, verbal fluency, psychomotor speed, memory and visuospatial functioning (Brandt and Butters, 1986, Craufurd and Snowden, 2002).

There often are subtle cognitive differences detectable more than a decade prior to predicted motor onset, which gradually decline as motor onsets approaches (Paulsen et al., 2008, Paulsen et al., 2001b, Stout et al., 2012, Tabrizi et al., 2012) (Figure 1). There are abnormalities on MRI such as caudate atrophy which can be seen in cross-sectional studies up to 15 years prior to predicted motor onset (Tabrizi et al., 2009a, Tabrizi et al., 2012). The
presence of longitudinal change in premanifest disease enables disease progression to be assessed, even before a patient has overt symptoms.

**Figure 1.2:** Longitudinal changes in cognitive measures from the Track-HD study over 24 months. Significant change differences relative to controls over 0-12, 12-24, and 0-24 months are represented by *p<0.05, **p<0.01 and ***P<0.001. Groups determined at start of study; PreHD-A: more than 10.8 years from predicted onset; PreHD-B: less than 10.8 years from predicted onset; HD1: early HD & less symptomatic on total functional capacity scale (TFC); HD2: early HD and more symptomatic on TFC. Adapted from (Tabrizi et al., 2012), Image reproduced with permission of the rights holder, Elsevier Inc.

### 1.2.5 Disease onset

By consensus, disease onset is defined as the point when a person who carries a CAG-expanded HTT allele develops ‘the unequivocal presence of an otherwise unexplained extrapyramidal movement disorder’ (eg chorea, dystonia, bradykinesia, rigidity) (Huntington’s et al., 1993, Hogarth et al., 2005). However, the transition from premanifest to manifest HD is not abrupt, making the clear delineation of this event more challenging than previously assumed, and more open to individual physician or investigator interpretation. There may be more subtle features evident to the careful observer prior to this in the peri-symptomatic or prodromal phases of HD. These include delayed initiation of saccades, slower saccades particularly on vertical eye movements, irregular finger tapping and a generalised restlessness. Psychiatric symptoms and cognitive changes often occur before motor onset (Tabrizi et al., 2009a, Tabrizi et al., 2011, Tabrizi et al., 2012).

The lack of clear transition between premanifest and manifest states, combined with different approaches from clinicians about making a formal diagnosis of manifest HD have led to concerns with using age at onset data. Lahiri (Lahiri, 2013) found that motor AAO in the very closely monitored TRACK-HD study is two years earlier than the less intensive EHDN Registry
study; this difference remains significant when analysis is restricted to matching populations, and is not accounted for by CAG.

### 1.2.6 HD Genetics

HD is inherited in an autosomal dominant manner, and is caused by a trinucleotide CAG repeat expansion in the huntingtin (HTT) gene on the short arm of chromosome 4 at 4p16.3. The expansion is translated into a polyglutamine stretch in the mutant Huntingtin protein (mHTT).

Non-disease-associated alleles vary from 10 to 35 repeats, whilst disease-associated alleles exceed 35 CAG repeats, with penetrance increasing to ~100% by 40 repeats (Quarrell et al., 2007) (Table 1.1). Up to 121 CAG repeats have been reported, but there is a marked skew to the right in the distribution and most people have 40-44 repeats (Langbehn et al., 2004).

<table>
<thead>
<tr>
<th>CAG repeat length</th>
<th>&lt;27</th>
<th>27 – 35</th>
<th>36 – 39</th>
<th>≥40</th>
<th>≥55 – 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical manifestation</td>
<td>Normal</td>
<td>Intermediate repeat allele</td>
<td>Reduced penetrance but pathogenic</td>
<td>Fully penetrant</td>
<td>Usually have juvenile onset</td>
</tr>
<tr>
<td></td>
<td>Not generally pathogenic</td>
<td>May expand into disease range in future generations in paternal line</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 1.1: Relationship between size of CAG repeat expansion and clinical outcome.*

### 1.2.7 Role of CAG repeat length in the phenotype of Huntington’s disease

There is an inverse relationship between the size of the CAG repeat and the onset and clinical manifestation of HD as outlined in Table 1.1, with those with very high CAG repeats developing a severe, juvenile onset form of the disease. Age of onset (AAO) of HD has a genetic component with 50-70% of the variance attributable to HTT CAG repeat length (Duyao et al., 1993, Brinkman et al., 1997, Wexler et al., 2004b, Langbehn et al., 2004).

### 1.2.8 Disease burden score and cumulative probability of disease onset

In order to explore Huntington’s disease related changes over time several approaches have been used to encapsulate the expected burden of pathology, relative to the subject’s age and CAG repeat score. The most notable of these are the ‘disease burden score’ and the ‘cumulative probability of disease onset’. The disease burden score is relatively calculable
(DBS= age x [CAG-35.5]) but is based on a small number of neuropathological samples (Penney, 1997, Sanchez-Pernaute et al., 1999).

A more widely used measure of the combined contributions of age and CAG on when an individual will develop onset is the parametric survival model developed by Langbehn and colleagues (Langbehn et al., 2004). This is based on a cohort of 2913 manifest and premanifest HD patients. The model predicts the probability of motor symptom onset at different ages for individual patients with narrow confidence intervals (Figure 1.3).

Advantages include being based on a large population sample, making no assumption of linearity, and taking into account the current age of a subject when predicting their future onset probabilities.

**Figure 1.3:** Cumulative probability of Huntington’s disease onset curves for various CAG lengths. Numbers indicate CAG repeat length. Cum. prob. onset = Cumulative probability of onset of Huntington’s disease. From (Langbehn et al., 2004), Image reproduced with permission of the rights holder, Wiley-Blackwell.

### 1.2.9 Intergenerational and somatic instability of the HTT CAG repeat

CAG repeat lengths vary from generation to generation, with both expansion and contraction of the number of repeats occurring, but with an overall tendency towards expansion. Large expansions are associated with transmission down the male line (Telenius H, 1993), and there is a familial tendency towards large expansions. The tendency of the CAG expansion to expand during transmission underlies the phenomenon of anticipation observed in Huntington’s and other neurodegenerative conditions such as SCAs 1, 2, 3, 6, 7 and DRPLA.
The HD CAG repeat expansion also exhibits somatic mosaicism which tends to be expansion-biased and age-dependent (Kennedy et al., 2003). Repeat instability is also found in other repeat disorders such as myotonic dystrophy type 1 (DM1), SBMA and SCAs 1, 2, 3, 7, 12 (further detail in Chapter 4). Much of the work on repeat instability has been done on DM1 model systems; DM1 is a multisystem disorder caused by an expanded CTG repeat (CAG on the non-coding strand) in the 3’-untranslated region of the DM protein kinase (DMPK) gene. Somatic instability is tissue-specific, with particularly high levels found in striatum and cortex of people with HD (Kennedy and Shelbourne, 2000, Kennedy et al., 2003), but also is observed in liver (Tome et al., 2013a). CAG repeat instability occurs in terminally differentiated, post mitotic neurons in several HD mouse models (Gonitel et al., 2008) suggesting a replication independent mechanism. Striatum, the brain area most affected in HD, exhibits the highest levels of CAG somatic instability in both mouse models and humans, whereas CAG expansion is minimal/absent in the cerebellum (Halliday et al., 1998, Telenius et al., 1994, Kennedy et al., 2003, Kennedy and Shelbourne, 2000). Striatal HTT CAG repeat size instability increased in an expansion-biased and age-dependent manner (Kennedy and Shelbourne, 2000). Notably, the degree of somatic expansion of the CAG repeat in HD patient brain predicts onset (Swami et al., 2009).

1.3 C9orf72 associated Amyotrophic Lateral Sclerosis / Frontotemporal dementia

1.3.1 Clinical Features

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterized clinically by upper and motor neuron weakness causing progressive paralysis leading to death from respiratory failure, typically within two to three years of symptom onset (Kiernan et al., 2011). Frontotemporal dementia (FTD) is a clinically and pathologically heterogeneous group of non-Alzheimer dementias characterised collectively by relatively selective, progressive atrophy involving the frontal or temporal lobes, or both (Warren et al., 2013). There are three main clinical syndromes of FTD: behavioural variant FTD, primary progressive aphasia and semantic dementia, and there is variable overlap between the syndromes, atypical parkinsonism and motor neuron disease.

1.3.2 Genetics
There were several genes known to be associated with FTD including MAPT (microtubule-associated protein tau) and progranulin (GRN) when, in 2011, an expanded hexanucleotide GGGGCC repeat in the C9orf72 gene was identified in large kindreds with FTLD and ALS (DeJesus-Hernandez et al., 2011, Renton et al., 2011). This expansion is now recognised as the commonest genetic cause of ALS and FTLD in many populations (DeJesus-Hernandez et al., 2011, Renton et al., 2011, Smith et al., 2012, Mahoney et al., 2012).

In addition to C9orf72, mutations in SOD1, TARDBP, FUS, ANG, ALS2, SETX, and VAPB also cause familial ALS and contribute to the development of sporadic ALS. The spectrum of genes causing ALS / FTD, and the pathological inclusions observed are summarized in Figure 1.4. While multiple pathways are involved in disease initiation and progression in ALS and FTD, RNA homeostasis has emerged as as a convergent underlying mechanism between ALS and FTD (Ling et al., 2013).

A. Genetics of ALS and FTD

![Genetics of ALS and FTD](image)

B. Pathological inclusions in ALS and FTD

![Pathological inclusions in ALS and FTD](image)

Figure 1.4: Clinical, genetic and pathological overlap of ALS and FTD. (A) ALS and FTD represent a continuum of a broad neurodegenerative disorder with each presenting as extremes of a spectrum of overlapping clinical symptoms (ALS in red and FTD in purple). Major known genetic causes for ALS and FTD are plotted according to the ratio of known mutations that give rise to ALS or FTD. (B) Pathological protein inclusions in ALS and FTD,
according to the major protein misaccumulated. Inclusions of TDP-43 and FUS/TLS in ALS and FTD reflect the pathological overlap of ALS and FTD. From (Ling et al., 2013), image reproduced with permission of the rights holder, Cell Press.

### 1.4 The Spinocerebellar ataxias

The spinocerebellar ataxias are a heterogeneous group of genetic disorders united by occurrence of slowly progressive incoordination of gait, fine motor skill tasks, speech, and eye movements (Table 4.1) (Harding, 1984, Jayadev and Bird, 2013). Seven SCAs are due to CAG repeat expansions, collectively known as the polyglutamine spinocerebellar ataxias. The phenotypes of these diseases differ. Atrophy of the cerebellum is observed on a frequent basis, and the dysfunction of the cerebellum and its associated systems is at the core of the clinical symptoms, thus ataxia is a cardinal feature. There is variable involvement of additional systems leading to changing frequencies of accompanying features such as optic atrophy, neuropathy, retinopathy, extrapyramidal and pyramidal symptoms, seizures, intellectual disabilities, dementia, sensorineural deafness, endocrine manifestations and more (Kawai et al., 2009, Jayadev and Bird, 2013).

Common features among the polyglutamine spinocerebellar ataxias include autosomal dominant inheritance, genetic anticipation, disruption of the normal conformation and function of the protein above a threshold repeat size, neuronal involvement and intracellular inclusions containing the cognate polyglutamine protein. The nature and temporal and regional expression pattern of the repeat-containing proteins probably leads to the clinical variability between these diseases, but the substantial phenotypic variation seen within each disease remains only partly explained (Gatchel and Zoghbi, 2005).

A more extensive discussion of the polyglutamine spinocerebellar ataxias along with a table summarizing the clinical characteristics of each disease and the causative mutation is given in the Introduction to Chapter 4.

### 1.5 Genetic analysis

There are various approaches which can be used to determine what chromosomal location or gene is responsible for a particular phenotype. The typical genetic study involves collecting a sample of subjects with phenotypic information, genotyping these subjects and then analysing the data to determine whether the phenotype is related to the genotypes at various loci (Sham and Purcell, 2014). Genetic linkage analysis was used for years to identify many
disease causing genes including the Huntingtin gene (Huntington's et al., 1993, Gusella, 1984): it is well suited to identifying the genetic underpinnings of Mendelian disorders which are largely caused by protein-coding changes with large effect sizes (Botstein and Risch, 2003). Linkage analysis is based on the observation that genes that reside physically close on a chromosome remain linked during meiosis, and can be quantified using a LOD score. This technique was developed by Newton Morton, and compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance (Morton, 1955). Genetic maps were made by looking at associations between genetic variants and diseases or traits, with the distances given in recombination units (the centiMorgan [cM]).

To determine whether an association is statistically significant various approaches have been used, the most popular is the frequentist significance testing approach, which was proposed by Fisher (Fisher, 1925) and further developed by Neyman and Pearson (Neyman and Pearson, 1933). While some point to the limitations of the use of p-values, and argue for a Bayesian approach given that it provides a more natural and logically consistent framework for drawing statistical inferences, the requirements for prior distributions to be specified for model parameters and intensive computation make this challenging (Sham and Purcell, 2014). Ensuring that a study has sufficient statistical power to detect an association is important: the probability of rejecting $H_0$ when the alternative hypothesis ($H_1$) is true is formalized as the statistical power in the Neyman–Pearson hypothesis testing framework. Technological advances mean that we are now able to adopt unbiased approaches in genetic analysis, however maximizing power for a given amount of sequencing/genotyping remains important. Many factors influence the statistical power of genetic studies. Some are outside the investigator’s control including the complexity of the genetic architecture of the phenotype, the effect sizes and allele frequencies of the underlying genetic variants, the inherent level of temporal stability or fluctuation of the phenotype, and the history and genetic characteristics of the study population (Sham and Purcell, 2014). While factors the investigator may manipulate to boost study power include the selection of study subjects, sample size, methods of phenotypic and genotypic measurements, and methods for data quality control and statistical analyses (Sham and Purcell, 2014). Thus optimal subject selection and careful phenotyping can boost study power as well as increasing sample size. In this thesis I have used the approach of careful subject selection and deep clinical phenotyping to facilitate genetic analysis.
Genetic variants are variable in both their risk allele frequency and the strength of genetic effect they have on phenotype/risk of disease (Figure 1.5), meaning that different techniques are variably suited to identifying variants with these effect/frequency profiles.

**Figure 1.5**: Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). From (Manolio et al., 2009), image reproduced with permission of the rights holder, Nature Publishing Group.

Whole exome sequencing uses Next Generation Sequencing technologies to provide sequence information on the protein-coding genome with high coverage. It is well suited for the identification of variants when there is substantial locus heterogeneity, to identify rare structural or coding variants of relatively large effect. While many WES studies adopt the trio design to filter out non-causative variants, in the exploratory study described in Chapter 5 I used a case control design to see if any variants were enriched in people who progress rapidly with HD compared with more slowly progressing subjects.

Linkage based approaches have had limited success in complex diseases due to their low power and resolution for variants of moderate or small effect (Pulst, 1999, Sham and Purcell, 2014). Candidate gene studies can be used to detect association between genetic variation within pre-specified genes of interest and phenotypes or disease states, the work in Chapter 4 of this thesis is an example of a candidate gene study. While candidate gene studies have been used to investigate complex traits, but by their very nature are incapable of identifying new molecules or pathways, and are at best a way of ‘proving’ a suspected molecules’ candidature (Gandhi and Wood, 2010).
The technique of looking at associations between hundreds of thousands of common genetic variants (polymorphisms) in the genome with a phenotype or disease status in people: the genome wide association study (GWAS), is based on the principle that common allelic variation(s) in a population will underlie the heritability of common diseases. In Chapter 3 of this thesis I present the findings of a GWAS to identify modifiers of disease progression in HD. GWASes have yielded many important findings over the past 13 years since the discovery of a Complement Factor H Polymorphism associated with Age-Related Macular degeneration (Klein et al., 2005). Some of the results were surprising, and highlighted areas of biology which are critical to the pathogenesis of the disease. While early GWASes tended to be small, over time the number of subjects included has grown with the number of GWAS published (Figure 1.6). A turning point for GWAS came in 2007 with the seminal Wellcome Trust Case Control Consortium (WTCCC) study (Wellcome Trust Case Control, 2007), which compared the sequences of hundreds of thousands of common genetic variants in people with and without seven diseases to look for variants associated with these diseases. This study had large sample sizes (2000 in each disease group and 3000 shared controls), necessarily requiring a high level of collaboration between groups (Wellcome Trust Case Control, 2007). Since then the trend towards increasing sample size has continued: while Klein et al (Klein et al., 2005) detected an association with just 96 cases and 50 controls, many more recent studies have over 200,000 subjects (Manolio, 2017).

Figure 1.6: Ever-increasing sample sizes for genome-wide association studies (GWAS). This graph shows the cumulative number of GWAS involving 10,000 samples or more published per year, with those involving different sample sizes indicated in different colours. Graph from
The reason for the increased size for these genome wide association studies can be understood if one considers the genetic architecture of the traits for which associations are being sought. If the effect size of the genetic variant is large, it requires fewer samples to be significantly detected; while if the effect size is small, more samples are needed. As increasing numbers of GWASes were performed it became evident that even the most important loci in the genome have small effect sizes, and for some time people were perplexed that the significant hits only explain a modest fraction of the predicted genetic variance. This was referred to as the mystery of the “missing heritability” (Manolio et al., 2009, Maher, 2008).

The concept of missing heritability is based on the observation that the portion of phenotypic variance in a population attributable to additive genetic factors: the heritability, is higher than the combined contribution of identified genetic factors. For example height has about 80% heritability (Visscher et al., 2006), but 40 loci associated with height were found to explain around 5% of the phenotypic variance despite studies of tens of thousands of people (Visscher, 2008).

It has been observed that common single-nucleotide polymorphisms (SNPs) with effect sizes well below the genome wide statistical significance level account for most of the “missing heritability” of many traits (Yang et al., 2010a, Shi et al., 2016, Boyle et al., 2017). These SNPs are frequently noncoding variants that are thought to affect gene regulation, which is subject to many stages and influences (Pickrell, 2014, Li et al., 2016, Hardy and Singleton, 2009).

Using a network model Pritchard and colleagues (Boyle et al., 2017) explain that for a variety of traits, the largest-effect variants are modestly enriched in specific genes or pathways that may play direct roles in disease. These are ‘core’ genes and pathways and their direct regulators: modest in number and with specific roles in disease aetiology. Core genes are likely to be those that harbour common variants with large clinical or biological effects, and genes with a series of disease-associated alleles. They are also the genes most likely to be amenable to targeting therapeutically. Rather than coming from core genes, the SNPs that contribute the bulk of the heritability tend to be spread across the genome and are not near genes with disease-specific functions. This is described as the omnigenic model. Given the key role of core genes, one could argue that variants only picked up with extremely large GWASes may be less relevant to therapeutic development.
The idea that many variants effect phenotype is not a new one: the “infinitesimal model” of complex trait genetics was established by Fisher (Fisher, 1918). The omnigenic model is an extension of the infinitesimal model, differing primarily in the mechanistic hypothesis as to how a large number of genes with small effects act to influence disease: via regulatory networks that act outside of core genes (Plenge, 2017). Plenge suggests that the Omnigenic model has important implications for drug discovery and development: (1) “core genes” represent good drug targets (2) regulatory networks identified by “peripheral genes” point to specific cell types and mechanism that can be used for phenotypic screens; and (3) new approaches are needed to drug cellular networks as the bulk of drug discovery today is an attempt to reduce complex mechanisms to individual drug targets.

In addition to having relevance to how we look for associations between genetic variants and disease in GWAS, through a focus on regulatory networks and expression the Omnigenic model also points to the value of integrating transcriptomic and genetic analysis: something that I have done in this thesis.

1.6 Previous work on Genetic Modifiers of Huntington’s disease

Though the primary determinant of the Huntington’s disease phenotype is the CAG repeat length, kindred studies suggest at least 40% of the residual age-of-onset variability not accounted for by disease burden (age(CAGn-35.5)) is determined by other genetic factors (Djousse et al., 2003, Wexler et al., 2004b, Penney, 1997). Around two-thirds of the rate of functional, motor, and cognitive progression in HD is determined by the same factors that also determine age at onset, with CAG repeat–dependent mechanisms having by far the largest effect, while around a third of the factors governing progression differ from those determining onset (Aziz et al., 2018, Rosenblatt et al., 2012, Rosas et al., 2011).

The huntingtin gene itself has been a region of interest in the search for factors that modify Huntington’s disease. Djousse et al (Djousse et al., 2003)’s work suggested that the smaller wild-type HTT allele influences onset in people with large HD repeat sizes (CAG of 47-83), but careful statistical analysis revealed that the methods used were prone to false-positive results due to susceptibility to outliers (Guesella et al., 2014, Ramos et al., 2012), and a more statistically rigorous study of more than 4000 subjects demonstrated no impact of the size of the smaller CAG repeat (Lee et al., 2012b).

Other sequence variation at the HTT locus has been defined beyond the polymorphic/expanded CAG repeat, including differences that alter the coding sequence
(including a polymorphic CCG repeat following the CAG repeat, a deletion polymorphism at codon 2642 (Novelletto et al., 1994)), the transcript’s untranslated sequence, intron sequences, and sequences flanking the centromeric and telomeric ends of the gene. These have been used to define HTT haplotypes (Wall and Pritchard, 2003), each of which represents the group of sequence variants found on a particular chromosome that is passed on largely intact to subsequent generations because of the lack of recombination events in this relatively small segment of the genome (Lee et al., 2012b). The haplotypes that carry expanded alleles in HD subjects have revealed that approximately 50% of Europeans with HD share a common ancestor, but that multiple independent mutations occurring on different chromosomal backbones account for the rest, in both people of European and non-European backgrounds (Kay et al., 2016a, Kay et al., 2016b). None of the most frequent haplotypes, either on HD chromosomes or on the normal chromosomes in HD heterozygotes, appears to modify age at motor diagnosis. Thus, natural sequence variation at HTT has not thus far been shown a major source of disease modification in HD (Lee et al., 2012b).

Moving from the HTT gene itself to its regulatory regions, work by Djousse et al suggested the presence of an AAO modifier in HD to be linked to the HD gene itself in 4p16 (Djousse et al., 2004). Bečanović et al identified a SNP in the HTT promoter which alters NF-κB binding and regulates HTT promoter transcriptional activity, and is associated with age at onset in HD (Becanovic et al., 2015). The rs13102260 minor variant on the HD disease allele was associated with delayed age of onset in a set of familial cases, whereas the presence of the rs13102260 (A) variant on the wild-type HTT allele was associated with earlier age of onset in HD patients in an independent extreme-based cohort.

Early studies looking for HD modifiers took a candidate gene approach, while various modifiers were proposed, no results were consistently replicated in larger studies. Candidate variants included:

- A polymorphic TAA repeat in the 3’UTR of GRIK2, the Glutamate receptor subunit (Rubinsztein et al., 1997, Zeng et al., 2006, Lee et al., 2012a).
- Apolipoprotein E (APOE) (Panas et al., 1999)
- Gln-Ala repeat length in the transcriptional co-activator CA150 (Holbert et al., 2001)
- Ser18Tyr polymorphism in the Ubiquitin carboxy-terminal hydrolase L1 (UCH-L1), an abundant neuron-specific deubiquitinating enzyme in the proteasome pathway (Naze et al., 2002, Metzger et al., 2006)
- Val471Ala polymorphism in the autophagy-related gene ATG7 (Metzger et al., 2013, Metzger et al., 2010)
• Val66Met polymorphism in the neurotrophic factor BDNF (Alberch et al., 2005)
• Several variants within the mitochondrial regulator PPARG1A (PGC-1 alpha) and its downstream transcription factors NRF-1 and TFAM (Taherzadeh-Fard et al., 2011, Ramos et al., 2012, Weydt et al., 2009, Che et al., 2011)
• Cys1976Thr polymorphism in the ADORA2A gene which encodes an adenosine receptor (Dhaenens et al., 2009)

A key study in the understanding of genetic modifiers of HD, and which largely superseded previous studies, was that by the Genetic Modifiers of Huntington’s disease (“GeM-HD”) Consortium study which looked at genetic modifiers of age of motor onset (GeM-HD-Consortium, 2015). In a study of 4082 people with Huntington’s disease they identified three genome-wide significant loci, one on chromosome 8 and two on chromosome 15, these are thought likely to be associated with RRM2B and FAN1, respectively. The chromosome 8 locus hastens onset by 1.6 years, while conditional analysis revealed that the effects at the chromosome 15 locus hasten or delay onset by 6 or 1.4 years respectively. Pathway analysis in this study implicated DNA handling in Huntington’s disease modification, as did near-significant association at the DNA repair gene MLH1.

1.7 DNA repair and Somatic Instability

As mentioned above, the CAG repeat tract is subject to somatic instability. Microsatellites, which like the HTT CAG repeats, are short tandem repetitive DNA elements, and are particularly susceptible to replication errors caused by DNA polymerase slippage over the repeat sequence (Mirkin, 2007). These errors are repaired by mismatch repair pathways (MMR), and are frequently observed in colon cancers where MMR proteins are deficient (Goellner et al., 1997). Evidence, primarily in mouse models, links somatic instability in repeat disorders to DNA mismatch repair proteins (Manley et al., 1999, Foiry et al., 2006, Dragileva et al., 2009, Kovalenko et al., 2012, Pinto et al., 2013, Mason et al., 2014, Pluciennik et al., 2013, Iyer et al., 2015, Wheeler et al., 2003, Tome et al., 2013a).

DNA mismatch repair is a conserved process that stabilizes the genome by correcting DNA replication errors (specifically of base-base mismatches and insertion and/or deletion loops), attenuating chromosomal rearrangements, and mediating the cellular response to certain types of DNA damage (Iyer et al., 2015).

There is a high level of interconnectedness between pathways involved in the DNA damage response, with proteins being involved in numerous pathways (Pearl et al., 2015).
example, MMR factors are also required for the repair of mismatches in heteroduplex DNA (hDNA) that form as a result of sequence heterologies between recombining sequences (Evans and Alani, 2000), and MMR also acts to inhibit recombination between moderately divergent (homeologous) sequences (Rogacheva et al., 2014, Evans and Alani, 2000). Similarly FAN1, a protein highly implicated by the GeM GWAS study as mentioned above (GeM-HD-Consortium, 2015) was initially linked to interstrand cross-link repair, but also interacts with MLH1, a protein generally linked to MMR.

Mismatch repair proteins have also been linked to disease progression/onset in model systems of Huntington’s disease and other repeat disorders, as will be discussed further in Chapter 3 (Wheeler et al., 2003, Kovalenko et al., 2012).

**Aims of this Thesis**

The overarching aim of this thesis is to better understand the genetic factors underpinning phenotypic diversity in neurodegenerative diseases, particularly those caused by repeat expansion mutations. Specifically, this thesis will:

1. Identify genetic modifiers of progression in people with Huntington’s disease using genome wide association analysis (Chapter 3)
2. Investigate whether DNA repair variants implicated as modifiers of age at onset in Huntington’s disease also modify onset in the polyglutamine spinocerebellar ataxias (Chapter 4)
3. Look for rarer variants of large effect modifying progression in Huntington’s disease using whole exome sequencing (Chapter 5)
4. Examine loci highlighted by genetic analysis (Chapter 5)
5. Examine the intergenerational stability of the C9orf72 repeat in families with normal range repeat lengths (Chapter 6)
6. Determine whether the repeat expansion in the C9orf72 associated with frontotemporal lobar degeneration (FTLD) and amyotrophic lateral sclerosis (ALS) also cause HD phenocopy presentations (Chapter 6)
7. Investigate the effect of disease status and stage on the transcriptome of Huntington’s disease expansion mutation carriers (Chapter 7).
8. Examine whether there is a transcriptomic signature associated with altered rate of progression in Huntington’s disease (Chapter 8).
Chapter 2: General Methods

2.1 Consent and ethics

All studies mentioned in this thesis were carried out at approved research institutions. Ethical approval to undertake these analyses was given by the local NHHN/ION, or University College London (UCL)/UCL Hospitals, Joint Research Ethics Committee. All experiments were carried out in accordance with the declaration of Helsinki, and informed consent for genetic studies was obtained from all participants.

2.2 Standard assessments commonly used to examine Huntington’s disease which are employed in this thesis

2.2.1 Total Functional Capacity

The Total Functional Capacity (TFC) Scale (Shoulson and Fahn 1979) is used crudely to ‘stage’ the progression of HD (Table 2.1). The scale reflects the progression of the disease, in particular the psychosocial and functional effects on the patient and their family. Points are assigned according to the individual’s ability to work, to manage money, to perform household chores, to perform activities of daily living, and to live at home or in supervised care.

<table>
<thead>
<tr>
<th>Stage</th>
<th>TFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early HD</td>
<td>1</td>
</tr>
<tr>
<td>Early HD</td>
<td>2</td>
</tr>
<tr>
<td>Moderate HD</td>
<td>3</td>
</tr>
<tr>
<td>Advanced HD</td>
<td>4</td>
</tr>
<tr>
<td>Advanced HD</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2.1: Total Functional Capacity Scale. HD: Huntington’s disease.

2.2.2. Unified Huntington’s Disease Rating Scale (UHDRS)

The UHDRS was developed by the Huntington Study Group as a clinical rating scale to assess four domains of clinical performance and capacity in HD: motor function, cognitive function, behavioural abnormalities and functional capacity (Group, 1996b).

2.2.2.1 UHDRS Functional assessment

The UHDRS Functional capacity score which is rated from 0 to 100 based on the ability to do various tasks, with higher scores indicating better functioning.
2.2.2.2. UHDRS Total Motor Score

The UHDRS total motor score (TMS), measures a range of motor features characteristically impaired in HD in a standardized manner, including gait, tongue protrusion, oculomotor function, chorea, dystonia and postural stability. Higher scores indicate more severe motor impairment than lower scores. UHDRS raters must be certified by the EHDN UHDRS-TMS online certification (www.euro-hd.net). This requires successful rating of three sample patients, filmed during UHDRS-TMS application, within a range defined as acceptable by experts in the field (as determined by a task force of the EHDN Motor working group).

2.2.2.3. UHDRS Cognitive assessment

There is no accepted cognitive battery for the cognitive assessment of HD although most HD centers rely on the UHDRS for routine clinical practice, which incorporates the symbol digit modality test, the Stroop colour word test, and a verbal fluency test as part of a comprehensive examination (Paulsen, 2011).

The Symbol Digit Modalities Test (SDMT) is a test of visuomotor integration, involving visual scanning, tracking, and motor speed. The examinee is given 90 seconds to match symbols and digits as quickly as possible using the key (specifying which number corresponds to each symbol) which is located at the top of the page (Smith, 1968). SDMT performance declines longitudinally in both Premanifest subjects close to predicted onset, and Early HD. In TRACK-HD, the SDMT has showed differences in rates of change at both 12 and 24 months in early HD, and in those close to onset had a significantly different rate of decline compared to controls over 36 months (adjusted mean loss 4.11 points [95% CI 1.49–6.73] greater than controls; p=0.003) (Tabrizi et al., 2013a).

The Stroop Test has three conditions that require visual scanning, cognitive control and processing speed. Because the Word Reading condition (the first condition normally presented) is the most sensitive in premanifest HD, it is the only Stroop condition used in the TRACK-HD Cognitive battery.

2.2.2.4 UHDRS Behavioural assessment

The behavioural assessment measures the frequency and severity of symptoms related to affect, thought content and coping styles. There are individual subscales for mood, behavior, psychosis and obsessiveness. Higher scores indicate more severe disturbance. (Group, 1996b).
2.3 Description of key studies from which data was used in this thesis

2.3.1. TRACK-HD

TRACK-HD was a prospective observational biomarker study collecting deep phenotypic data on subjects with early HD, premanifest HD gene carriers and controls. Data was collected at four study sites world-wide: London, Paris, Leiden and Vancouver. Assessments were performed annually between 2008 and 2011 within a one month window. Site staff training and quality control were rigorous, enabling the generation of highly sensitive and specific data.

TRACK-HD has been successful in the development of a battery of clinical endpoints which can be applied in clinical trials of putative therapeutics in Huntington’s disease (Tabrizi et al., 2009a, Tabrizi et al., 2011, Tabrizi et al., 2012, Tabrizi et al., 2013a). It has also generated a large body of high quality data about how Huntington’s disease subjects differ from controls, and change over time which are improving our understanding of the disease and provide avenues for further study such as in this thesis.

There were 366 subjects at baseline: 123 controls, 120 premanifest HD gene carriers and 123 Early HD subjects, of these 298 completed 36-month follow-up. Subjects with missing values or early drop-outs still contributed to the study if they had at least two study visits. Subjects had approximately 7 hours of assessments during one day annually, which included 3T MRI, quantitative motor, cognitive, oculomotor, neuropsychiatric, wet biomarker and genetic studies (Figure 2.1). While I did not collect data for the TRACK-HD study I was an investigator on the TrackOn-HD study, which followed similar protocols and is described below.
**Figure 2.1**: Study outline of TRACK-HD. Study sites, numbers of subjects in each disease group at baseline, principle assessment modalities and years assessed are shown.

2.3.2 **TrackOn-HD**

While TRACK-HD had successfully identified clinical endpoints for drug trials, there was ongoing interest in changes in Premanifest Huntington’s disease gene expansion mutation carriers, and the changes that occur in someone with Huntington’s disease around the time of diagnosis. To further explore these, and increase the longitudinal data available, TRACK-HD participants who were still pre-symptomatic at the end of the TRACK-HD project, and all TRACK-HD controls, were invited to participate in TrackOn-HD. Further subjects were recruited at each study site so that there were 30 premanifest HD subjects and 30 controls at each of the four study sites (London, Paris, Vancouver, Leiden) at the start of TrackOn-HD in 2012. I joined the London TrackOn-HD study team in late 2011 as the London site Clinical Fellow, my initial role within the study being to recruit subjects both from the existing TRACK-HD cohort, and new Premanifest subjects and controls. At the London site, new subjects were recruited from the multidisciplinary Huntington’s disease clinic at the National Hospital for Neurology and Neurosurgery.

I was responsible for the clinical evaluation and biosample processing at the London site for TrackOn-HD. Clinical evaluation of all subjects occurred at the start of the study visit. I checked that subjects were eligible for the study and gained written consent from all subjects. A medical history was performed, checking all previous and current medical problems and medications, and a detailed family history was taken. A Huntington’s disease Clinical Characteristics Questionnaire was completed. The UHDRS Motor assessment (Group, 1996b) was completed, along with the UHDRS functional assessment and the Shoulson and Fahn Total Functional Capacity score (Shoulson and Fahn, 1979). Blood samples and a buccal swab were collected. An outline of the study day is given in Table 2.2.

<table>
<thead>
<tr>
<th>Time</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00</td>
<td>Consent, Clinical Rating, Neuropsychiatric assessment</td>
</tr>
<tr>
<td>(45 min)</td>
<td></td>
</tr>
<tr>
<td>9:45</td>
<td>Imaging</td>
</tr>
<tr>
<td>(120-150 min)</td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>13:30</td>
<td>Cognitive</td>
</tr>
<tr>
<td>(60 min)</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Activity</td>
</tr>
<tr>
<td>--------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>14:30</td>
<td>Quantitative Motor</td>
</tr>
<tr>
<td>(70 min)</td>
<td></td>
</tr>
<tr>
<td>15:40</td>
<td>Oculomotor</td>
</tr>
<tr>
<td>(15 min)</td>
<td></td>
</tr>
<tr>
<td>15:55</td>
<td>Neuropsychiatric &amp; Functional assessments</td>
</tr>
<tr>
<td>(10 min)</td>
<td></td>
</tr>
<tr>
<td>16:10</td>
<td>Transcranial Magnetic Stimulation</td>
</tr>
<tr>
<td>(75 min)</td>
<td></td>
</tr>
</tbody>
</table>

*Table 2.2: Outline of TrackOn-HD assessment day.*

### 2.3.3 EHDN Registry Study

The EHDN is a non-profit research network committed to advancing research, facilitating the conduct of clinical trials, and improving clinical care in HD. Through the EHDN a platform has been created such that basic scientists, clinicians, patients and families can collaborate on academic and industry studies to fulfil its mission (EHDN, 2018).

The EHDN REGISTRY study (Orth et al., 2010) was a multisite prospective observational study which collected phenotypic data between 2003 – 2013 on over 13,000 subjects, mostly manifest HD gene carriers but also some controls. The data are less detailed, and follow up less complete than in TRACK-HD. The aim was for annual assessments +/- 3 months, though this was variable, and many subjects did not have annually collected data. The core data include: age, CAG repeat length, UHDRS Total Motor Score (TMS) and Total Functional Capacity (TFC); some patients have further assessments such as a cognitive battery (Orth et al., 2010). I recruited people to and performed assessments on subjects as a part of the EHDN Registry study.

### 2.3.4 Neuromics

During the course of my PhD I was involved, from the opening meeting in January 2014 to the closing meeting in May 2017, with Neuromics. Neuromics was a European Commission 7th Framework Programme funded project set up with the aim to revolutionize diagnostics and develop new treatments for ten major rare neuromuscular and neurodegenerative diseases. It brought together leading research groups in Europe, five highly innovative small and medium sized enterprises (SMEs), and overseas experts; using the most sophisticated Omics technologies to revolutionize diagnostics and to develop pathomechanism-based treatments for ten major neurodegenerative and neuromuscular diseases. Specifically the aims were to:
(i) use next generation WES to increase the number of known gene loci for the most heterogeneous disease groups from about 50% to 80%,
(ii) increase patient cohorts by large scale genotyping by enriched gene variant panels and NGS of so far unclassified patients and subsequent phenotyping,
(iii) develop biomarkers for clinical application with a strong emphasis on presymptomatic utility and cohort stratification,
(iv) combine -omics approaches to better understand pathophysiology and identify therapeutic targets,
(v) identify disease modifiers in disease subgroup cohorts with extreme age of onset
(vi) develop targeted therapies (to groups or personalized) using antisense oligonucleotides and histone deacetylase inhibitors, translating the consortiums expertise in clinical development from ongoing trials toward other disease groups, notably the polyglutamine repeat diseases and other neuromuscular diseases.

Much of the work in this thesis was performed to meet the objectives of the Neuromics project, including all HD Whole Exome Sequencing (Chapter 5), TrackHD SNP genotyping and GWAS (Chapter 3), work as a part of WP3 (Identification of modifying factors in cohorts enriched by deep phenotyping), and the TrackHD RNAseq (Chapter 7) as a part of WP4 (Identification of hypothesis-driven biomarkers for disease progression). I also contributed to sessions on clinical phenotyping which is discussed below.

### 2.4 Clinical Phenotyping

In addition to my contributions to the large scale HD studies described above, both my work on Polyglutamine diseases and Huntington’s disease phenocopies (Chapters 4 and 6 respectively) required detailed clinical phenotyping.

After consideration of what clinical data was pertinent to the studies and potential future studies, clinical notes were interrogated and data inputted into databases. It was important to ensure that all available notes were obtained; some subjects have multiple sets of notes. Important pieces of information, such as time of disease onset, were cross checked over multiple source documents within the notes were possible to ensure that the most accurate data was obtained.

The techniques above were adequate for the clinical phenotyping required for Chapters 4 and 6 however for many larger scale projects where multiple centres collaborate different
approaches are needed. This applies particularly when searching for the genetic causes of rare genetic conditions, one of the key objectives of the Neuromics consortium. Identifying a gene underpinning a particular condition requires resources to be pooled and phenotyping to be standardized: the deep phenotyping characterization can be seen as the counterpart to the analysis of the biomaterial samples of the respective study participants. A goal of Neuromics was to therefore develop a standardized phenotyping protocol for each of the disease groups studied. Essential clinical data was defined for each condition. The Human Phenotype Ontology (HPO) (Köhler et al., 2017) was used to map the clinical features in order to get standardized terms. The phenotyping protocols have been implemented in Phenotips (Girdea et al., 2013) and followed routinely. PhenoTips is a database which enables detailed phenotypic data to be captured, and offers opportunities for matching patients according to their disease, family background or symptoms.

2.5 Progression analysis

A key element of this thesis is the identification of genetic modifiers of Huntington’s disease progression. Progression scores were specifically developed to address this aim, initially in the TRACK-HD cohort, and then in the EHDN REGISTRY cohort with further exploratory analysis in the Leiden University Medical Centre HD cohort. While Professor Douglas Langbehn, University of Iowa performed the progression analysis, I was very involved in discussions about the approach, data usage and analysis from inception. Progression scores were derived using a combination of principal component analysis (PCA) and regression of the predictable effects of the HTT CAG repeat length in order to encapsulate the longitudinal change not accounted for by CAG and age.

2.5.1 Progression analysis for the TRACK-HD study

24 TRACK-HD variables were used in the analysis (Table 2.3). Among the wide variety of potential cognitive and quantitative-motor variables available, we analysed a subset of those that were previously used in the TRACK-HD 36-month predefined primary analysis (Tabrizi et al., 2013a). A small number of quantitative-motor variables that were substantively redundant were eliminated and those with more tractable metric properties were chosen. The 24 variables were divided a priori into 3 broad domains: (1) brain volume measures, (2) cognitive variables, and (3) quantitative-motor variables as shown in Table 2.3.

<table>
<thead>
<tr>
<th>TRACK-HD variable</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol digit modality test (number correct)</td>
<td>Cognitive</td>
</tr>
</tbody>
</table>

43
<table>
<thead>
<tr>
<th>Measure</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroop word reading (number correct)</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Paced Tapping 3 Hz (inverse standard dev)</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Spot the Change 5K</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Emotion Recognition</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Direct Circle (Log annulus length)</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Indirect Circle (Log annulus length)</td>
<td>Cognitive</td>
</tr>
<tr>
<td>Total brain volume</td>
<td>Brain imaging</td>
</tr>
<tr>
<td>Ventricular volume</td>
<td>Brain imaging</td>
</tr>
<tr>
<td>Grey matter volume</td>
<td>Brain imaging</td>
</tr>
<tr>
<td>White matter volume</td>
<td>Brain imaging</td>
</tr>
<tr>
<td>Caudate volume</td>
<td>Brain imaging</td>
</tr>
<tr>
<td>Metronome tapping, nondominant hand (log of tap initiation SD for all trials)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Metronome tapping, nondominant hand (inverse tap initiation SD for self-paced trials)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Speeded tapping, nondominant hand (log of repetition time SD)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Speeded tapping, nondominant hand (log of tap duration SD)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Speeded tapping, nondominant hand (mean intertap time)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Tongue force—heavy (log coefficient of variation)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Tongue force—light (log coefficient of variation)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Grip force, dom. hand, heavy condition (log of mean orientation)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Grip force, dom. hand, heavy condition (log of mean position)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Grip force, nondominant hand, heavy condition (log of coefficient of variation)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Grip force, dom. hand, light condition (log of coefficient of variation)</td>
<td>Quantitative motor</td>
</tr>
<tr>
<td>Grip force, nondominant hand, light condition (log of coefficient of variation)</td>
<td>Quantitative motor</td>
</tr>
</tbody>
</table>

**Table 2.3:** List of Variables to be used in TRACK-HD progression analyses. Further detail regarding these measures can be found in (Tabrizi et al., 2009a, Tabrizi et al., 2011, Tabrizi et al., 2012, Tabrizi et al., 2013a).

10 TRACK-HD subjects were excluded because they had no follow-up data. 15 further subjects were excluded because of missing brain MRI data.
For each variable the input for analysis was the subject’s random longitudinal slope from a mixed effects regression model with correlated random intercepts and slopes for each subject. The subject's random slope estimate is a "stabilized" version of the difference between observed change versus predicted change: all subjects were represented by one slope regardless of the number of visits completed, minimizing the effect of bias due to drop outs. This model regressed the observed values on clinical probability of onset statistic (CPO) derived from CAG repeat length and age, and its interaction with follow-up length. The subjects' random slope estimates thus provided a measure of atypical longitudinal change not predicted by age and CAG length.

Principal Component Analyses (PCA: see below) of the random slopes was then used to study the dimensionality of these age and CAG-length corrected longitudinal changes. Our models controlled for study site, gender, education, and their interactions with follow-up time, consistent with the models used in the TRACK-HD standard analyses which are described elsewhere (Tabrizi et al., 2009a, Tabrizi et al., 2011, Tabrizi et al., 2012, Tabrizi et al., 2013a).

2.5.1.1 Principal Component Analysis (PCA)

PCA is a technique to reduce the dimensionality of large datasets, while preserving as much statistical information (variability) as possible (Jolliffe and Cadima, 2016). This is done by finding variables that are linear functions of those in the original dataset, that successively maximize variance, and that are uncorrelated with each other (Jolliffe and Cadima, 2016). Given that PCA analysis was used to generate the progression scores which formed an important part of this thesis, I will briefly introduce the concept below, based on discussions and personal correspondence with Professor Douglas Langbehn (Langbehn, 2012).

Given a dataset of N non-redundant variables, a representation of that data can be given in N-dimensional space. A set of uncorrelated (right-angled) coordinate axes for the space can be created, and we can rotate the set of axes in an arbitrary direction. It is easiest to think of this paradigm using the intuition of 3-dimensional space, corresponding to a dataset of 3 variables. Think of rotating the x-y-z axes in a 3-D diagram without rotating the rest of the diagram. The axes can be rotated so that the variance of the data is greatest along the “x” axis. In a sense, this maximizes the average correlation of the original variables with a right-angled projection of those variables onto the axis. This axis is defined as the first principal component. It is described by the angle of rotation or equivalently by the correlation of each of the original variables with it.
Once the first axis is fixed, then we can further rotate the remaining axes so that, while remaining at right angles to each other, one of the axes again maximizes its correlation with the data, given that the first PC axis is already fixed. This next axis is the second principal component. This procedure can be repeated for subsequent components until the rotation of the entire axis system for the data-space is fixed.

We assume that the data variation projected along the first principal component is much greater than the variation along the second or subsequent components. It may be reasonable to assume that the first component may summarize the most relevant information within the data and subsequent components may reflect noise, however in other cases the higher numbered principal components may represent crucial fine detail. It should be noted that if the original variables have little correlation with each other, then little or no dimension-reduction can be gained via PCA. (Langbehn, 2012).

2.5.1.2 Assessing phenotypic clustering

In order to evaluate whether the data provided evidence for phenotypic clustering in HD we performed the analysis twice: firstly with the variables grouped a priori into 3 broad domains: (1) brain volume measures, (2) cognitive variables, and (3) quantitative-motor variables; and secondly with all variables grouped together. The results were inspected to look for evidence of phenotypic clustering.

2.5.2 Progression analysis in REGISTRY

1835 adult subjects from REGISTRY were included in this study on the basis of available genotype data (GeM-HD-Consortium, 2015). We collected the following phenotypic variables: UHDRS TMS, SDMT, verbal fluency, Stroop colour reading, word reading and interference measures, functional assessment score, and TFC.

Follow-up length and frequency was variable and missing data were substantial, making longitudinal progression analysis problematic. We therefore examined cross-sectional status at last visit, using a single unified motor-cognitive dimension of severity. In summary we performed multiple imputation to fill in missing data, derived PCA severity scores and regressed off the predictive effect of age, CAG length, and gender on the PCA severity scores derived from this data to obtain the measure of atypical severity at the last visit. This gives a single point “severity” score based on how advanced a subject is compared with expectations based on their CAG repeat and age, this score was used as the REGISTRY progression score.
In order to generate atypical severity scores, three sequential procedures were required: (i) Multiple imputation of missing data (ii) Principal Component Analysis (PCA) and severity scoring of the combined imputed data replications (iii) Regression of the predictive effect of age, CAG length, and gender on the PCA-derived severity scores so that we are left with a measure of atypical (or “unexplained”) severity. The steps were taken in the order above; given that these steps could be done in different orders we also confirmed that there were only minimal differences due to order. This analysis was performed by Professor Douglas Langbehn after discussions to which I contributed about how best to approach the analysis.

We looked only at 1835 subjects who had available genotypic data through the GeM consortium. Given that the GeM study focused initially on examining genetic modifiers of motor onset in HD, the majority of these participants in REGISTRY had manifest Huntington’s disease. 1773 subjects had adequate phenotypic data for progression scoring. We used a square-root transform of TMS to improve approximate multivariate normality of the data.

To deal with the missing data for clinical items, multiple imputation with 25 imputations was performed. Age, gender, and CAG expansion length were auxiliary variables for the imputations. Final parameter estimates and statistical significance were estimated by Rubin’s method (Rubin, 2008). We performed the above using the MI and MIANALYZE procedures of SAS/STAT 13.1 (Inc., 2013). We noted some evidence of study site effects in the eventual regressions. Thus we used a random effect for site in models adjusting for age and CAG.

Atypical severity was defined as the residual between each subject’s observed and marginal predicted value. The final averaged multiple imputation model used a 2 degree of freedom restricted cubic spline (Harrell, 2001) of cumulative probability of onset (CPO), plus main effects of gender and CAG length and a random effect for site. Marginal effects from this model, which represent the estimated effects after accounting for site fluctuations, were used for all predictions. The knot placement for the clinical probability of onset spline was defined a priori using a conventional standard at the 10th, 50th, and 90th percentiles of its observed distribution. The corresponding values were (0.131, 0.395, 0.885). Atypical severity was defined as the residual between each subject’s observed and marginal predicted value. Final parameter estimates, along with estimates of statistical significance adjusted for the multiple imputation procedure are shown in the Table 2.4.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>gender</th>
<th>Estimate</th>
<th>Std Error</th>
<th>95% Confidence Limits</th>
<th>DF</th>
<th>t for H0:</th>
<th>P Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td></td>
<td>2.075589</td>
<td>0.267283</td>
<td>1.55102</td>
<td>2.60016</td>
<td>897.01</td>
<td>7.77</td>
</tr>
<tr>
<td>cpo_1</td>
<td></td>
<td>-0.9142</td>
<td>0.21009</td>
<td>-1.32638</td>
<td>-0.50201</td>
<td>1191.6</td>
<td>-4.35</td>
</tr>
<tr>
<td>cpo_2</td>
<td></td>
<td>-7.00283</td>
<td>0.911001</td>
<td>-8.79025</td>
<td>-5.2154</td>
<td>1141.5</td>
<td>-7.69</td>
</tr>
<tr>
<td>cag</td>
<td></td>
<td>-0.01919</td>
<td>0.005133</td>
<td>-0.02927</td>
<td>-0.00912</td>
<td>862.96</td>
<td>-3.74</td>
</tr>
<tr>
<td>gender F</td>
<td></td>
<td>-0.13631</td>
<td>0.042605</td>
<td>-0.21992</td>
<td>-0.05271</td>
<td>1030.1</td>
<td>-3.2</td>
</tr>
<tr>
<td>gender M</td>
<td></td>
<td>0</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Table 2.4: Parameter estimates of variables in the model used to generate the REGISTRY cross-sectional severity score.

Multiple imputation adjusted estimates of statistical significance are given. CPO_1: clinical probability of onset; CPO_2: single transformation of clinical probability of onset. DF: degrees of freedom.

We inspected the potential biasing influence of the CAG repeats, by classifying the individual in short (CAG < 41) and long (CAG > 55) repeats. We found an overrepresentation of people with larger atypical severity scores among those with short CAG, which implies that those with a small number of repeats are more likely to be in the study if atypically severely affected. This is likely to be due to the disease only being partially penetrant in those with short CAG repeats, resulting in bias (Langbehn et al., 2004). This prompted us to exclude subjects with short CAG from the creation of the severity scores, while retaining those with long CAG repeats. However, we confirmed that the age-CAG severity function predicted using CAG > 41 gave sensible estimates for both the short and long ranges, enabling even those subjects with short CAG repeats to be used in the final analysis (Figure 2.2).
Figure 2.2: Age-CAG severity function against clinical probability of onset (CPO) in REGISTRY. A: plot showing predicted values for all subjects. B: plot of predicted values using only subjects in the CAG 41–55 range. C: Plot based on extrapolating the severity model to subjects with CAG in the 36-40 range (the appearance of two rather distinct lines are due to the gender effect, with women having lower predicted scores than men).

2.5.3 Progression analysis in Leiden University Medical Centre (LUMC) samples

Though a collaboration with Willeke van Roon-Mom via Neuromics we had access to a cohort of HD and control subjects from LUMC. The primary objective of the cohort was to investigate neuropsychiatric aspects of HD, but the samples from these subjects have been also extensively investigated by van Roon-Mom and her team. The subset of the cohort (Mastrokolias et al., 2015) that we used for our RNAseq work (Chapter 7) consisted of 18 premanifest gene carriers, 56 manifest HD subjects and 27 age and gender-matched controls. Motor onset was determined by an experienced neurologist using the same UHDRS standard as in TRACK-HD. All premanifest carriers showed no substantial motor signs, with a TMS of 5 or less and a UHDRS diagnostic confidence level less than 4. All controls were free of known medical conditions.

The phenotypic data available for the LUMC samples were: UHDRS TMS (total motor score), total functional capacity (TFC) alongside neuropsychiatric variables, age and CAG repeat size. Note that because the motor score has a floor at 0 (no motor symptoms = score of 0) and the TFC has a ceiling at 13 (functionally normal = score of 13), our ability to look at premanifest
HD is limited. Given that previous investigation in the TRACK-HD cohort led us to exclude neuropsychiatric variables in our progression analysis we did not use them here. Where available, longitudinal data had an interval of roughly 3 years.

We first considered a longitudinal analysis as data obtained at a roughly 3 year interval was available for some subjects. However there was little correlation between the TFC and TMS residual changes. We instead opted to look at cross sectional severity scores in a similar approach to that used for the REGISTRY progression analysis described above. To do this we tested a variety of models for predicting the severity component, based on various combinations of CAG length, age gender, interactions and nonlinear functions. Results were robust to the particular choice of model. We therefore selected a method similar to the REGISTRY last-visit cross-sectional model. The main difference from the REGISTRY method is that subjects' values from both visits were used, whereas only the last visit was in REGISTRY. The concern in REGISTRY was that visits tended to be unevenly spaced and scheduled for unclear, possibly inconsistent reasons. In contrast, most Leiden subjects had a baseline and a planned 3 year follow-up. 78 subjects had adequate data to generate the LUMC atypical severity score.

The severity factor was based on a principal component analysis of only two variables, total functional capacity (TFC) and (square root) motor score. After standardizing each variable to mean of 0 and standard deviation = 1, both of them receive equal weighting in calculating this score since equal weighting is inherent when only 2 variables are used for a PCA. The principal component has a correlation of .949 with both the TFC and square root motor score.

2.6 Assessment of Relatedness

Family history data was collected as a part of the TRACK-HD clinical evaluation. I obtained these data in the form of family history diagrams, and relationship descriptions. To enable further analysis I converted the data of 38 family histories in which there was more than one family member in the study into standard family history formatting (Table 2.5).
<table>
<thead>
<tr>
<th>Column number</th>
<th>Column entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Individual’s family ID ('FID')</td>
</tr>
<tr>
<td>2</td>
<td>Individual’s within-family ID ('IID'; cannot be '0')</td>
</tr>
<tr>
<td>3</td>
<td>Within-family ID of father ('0' if father isn't in dataset)</td>
</tr>
<tr>
<td>4</td>
<td>Within-family ID of mother ('0' if mother isn't in dataset)</td>
</tr>
<tr>
<td>5</td>
<td>Sex code ('1' = male, '2' = female, '0' = unknown)</td>
</tr>
</tbody>
</table>

*Table 2.5: Format for family history encoding*

A simple example is shown in Figure 2.3 below.

![Family History Diagram and Table](image)

*Figure 2.3: Family history encoding. A: family history diagram, and B: encoded format of family history data for family 1, comprising a mother and daughter pair. Grandchildren not coded as not required for analysis in this case.*

Of those with family members in TRACK-HD, 28 individuals had at least one family member also included in the genome-wide association analysis.

### 2.7 General genetics methods

#### 2.7.1 Genotyping

Genotyping is the measurement of genetic variation. Historically, in order to do genetic mapping it was necessary to develop techniques for genotyping. The first type of DNA marker to be studied were restriction fragment length polymorphisms (RFLPs). Restriction fragments are produced when a DNA molecule is treated with a restriction endonuclease that cut the
DNA at a defined point, eg EcoR1. The RFLPs can then be detected using Southern hybridisation or PCR. Southern blotting is discussed in the Materials and Methods section of Chapter 6.

SNP genotyping is the measurement of genetic variations of single nucleotide polymorphisms (SNPs) between samples. SNPs are biallelic and occur approximately every 1,000 base pairs (bp) throughout the human genome. SNPs can be readily genotyped using techniques that discriminate any two-way combination of adenine, guanine, cytosine, and thymine nucleotide bases. SNP detection is based on oligonucleotide hybridization analysis: the synthetically generated oligonucleotide binds to DNA if it is complementary to the DNA, if there is a mismatch (alternative version of the SNP) it does not. There are many different techniques that can be employed to detect the SNP genotype of a subject, the choice often being guided by the number of samples and number of SNPs to genotype. To genotype a handful of SNPs techniques include microtitre plate based techniques such as Taqman, whereas for large numbers of samples high throughput arrays tend to be used. These genotyping techniques may be universal, based on standard SNPs, or customized to the particular SNPs of interest to the researcher.

Genotyping by allelic discrimination using the 5’ nuclease (TaqMan®) assay in conjunction with Minor Groove Binding probes was used to genotype samples for rs3849942 in Chapter 6. In this technique, a wild-type SNP Allele “A” is amplified separately from the alternative Allele “B” using region specific forward and reverse primers and two allele-specific TaqMan® probes designed to target the polymorphism (Malkki and Petersdorf, 2012). The amplification is performed using a thermal cycler or a real-time PCR system and fluorescent signals are interpreted automatically using sequence detection software dedicated to real-time PCR instrumentation (Malkki and Petersdorf, 2012).

Custom KASP assays were used for the genotyping of DNA repair gene variants for Chapter 4 of this thesis, enabling a set of specific SNPs to be examined. In KASP, the SNP-specific KASP Assay mix and the universal KASP Master mix are added to DNA samples, a thermal cycling reaction is then performed, followed by an end-point fluorescent read (LGC, 2018). The KASP Assay mix contains three assay-specific non-labelled oligonucleotides: two allele-specific forward primers and one common reverse primer. The allele-specific primers each harbour a unique tail sequence that corresponds with a universal FRET (fluorescence resonant energy transfer) cassette, when not quenched the cassette emits fluorescence. Bi-allelic
discrimination is achieved through the competitive binding of the two allele-specific forward primers.

For the GWAS (Chapter 3) samples were instead genotyped on a chip array: the Illumina Omni 2.5-8 v1.1 array. This is an example of a high throughput SNP genotyping technology, in which the genotyping is multiplexed, enabling many more SNPs to be genotyped simultaneously on a bead array platform both accurately and cost effectively, thus transforming what is possible in genetic studies. Illumina bead array microarray technology is based on 3-micron silica beads that self-assemble in microwells. Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that acts as the capture sequence, in this case the arrays have around 2.5 million markers, chosen to provide a comprehensive set of both common and rare SNP content from the 1000 Genomes Project (MAF>2.5%) for diverse world populations (Illumina, 2017). The beads are randomly deposited into the wells on a substrate, and the array must be decoded to determine which oligonucleotide-bead combination is in which well. This decoding is done using the address segment of the oligonucleotide, and, involves sequential hybridization of differentially labelled probes (OHSU, 2017). The differential labelling uses three states – carboxyfluorescein (FAM) labelled green, cyanine 3 (Cy3) labelled red, and not labelled. During any given cycle of the process, a bead is green, red, or blank. Labelled oligonucleotides are hybridized to the arrays at high concentrations which allows for rapid hybridizations, followed by washing to remove non-specific signal and background. Each round of hybridization adds another digit to the number (Gunderson et al., 2004), until there are sufficient digits to uniquely identify each probe (OHSU, 2017).

2.7.2 Genotyping of polymorphic repeats using fragment analysis

Repetitive regions of DNA are challenging to genotype and are still not covered by standard technologies such as those described above. Investigation of the hexanucleotide repeat C9orf72 for this thesis was done using repeat-primed PCR (More detail is given in Chapter 6). This involves a forward primer unique to a sequence near the repeat and a reverse primer composed of several repeat units which can bind anywhere in the repeat region, thus creating amplicons of varying sizes. The reverse primer is used in lower concentrations and is exhausted in a few cycles, after which an anchor primer takes over as the reverse end starting point. Fragment analysis is then performed, the presence of a characteristic stutter pattern indicating the presence of an expansion at the locus of interest.

2.7.3 Sanger Sequencing
Sanger sequencing is the process of selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. While these days next generation sequencing (NGS) is more widely used Sanger sequencing remains valuable to confirm variants identified by NGS. In Chapter 5 I used Sanger sequencing to investigate a genetic locus highlighted by next generation technologies.

Classical Sanger sequencing requires a single-stranded DNA template, a DNA polymerase, a DNA primer, normal deoxynucleosidetriphosphates (dNTPs), and modified nucleotides (ddNTPs) that terminate DNA strand elongation. For the *MSH3* sequencing described in Chapter 5 Primer3 (Untergasser et al.) was used to generate the primers. The ddNTPs lack a 3'-OH group that is required for the formation of a phosphodiester bond between two nucleotides, causing the extension of the DNA strand to stop when a ddNTP is added. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard dNTPs, the DNA polymerase, and only one of the four ddNTPs for each reaction. After rounds of template DNA extension, the DNA fragments that are formed are denatured and separated by size using gel electrophoresis with each of the four reactions in one of four separated lanes. While gels or X-ray film may be used to read the sequence, in the method I used ddNTPs which were fluorescently labelled for detection in an ABI automated sequencing machine, and the output was analysed using Sequence Scanner 2 software. Output sequences were then aligned to the reference genome using BLAST.

### 2.7.4 Next generation sequencing (NGS)

Next-generation sequencing refers to non-Sanger-based high-throughput DNA sequencing technologies. Millions or billions of DNA strands can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods.

There are several different NGS technologies including:

- Illumina sequencing
- Roche 454 sequencing
- Ion torrent: Proton / PGM sequencing
- SOLiD sequencing

These vary on the read lengths and chemistry used.

Illumina Nextera library pooling method was used to perform the Whole Exome Sequencing in this thesis (Chapter 5). This technology uses the following steps (Illumina, 2018b):

1. **Fragmentation**- DNA is simultaneously tagged and fragmented by a transposome
2. **Tagmented DNA is amplified and sequencing indexes are added by PCR
3. **Library pooling of up to 12 libraries (enabling high throughput)**
4. Biotin-labeled probes specific to the targeted regions are used for two rounds of hybridization. The pool is enriched for the desired regions using streptavidin beads that bind to the biotinylated probes. Biotinylated DNA fragments bound to the streptavidin beads are magnetically pulled down from the solution.

5. Second amplification with PCR

6. Amplified libraries are cleaned up: fragments are eluted from the beads

7. Sequencing. The whole exome sequencing was performed on the Illumina HiSeq 2000 (Illumina, 2018a) which is based on a proprietary reversible terminator-based method that detects single bases as they are incorporated into the growing DNA strands.

2.7.5 Expression analysis

The expression analysis in this thesis was done using RNAseq, which employs next generation sequencing technology (described above) to reveal the presence and quantity of RNA in a biological sample at a given moment. This enables differential expression between different individuals / samples / conditions to be explored, along with novel isoforms and splice variants to be identified.

The detailed methods of the RNAseq in this thesis are described in Chapter 7 in which RNAseq is discussed.

2.7.6 Association testing

Association analysis is the statistical method in which the association between genotype and phenotype is examined. The classical approach to hypothesis testing developed by Neyman and Pearson (Neyman and Pearson, 1933) involves setting up a null hypothesis (\(H_0\)) and an alternative hypothesis (\(H_1\)), calculating a test statistic (\(T\)) from the observed data and then deciding on the basis of \(T\) whether to reject \(H_0\). In genetic studies, \(H_0\) typically refers to an effect size of zero, whereas \(H_1\) usually refers to a non-zero effect size (for a two-sided test) (Sham and Purcell, 2014). If the study were to be repeated many times, each drawing a different random sample from the population, then a set of many different values for \(T\) would be obtained, which can be summarized as a frequency or probability distribution. The \(P\) value, which was introduced earlier by Fisher (Fisher, 1925) in the context of significance testing, is defined as the probability of obtaining — among the values of \(T\) generated when \(H_0\) is true — a value that is at least as extreme as that of the actual sample (denoted as \(t\)) (Sham and Purcell, 2014).

2.7.7 Genome wide association analysis
In a genome wide association analysis (GWAS), association analysis is performed, with a separate statistical test being performed at each locus to examine whether the locus is associated with the variable being tested.

The GWAS approach relies on the foundation of data produced by the International Human HapMap Project and the identification of millions of single nucleotide polymorphisms (SNPs) in the human genome, and the fact that due to linkage, genetic variance at one locus can predict with high probability genetic variance at an adjacent locus (Hardy and Singleton, 2009, Gandhi and Wood, 2010).

Some of the earliest genome wide association analyses involved testing for linkage at loci spanning a large portion of the genome, but over time more and more markers have been included. As many SNPs are being tested, keeping the significance threshold at the conventional value of 0.05 would lead to a large number of false-positive significant results; to avoid this, the threshold for significance in linkage analysis was typically chosen so that the probability of any single false positive among all loci tested is ≤0.05. Simulation studies using data on HapMap Encyclopedia of DNA Elements (ENCODE) regions to emulate an infinitely dense map gave a genome-wide significance threshold of $5 \times 10^{-8}$. Other thresholds have been suggested however: by subsampling genotypes at increasing density and extrapolating to infinite density, a genome-wide significance threshold of $7.2 \times 10^{-8}$ was obtained; sequence simulation under various demographic and evolutionary models found a genome-wide significance threshold of $3.1 \times 10^{-8}$ for a sample of 5,000 cases and 5,000 controls, in which all SNPs were selected with minor allele frequency of at least 5%, for a European population; a detailed study of the Icelandic population suggested that sequence variants should be weighted base on their annotation, and that variable significance thresholds should be used based on the annotation (Sveinbjornsson et al., 2016); others propose that q values which are similar to the p value, except that it is a measure of significance in terms of the false discovery rate rather than the false positive rate (Storey and Tibshirani, 2003) should be used for significance testing.

### 2.7.8 Gene-set and pathway analysis

The association analysis described above is based on single genetic markers, by contrast pathway analysis aggregates signal from a set of markers. It was first developed for the analysis of transcriptome data, and then transferred to the analysis of GWAS data (Holmans, 2010). The motivation being that it was noted that genetic variants that confer small disease risks are likely to be missed in the most-significant SNPs/genes approach after adjustment for
multiple testing, and even those variants that confer a larger effect might not always rank among the top 20–50 among hundreds of thousands of markers tested (Wang et al., 2007). In some cases the association signal is spread out over a gene or biological pathway, thus methods to aggregate the association signal over a gene or set of genes which form a biological pathway can prove a valuable addition to single variant based methods (Wang et al., 2007).

In pathway analysis, a set of genes (the “pathway”) is tested for enrichment of association signal with a trait (Holmans, 2010, Mooney and Wilmot, 2015). There are two types of pathway analysis, depending on the null hypothesis being tested. Competitive tests compare the association between a gene-set and disease with that of all other gene sets being studied, whereas self-contained tests test whether there is significant association between the gene-set and disease (Holmans, 2010). There are various different statistical methods used for pathway analysis including -

- Overrepresentation analysis- a comparative test in which the proportion of genes in a pathway is compared with the proportion of genes not in the pathway eg DAVID, ALIGATOR (Holmans et al., 2009). The disadvantage is that the threshold to define the list of genes/SNPs needs to be set.
- Gene-set enrichment analysis- a competitive test approach which instead ranks the genes in order of significance, then tests for differences between the ranks of genes in a pathway compared to other genes eg Bioconductor (Gentleman RC, 2004), GSEA-P (Holmans, 2010, Holmans et al., 2009), Gorilla (Eden et al., 2009).
- Set-based methods- aggregate the association evidence across all genes / SNPs in a pathway into one combined test statistic, and then test whether this statistic is larger / smaller than expected under the null hypothesis
- Modelling methods- attempt to use more sophisticated models of the relationship of phenotype and genes/SNPs
- Network-based methods. In contrast to the methods above which require that the pathway is already specified, network-based methods derive pathways or gene networks from the data itself, clustering genes based on their co-expression into modules/ clusters. For example weighted co-expression network analysis (WCNA) (Zhang and Horvath, 2005, Langfelder and Horvath, 2008) which is employed in Chapter 7.

2.7.8 MAGMA analysis
MAGMA analysis (de Leeuw et al., 2015) is a recently developed technique of gene analysis which uses a multiple regression approach to incorporate linkage disequilibrium between markers and to detect multi-marker effects. The MAGMA model of gene-set analysis is divided into two separate parts: firstly the gene analysis quantifies the degree of association each gene has with the phenotype, and the correlations between genes are estimated; secondly these gene p-values and gene correlation matrix are used in the gene-set analysis (de Leeuw et al., 2015). It can be used for both aggregating signal across a gene, or a biological pathway: both strategies have been employed in this thesis (Chapter 4).
Chapter 3: Identification of genetic variants associated with Huntington’s disease progression: a genome-wide association study

3.1 Introduction

While HD is characterised by movement, cognitive and psychiatric problems, the symptoms, age of disease onset (AAO) and rate of disease progression vary from person to person (Ross and Tabrizi, 2011). It is the aim of this chapter to identify genetic factors which modulate the course of HD.

There is a strong inverse correlation between HTT CAG repeat length and age at motor onset which accounts for 50-70% of the observed variance in onset (Chapter 1) (Langbehn et al., 2004). However studies of extended Venezuelan kindreds suggested that there was residual heritability accounting for this difference in onset age, even after accounting for CAG repeat (Wexler et al., 2004b), suggesting that other genetic factors may modulate onset. This was confirmed by a recent genome wide association study from the Genetic Modifiers of Huntington’s Disease (GeM-HD) Consortium which identified genes in the DNA damage response as likely to modify the onset of HD (GeM-HD-Consortium, 2015).

Extensive investigation has shown that the presence of the HD CAG repeat expansion within huntingtin perturbs the cellular and physiological system in numerous ways (Bates et al., 2015, Ross et al., 2014), however it is not established which of these events are critical in humans to making the patient unwell, and which can be regarded as epiphenomena. Identifying genetic modifiers of progression in HD is likely to illuminate key events, since such genetic modifiers by definition are sufficiently pivotal to alter the manifestation of disease. And importantly, the genetic variants provide proof of concept that biological factors can be manipulated in people to result in a change of the disease trajectory. Thus they may be good drug targets (Plenge et al., 2013).

AAO (Huntington’s et al., 1993, Hogarth et al., 2005) reflects the trajectory of disease pathology up to the point of motor onset: onset of disease is preceded by a long prodromal phase accompanied by substantial brain cell death. However, as discussed in Chapter 1 the transition from premanifest to manifest HD is gradual and fluctuant rather than abrupt (Long et al., 2013, Tabrizi et al., 2013a): for example subtle early chorea may be more apparent if
the patient is anxious than relaxed. In the prodromal phase, some clinicians will first introduce the patient gently to the idea of them having HD during several consultations rather than formally diagnosing HD the first time chorea is noticed. All these factors make HD onset challenging to define, particularly retrospectively from case notes. In addition to its likely inaccuracy, onset is only clinically confirmed in those with unequivocal motor signs. This is likely to cause problems in treatment trials in subjects close to, or before, clinical onset of disease, which will be necessary if the course of neurodegeneration is to be slowed or halted in HD.

The need for robust biomarkers of disease progression in both manifest and premanifest HD has motivated a raft of observational studies (Tabrizi et al., 2013a, Orth et al., 2010, Paulsen et al., 2008). These provide the opportunity to investigate the relationship between onset and progression, whether they are influenced by the same biology, and also permit the study of subjects before clinical onset.

In comparison to the well-established relationship between HTT CAG repeat size and AAO, the relationship between HTT CAG repeat size and progression is less clear. In an analysis of 335 subjects, significant associations between CAG repeat length and worsening on several motor, cognitive, and functional outcomes were found, however when age was controlled for, these effects were not significant (Ravina et al., 2008). However in a later study of 569 subjects and longer follow-up times, CAG repeat length showed a significant but small effect on the progression of clinical measures (motor, cognitive and functional), and when age was controlled for the correlation increased (Rosenblatt et al., 2012). Intriguingly it was recently demonstrated that mutant HTT CAG repeat size is strongly associated with both age at onset and age at death in patients with HD, but not with disease duration defined as the difference between the ages at onset and death (Keum et al., 2016). A recent analysis of 5,821 Enroll patients (Aziz et al., 2018, Landwehrmeyer et al., 2016) found that around two-thirds of the rate of functional, motor, and cognitive progression in HD is determined by the same factors that also determine AAO, and that CAG repeat size alone could account for about half of the variation in the rate of deterioration in these domains. Their data suggest that factors that are represented by the age at onset influence progression through their interaction with HTT CAG repeat dependent mechanisms (Aziz et al., 2018). By contrast HTT CAG repeat accounted for only a minimal effect on weight loss, while residual onset had no effect (Aziz et al., 2018), leading the authors to postulate that weight loss and the pathological process which drive it may be linked to age of death in HD in an effort to reconcile their data with Keum et al (Aziz et al., 2018, Keum et al., 2016).
While the most clearly distinct phenotypic subtype of Huntington’s disease is juvenile onset Huntington’s disease (Quarrell, 2014) there has also been discussion in the field about the possibility of subtypes within the more typical adult-onset disease (Roos, 2014, Kim et al., 2015). Furthermore, some have attempted to identify genetic modifiers of specific disease subtypes in small candidate gene studies (Vinther-Jensen et al., 2016). It was therefore important for us to establish whether we should be looking for genetic modifiers predisposing towards a particular subtype of disease, or whether we should be looking for genetic modifiers of Huntington’s disease overall. We therefore looked for phenotypic clustering when performing the progression analysis- questioning whether there is evidence for motor dominant vs cognitive dominant HD for example.

TRACK-HD represents the most deeply phenotyped cohort of premanifest and symptomatic Huntington’s disease, with annual visits involving clinical, cognitive and motor testing alongside detailed brain imaging (Tabrizi et al., 2009a, Tabrizi et al., 2013a). In this chapter the detailed data in TRACK-HD (Tabrizi et al., 2009a, Tabrizi et al., 2013a) is explored to establish whether distinct subphenotypes of Huntington’s disease exist. A novel unified Huntington’s disease progression measure was developed (detailed in Chapter 2) and used to explore the relationship between HD progression and onset. We examined whether we could detect any genetic association in TRACK-HD using the unified HD progression measure as a quantitative trait in the analysis. We developed a similar measure in subjects from the REGISTRY study to replicate our findings (Orth et al., 2010).

3.2 Materials and Methods

3.2.1 Study design

The overall study design is illustrated in Figure 3.1. Most of the material discussed here was recently published in Hensman Moss, Pardiñas et al (Hensman Moss et al., 2017b); the work presented was part of a collaborative project and I will indicate my involvement with different aspects in the text below.

We first performed progression analysis in TRACK-HD: this analysis is detailed in the General Methods (Chapter 2.5.1). We next performed a GWAS in TRACK-HD using cross domain progression as the analytical variable (Basic principles of association testing and Genome Wide Association Analysis are introduced in Chapter 2.7.6 and 2.7.7). To validate these findings and investigate further we performed progression analysis in REGISTRY (Chapter
2.5.2) then performed a GWAS in REGISTRY using the REGISTRY progression score as the analytical variable. We meta-analysed the results of the TRACK-HD and REGISTRY association analyses. In addition to single variant based analysis we performed pathway based approaches.

![Study Design Diagram]

**Figure 3.1: Study Design.** After establishing that brain imaging, quantitative motor and cognitive variables are correlated and follow a similar trajectory, we scored the TRACK-HD subjects using principal component 1 as a unified progression measure, and used this measure to look for genome-wide associations with HD progression. We replicated our findings in the EHDN Registry subjects by looking at how far their disease had progressed compared with
expectations based on CAG/age, and used this progression measure to look for genome-wide associations in REGISTRY. 1835 Registry subjects had genotype data (8). UHDRS TMS: Unified Huntington’s Disease Rating Scale Total Motor Score. SDMT: symbol digit modality test. TFC: Total Functional Capacity. (Figure made by me).

3.2.2 Standard Protocol Approvals, Registrations, and Patient Consents

All experiments were performed in accordance with the Declaration of Helsinki and approved by the University College London (UCL)/UCL Hospitals Joint Research Ethics Committee; ethical approval for the REGISTRY analysis is outlined in (Consortium, 2015a). Peripheral blood samples were donated by genetically-confirmed HD gene carriers, and all subjects provided informed written consent.

3.2.3 Case ascertainment

Subjects for this chapter came from two studies: TRACK-HD and REGISTRY which are more extensively described in the Methods (Chapter 2).

TRACK-HD was a prospective observational biomarker study collecting deep phenotypic data including imaging, quantitative motor and cognitive assessments on adult subjects. It provided annually collected high quality longitudinal prospective multivariate data over three years (2008-2011) with 243 adult subjects at baseline: 123 early HD, 120 premanifest HD gene carriers and 123 controls (Tabrizi et al., 2009a, Tabrizi et al., 2013a, Tabrizi et al., 2012, Tabrizi et al., 2011). 218 Huntington’s gene carriers from TRACK-HD were included in this study on the basis of adequate longitudinal data. I was clinical fellow for the TrackOn-HD study, a three year extension of the TRACK-HD study focusing on pre- and peri-manifest HD subjects.

REGISTRY (Orth et al., 2010) was a multisite prospective observational study which collected phenotypic data between 2003 – 2013 on over 13,000 subjects, mostly manifest HD gene carriers. The core data include: age, CAG repeat length, UHDRS Total Motor Score (TMS) and Total Functional Capacity (TFC); some patients have further assessments such as a cognitive battery (Orth et al., 2010). 1835 adult subjects from REGISTRY were included in this study on the basis of available genotype data from the GeM GWAS (Consortium, 2015a). I enrolled people for and did study visits for the REGISTRY study, though the data used here was obtained from a large data-cut on subjects also used for the GeM GWAS study.

3.2.4 Relationship between progression scores used in TRACK-HD and REGISTRY
Using the methods described in Chapter 2 (Methods) we generated scores of cross-domain progression using data from both the TRACK-HD and REGISTRY studies. While the principal component analysis was performed by Prof Douglas Langbehn I was involved in extensive discussions to decide which approach to use for the progression analysis for both TRACK-HD and REGISTRY. To ensure that the unified TRACK-HD progression measure and the unified REGISTRY progression measure encapsulated similar clinically relevant information we explored the relationship between them.

Four measures were common between the TRACK-HD and REGISTRY studies: TMS, symbol digit modality score, Stroop word reading score and TFC. Using these we were able to construct a progression score using the REGISTRY cross sectional scoring method with the TRACK-HD dataset (the TRACK-HD severity score), and compare this with the score generated by the TRACK-HD longitudinal progression analysis method.

We conducted a principal component analysis of the four shared measures at the last TRACK-HD visit: first principal component accounted for 79.4% of the variance in the PCA and correlated approximately equally with each of the four observed variables (Table 3.1).

<table>
<thead>
<tr>
<th>Factor Pattern</th>
<th>Factor1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Square root of raw UHDRS total motor score</td>
<td>-0.91567</td>
</tr>
<tr>
<td>Symbol digit modality test (number correct)</td>
<td>0.90797</td>
</tr>
<tr>
<td>Stroop word reading test (number correct)</td>
<td>0.87904</td>
</tr>
<tr>
<td>UHDRS Total functional capacity</td>
<td>0.86045</td>
</tr>
</tbody>
</table>

Table 3.1: Proportion of variance among variables present in TRACK-HD and REGISTRY which are accounted for by the first PC in the combined analysis.

To calculate the measure of severity unaccounted for by age and CAG length in TRACK, we regressed these principal component scores on the same predictors used for the unified REGISTRY progression measure. The residuals served as the TRACK-HD severity scores.

3.2.5 Relationship between progression scores and other clinical measures
UHDRS TMS and TFC were not included in the TRACK-HD progression analysis. To confirm that the TRACK-HD progression measure correlated with these, we examined the residual change relationships between the progression score and UHDRS TMS change and TFC change after controlling for the clinical probability of onset (CPO).

3.2.6 Genotyping

DNA was obtained from blood samples of the 218 TRACK-HD study participants who had complete serial phenotype data. Blood was drawn using an aseptic technique from the antecubital fossa; blood for DNA extraction was collected in ACD tubes and shipped on the day of collection at ambient temperature to BioRep, Milan, Italy for processing. (While I did not collect blood samples for TRACK-HD I used the same technique when collecting samples for TrackOn-HD). BioRep carried a manual salting out method for DNA extraction: this makes use of high salt conditions to selectively precipitate out proteins, leaving DNA in solution to be subsequently precipitated with alcohol. The purified DNA was then stored in TE buffer (10mM Tris, 1mM Na2EDTA, pH8). Spectrophotometric analysis (Nanodrop), agarose gel electrophoresis and genotyping for sample identity confirmation (PCR + capillary electrophoresis) were done to assure quality control.

I obtained TRACK-HD DNA samples from BioRep and prepared them for genotyping. Genotyping was performed in Illumina Omni2.5-8 v1.1 arrays at UCL Genomics, in accordance with the Infinium LCG Assay (15023141_A, June 2010) protocol (Illumina Inc, San Diego, USA), details of this technique are given in Chapter 2.

I sent the genetic data to Antonio Pardiñas (Cardiff University) who carried out standard quality control procedures (Anderson et al., 2010) by using PLINK v1.9 (Chang et al., 2015), including controlling for:

- Coverage and call rates: 5% of missing data allowed per SNP and individual
- Inbreeding (F < 0.2 required)
- Hardy-Weinberg equilibrium (SNPs with p < 10^-6 in an exact test were removed).

With these criteria, and after removing one individual of an identical twin pair, a total of 216 gene positive TRACK-HD subjects were left in the sample who were genotyped for 2.34 million genome-wide markers (Figure 3.1).

3.2.7 Relatedness and Population genetic analysis

Of those with family members in TRACK-HD, using the family history data (Chapter 2) I identified 28 individuals who reported at least one family member also included in the
genome-wide association analysis. Relatedness was also examined by Dr Pardiñas using the genetic data (Weir et al., 2006): identity-by-descent analysis showed 9 pairs of individuals with a relatedness coefficient ($\hat{r}$) higher than 0.15, which included 6 putative first degree relatives, 2 putative second degree relatives and 1 putative pair of third degree relatives. ADMIXTURE analysis with a subset of the 1000 Genomes (1000 Genomes Project Consortium, 2012) populations revealed 6 individuals with more than 25% of non-European ancestry (Figure 3.2).

![Figure 3.2](image)

**Figure 3.2**: Ancestry analysis of the TRACK-HD cohort (left hand box) in comparison to Finnish, Chinese, and Yoruban populations from left to right respectively. 6 subjects had >25% non-European ancestry. *Figure prepared by Dr Pardiñas.*

One member of an identical twin pair was removed from the analysis; otherwise all individuals were retained in the TRACK-HD sample, as their relatedness and admixture can be accommodated well by using association methods based on mixed linear models (Thornton et al., 2014, Shin and Lee, 2015, Yang et al., 2014a).

### 3.2.8 Imputation

As described above almost 2.5 million markers were genotyped using the Illumina arrays. The technique of genetic imputation, in which short stretches of haplotype are used to provide useful information about untyped genetic markers, was used to determine information about the untyped markers, thus increasing the power of the subsequent GWAS (Li et al., 2009). Simplistically, study samples are genotyped, and these genotypes are compared to a reference imputation panel of haplotypes that includes detailed information on a much larger
number of markers, this information can be used to predict the genotype of markers that were not directly genotyped (Figure 3.3). In this case, study samples genotyped for almost 2.5 million genetic markers, were compared to a reference panel of haplotypes that includes detailed information around 10 million markers.

Figure 3.3: Genotype imputation in a sample of apparently unrelated individuals. A: the observed data which consists of genotypes at a modest number of genetic markers in each sample being studied and of detailed information on genotypes (or haplotypes) for a reference sample. B: the process of identifying regions of chromosome shared between a study sample and individuals in the reference panel. When a typical sample of European ancestry is compared to haplotypes in the HapMap reference panel, stretches of >100kb in length are typically identified. C: observed genotypes and haplotype sharing information have been
combined to fill in a series of unobserved genotypes in the study sample. (Figure from (Li et al., 2009), Image reproduced with permission of the rights holder, Annual Reviews)

TRACK-HD was imputed by our collaborators Dr Pardiñas and Professor Peter Holmans in the Cardiff University high-performance computing cluster RAVEN (Advanced Research Computing @ Cardiff (ARCCA)), using the SHAPEIT/IMPUTE2 algorithms (Howie et al., 2012, Delaneau et al., 2013) and a standardised pipeline (van Leeuwen et al., 2015). The 1000 Genomes phase 3 panel provided by the IMPUTE2 authors (release October 2014), was used as the reference imputation panel. Imputation probabilities (“dosages”) were converted to best-guess genotypes in fcGENE v1.07 (Roshyara and Scholz, 2014) using a minimum probability threshold of 80% and a per-SNP missingness threshold of 5% of the sample. After this process an INFO score cut-off of 0.8 was applied in order to select well-imputed variants, and all monomorphic and singleton markers were excluded. With these filters 9.65 million biallelic markers remained in the dataset.

Genotypes for the REGISTRY subjects were obtained from the GeM-HD Consortium (Consortium, 2015a) where details of their genotyping and imputation are more extensively provided. DNA samples from the EHDN Registry study were obtained from the BioRep Inc. repository (Milan, Italy) and phenotypic data recorded from each of the EHDN Registry sites were provided from the central EHDN database. Genotyping was performed at the Broad Institute using Illumina Omni2.5 arrays. A standard quality control was used: SNPs with genotyping call rate >95%, minor allele frequency >1%, Hardy-Weinberg Equilibrium p-value >1E-6, and samples with genotyping call rate >95% were identified for subsequent genotype imputation. In addition, when data were available, samples with ambiguous gender, DNA contamination, and significant discordant genotype between fingerprint data and full data were excluded. The MACH program (Abecasis, 2017) was used for haplotype phasing and MINIMAC program for genotype imputation (Michegan, Howie et al., 2012). Resulting dosage data were transformed into PLINK program compatible genotype data. SNPs with imputation quality score (i.e., Rsq) >0.5 were used for the subsequent association analysis. From the GeM dataset we selected all samples from REGISTRY who passed the genetic QC and had adequate phenotypic data to generate a progression score on the subject. The resulting REGISTRY dataset harboured 8.94 million biallelic markers of 1,773 individuals (Figure 3.1).

3.2.9 Mixed linear model GWAS
Association analyses was conducted with the mixed linear model (MLM) functions included in GCTA v1.26 (Yang et al., 2011), specifically the leave-one-chromosome-out (LOCO) procedure (Yang et al., 2014b). MLMs are useful tools for conducting association mapping in the presence of sample structure, including geographic population structure, family relatedness and/or cryptic relatedness, as the model takes these characteristics into account (Yang et al., 2014a). The basic approach involves building a genetic relationship matrix (GRM) that models genome-wide sample structure, estimating the contribution of the GRM to phenotypic variance using a random effects model (a kind of hierarchical linear model) and computing association statistics that account for this component of phenotypic variance. Phenotypic variables already controlled for the relevant clinical co-variates (in the progression analysis). Therefore, no covariates were added to the analyses.

In order to transform the results into independent GWAS signals, PLINK was again used to perform linkage disequilibrium (LD) clumping ($r^2 = 0.1$, $p < 1 \times 10^{-4}$; window size < 3 Mb). Due to the relatively small size of the TRACK-HD and REGISTRY samples, analyses were restricted to SNPs with minor allele frequency >1%. Small sample sizes also meant that calculation of SNP-based heritability ($h^2_{SNP}$) for our tested phenotypes was not possible using either genotyped or imputed markers (Yang et al., 2010b, Yang et al., 2015).

Meta-analysis of the GWAS summary statistics from the TRACK-HD and REGISTRY studies was carried out by Dr Pardiñas and Professor Peter Holmans using the fixed effects method with inverse-variance weights as implemented in METAL (Willer et al., 2010). To control for spurious results due to scale differences between the TRACK-HD and REGISTRY progression phenotypes, effect sizes from both summary statistics were standardised to have equal variances before meta-analysis.

QQ plots of observed log p-values (sorted by value) for each SNP versus their expected values in the absence of association are shown for TRACK-HD, REGISTRY and the meta-analysis are shown in Figure 3.4. If there is no association, and no systematic inflation in the test statistics (for example, from population stratification), the observed log p-values would follow their expected values (the red line in Figure 3.4) exactly. Indeed, this is what is observed for the majority of data points, which do not show association. The extent to which such systematic inflation exists is measured by the genomic inflation factor $\lambda$ (Devlin and Roeder, 1999), which is the median of the observed test statistics divided by 0.456 (the median of a chi-squared distribution on 1df). Values of $\lambda$ close to 1 – as is the case here – indicate a lack of inflation.
The 95% confidence interval for log p-values in the absence of association is shaded grey, and the points lying above this in the top right corner indicate genuine associations.
Figure 3.4: QQ plots of the (A) TRACK-HD and (B) REGISTRY genome wide association studies. And (C) meta-analysis. \( \lambda \) close to 1 shows there is no systematic inflation of test statistics. (Figure prepared by Dr Pandiñas).

Conditional analyses of GWAS summary statistics were carried out using the COJO procedure included in GCTA v1.26 (Yang et al., 2012).

### 3.2.10 Co-localisation analyses

In order to discern if our top GWAS signals were mediated by the same SNPs in both TRACK-HD and REGISTRY, the co-localisation method of Giambartolomei et al. (Giambartolomei et al., 2014), as implemented by Dr Pardiñas in GWAS-pw v0.21 (Pickrell et al., 2016) was used after discussion with Dr Vincent Plagnol. In summary, the GWAS summary statistics of our two samples were first divided into approximately independent LD blocks (Berisa and Pickrell, 2016), and each block was then scanned to estimate the probability (in a hierarchical Bayesian framework) of harbouring an association common to the two samples. In contrast to the original algorithm, the model priors do not need to be pre-specified in GWAS-pw, as they are estimated directly from the summary statistics. This implementation has been thoroughly tested by simulation and applied to real data from heterogeneous sources (Pickrell et al., 2016). By testing the entire genome instead of a small number of candidate regions arising from the GWAS clumps, a conservative approach is followed towards estimating co-localisation, which also has the desirable property of allowing us to compare our candidates (to the resolution of single SNPs) with every other region in the genome.
A similar procedure was used to test for co-localisation between the region on chromosome 5 containing GWAS signal in TRACK-HD and REGISTRY and SNPs influencing expression (eQTLs), since this may indicate which gene in an association region is causal. Given that eQTLs close to the gene (cis-eQTLs) tend to replicate more reliably than those from other parts of the genome (Ramasamy et al., 2014), these analyses were restricted to the regions of GWAS signal and genes within 1Mb of these regions. These analyses used expression data from 53 tissues, accessed through GTex (Consortium, 2015b). To minimise multiple testing, the two tissues showing the most significant eQTLs for each gene were used for the co-localisation analysis. Additionally, for DHFR and MSH3, analyses were performed using three brain tissues (caudate, cerebellum and cortex), since these are the most biologically relevant to HD a priori.

### 3.2.11 Gene based analyses

Gene-wide p-values were calculated by Dr Pardiñas and Professor Peter Holmans using MAGMA v1.05 (de Leeuw et al., 2015) on the TRACK-HD and REGISTRY summary statistics, by summing the p-values of all SNPs inside each gene. MAGMA aggregates the association evidence across all SNPs in a gene, while correcting for LD between SNPs (See Chapter 2, General Methods) for an introduction to MAGMA; in this case the European data from Phase 3 of the 1000 Genomes Project were used as reference). This analysis increases power when a gene contains multiple causal SNPs (e.g. as a result of allelic heterogeneity), or when the causal SNP is not typed and its signal is partially captured by multiple genotyped SNPs in LD with it. We set a window of 35 kb upstream and 10 kb downstream of each gene in order to capture the signal of proximal regulatory SNPs (Maston et al., 2006, The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium, 2015).

### 3.2.12 Gene-set analyses

The principles behind gene-set analysis and different types of gene-set analysis are described in the General Methods (Chapter 2). To maximise comparability with the GeM GWAS (Consortium, 2015a), our primary gene-set analyses used Setscreen (Moskvina et al., 2011). Setscreen sums the (log-) p-values of all SNPs in the gene set, similar to Fisher’s method, but adjusts the distribution to allow for non-independence of SNPs due to linkage disequilibrium (Brown, 1975a). Significant enrichments from the Setscreen analyses were confirmed using the competitive gene-set analysis procedure implemented in MAGMA. This more conservative approach tests whether genes in a gene set have more significant gene-wide p-values than other genes, correcting for gene size, SNP density and intergenic linkage disequilibrium (de Leeuw et al., 2015), but may be less powerful than the Setscreen analysis for small gene sets.
Initially, gene set analyses were performed on the 14 pathways found to be significantly enriched for association signal in the GeM GWAS (Consortium, 2015a). Many of these pathways relate to DNA repair, so we investigated the biological specificity of this signal further by analysing 78 gene-sets taken from a recent review of DNA repair (Pearl et al., 2015).

As a secondary analysis, to potentially uncover areas of novel disease-related biology, the same broad list of gene sets used by GeM-HD Consortium (2015) was tested. This comprises a collection of 14,706 pathways containing between 3 and 500 genes from the Gene Ontology (GO)(Consortium, 2015d), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016), Mouse Genome Informatics (MGI) (Eppig et al., 2015), National Cancer Institute (NCI) (Schaefer et al., 2009), Protein ANalysis THrough Evolutionary Relationships (PANTHER) (Mi et al., 2013), BioCarta (Nishimura, 2001) and Reactome (Fabregat et al., 2016). Multiple testing correction was carried out for this analysis by calculating q-values (Storey and Tibshirani, 2003).

### 3.2.13 Linking genetic variation to clinical measures

To explain how our TRACK-HD lead variant (rs557874766) affected commonly used clinical measures of HD severity we first correlated TRACK-HD progression score with UHDRS Total Motor Score (TMS) and UHDRS Total Functional Capacity (TFC). We defined “raw” TMS rate as TMS change divided by follow-up years and “adjusted” TMS rate as the residual of raw TMS rate after regressing off effects of initial TMS, age, sex, CAG. We followed the same procedure for TFC.

Regressing these measures on progression gives the following estimates of the amount of change for one unit increase in progression (Table 3.2)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Effect of one unit change in subject’s progression score on this variable</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw TMS rate</td>
<td>0.71</td>
<td>0.19</td>
</tr>
<tr>
<td>Adjusted TMS rate</td>
<td>0.57</td>
<td>0.18</td>
</tr>
<tr>
<td>Raw TFC rate</td>
<td>0.21</td>
<td>0.047</td>
</tr>
<tr>
<td>Adjusted TFC rate</td>
<td>0.20</td>
<td>0.044</td>
</tr>
</tbody>
</table>

*Table 3.2: Relationship between change in progression score and rate of change in Total Motor Score (TMS) and Total Functional Capacity (TFC).*
3.3 Results

3.3.1 Phenotypic clusters of Huntington’s disease were not observed

We first compared the results when all phenotypic variables were combined in a common analysis to the results when variables were grouped into brain imaging, quantitative motor and cognitive domains.

We performed individual PCA of each domain and found that first PC scores were highly correlated between the domains (P < 0.0001 in all cases, Table 3.3). No phenotypic subtypes of symptom clusters in motor, cognitive or brain imaging domains were observed; rather, longitudinal change in TRACK-HD not predictable by CAG-age was distributed on a correlated continuum (Figure 3.5). We therefore repeated PCA of the measures combined across all domains. The first PC of this combined analysis accounted for 23.4% of the joint variance, its dominance is shown in the Scree plot (Figure 3.6). This first PC was also at least moderately correlated (r>0.4) with most of the variables that contributed heavily to each domain-specific first PC (Table 3.4).

We did consider whether to include psychiatric variables in the progression analysis, however in exploratory analysis the first psychiatric PC has notably lower correlation with motor and cognitive domains and CPO variables, psychiatric variables were therefore excluded.
Table 3.3: Correlations among Domain-Specific Residual Principal Components in the TRACK-HD analysis, showing that the first principal components of each domain are significantly correlated.

The prefaces “brain”, “cog”, and “mot” indicate the domain. The suffix f1, f2, etc, numbers the principal components within each domain.
Having approximated the residual longitudinal variability within each of the three domains via principal components, we then examined cross-domain relationships among these components. For example, after accounting for CAG-age-risk, testing whether residual longitudinal change in the brain measures correlated with the Q-motor measures.

**Figure 3.5:** Distribution of progression measure in 218 members of TRACK-HD cohort. Curve is the normal distribution approximations of the severity score distributions. (Figure prepared by Prof Langbehn, edited by me, version of this figure used in Figure 2 in (Hensman Moss et al., 2017b)).
**Figure 3.6**: The first principal component accounts for a high proportion of the variance in the TRACK-HD progression analysis. (A) Scree Plot and (B) Plot showing proportion of variance explained in the TRACK-HD progression principal component analysis: the dominance of the first PC is illustrated. Figure prepared by Antonio Pardiñas.

<table>
<thead>
<tr>
<th>Measure</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
<th>PC8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol Digit</td>
<td>-0.505</td>
<td>-0.027</td>
<td>0.135</td>
<td>0.194</td>
<td>0.034</td>
<td>0.047</td>
<td>-0.394</td>
<td>-0.121</td>
</tr>
<tr>
<td>Stroop Word</td>
<td>-0.391</td>
<td>-0.017</td>
<td>0.361</td>
<td>0.468</td>
<td>0.078</td>
<td>-0.232</td>
<td>0.087</td>
<td>0.123</td>
</tr>
<tr>
<td>Paced Tapping 3 Hz (inverse std dev)</td>
<td>-0.054</td>
<td>-0.123</td>
<td>-0.031</td>
<td>-0.066</td>
<td>0.032</td>
<td>0.621</td>
<td>-0.420</td>
<td>0.233</td>
</tr>
<tr>
<td>Spot the Change 5K</td>
<td>0.224</td>
<td>-0.123</td>
<td>0.113</td>
<td>-0.223</td>
<td>-0.016</td>
<td>0.190</td>
<td>0.427</td>
<td>0.479</td>
</tr>
<tr>
<td>Emotion Recognition</td>
<td>-0.226</td>
<td>0.188</td>
<td>0.228</td>
<td>0.086</td>
<td>-0.090</td>
<td>-0.415</td>
<td>0.098</td>
<td>0.264</td>
</tr>
<tr>
<td>Direct Circle (Log annulus length)</td>
<td>-0.374</td>
<td>-0.101</td>
<td>0.419</td>
<td>0.199</td>
<td>0.488</td>
<td>0.258</td>
<td>0.060</td>
<td>-0.027</td>
</tr>
<tr>
<td>Indirect Circle (Log annulus length)</td>
<td>-0.406</td>
<td>-0.076</td>
<td>0.407</td>
<td>0.418</td>
<td>0.161</td>
<td>0.336</td>
<td>0.036</td>
<td>0.130</td>
</tr>
<tr>
<td>Total brain volume</td>
<td>0.749</td>
<td>-0.457</td>
<td>0.168</td>
<td>0.077</td>
<td>-0.046</td>
<td>-0.100</td>
<td>-0.115</td>
<td>-0.079</td>
</tr>
<tr>
<td>Ventricular volume</td>
<td>-0.545</td>
<td>0.509</td>
<td>-0.079</td>
<td>-0.125</td>
<td>0.094</td>
<td>0.131</td>
<td>0.274</td>
<td>0.043</td>
</tr>
<tr>
<td>Grey matter volume</td>
<td>0.631</td>
<td>-0.491</td>
<td>0.173</td>
<td>-0.050</td>
<td>-0.088</td>
<td>-0.137</td>
<td>0.038</td>
<td>-0.022</td>
</tr>
<tr>
<td>White matter volume</td>
<td>0.699</td>
<td>-0.409</td>
<td>0.252</td>
<td>-0.085</td>
<td>-0.019</td>
<td>-0.048</td>
<td>0.062</td>
<td>0.044</td>
</tr>
<tr>
<td>Caudate volume</td>
<td>0.584</td>
<td>-0.426</td>
<td>0.082</td>
<td>0.223</td>
<td>0.086</td>
<td>0.083</td>
<td>-0.055</td>
<td>0.046</td>
</tr>
<tr>
<td>Metronome tapping, nondominant hand (log of tap initiation SD for all trials)</td>
<td>0.433</td>
<td>-0.033</td>
<td>-0.206</td>
<td>-0.338</td>
<td>0.104</td>
<td>0.392</td>
<td>0.037</td>
<td>-0.081</td>
</tr>
<tr>
<td>Metronome tapping, nondominant hand (inv tap initiation SD for self-paced trials)</td>
<td>-0.033</td>
<td>-0.212</td>
<td>0.013</td>
<td>0.144</td>
<td>0.116</td>
<td>0.133</td>
<td>0.347</td>
<td>-0.705</td>
</tr>
<tr>
<td>Speeded tapping, nondominant hand (log of repetition time SD)</td>
<td>0.380</td>
<td>-0.022</td>
<td>-0.483</td>
<td>0.315</td>
<td>0.554</td>
<td>-0.206</td>
<td>-0.058</td>
<td>0.123</td>
</tr>
<tr>
<td>Speeded tapping, nondominant hand (log of tap duration SD)</td>
<td>0.594</td>
<td>0.028</td>
<td>-0.335</td>
<td>0.182</td>
<td>0.437</td>
<td>-0.061</td>
<td>0.027</td>
<td>0.206</td>
</tr>
<tr>
<td>Speeded tapping, nondominant hand (mean)</td>
<td>0.316</td>
<td>0.373</td>
<td>-0.219</td>
<td>0.006</td>
<td>0.411</td>
<td>-0.036</td>
<td>-0.002</td>
<td>-0.120</td>
</tr>
</tbody>
</table>
### Table 3.4: PCA of Residual Longitudinal Change Among Variables from All 3 Domains in the TRACK-HD analysis showing that the variables that correlated with the domain specific analyses also correlated with the common principal component analysis. Dom- dominant; nondom- nondominant; std dev- standard deviation.

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Variance Explained (%)</th>
<th>Intertap Time</th>
<th>Tongue Force—Heavy (log coefficient of variation)</th>
<th>Tongue Force—Light (log coefficient of variation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grip force, dom. hand, heavy condition (log of mean orientation)</td>
<td>23.4</td>
<td>9.5</td>
<td>7.1</td>
<td>6</td>
</tr>
<tr>
<td>Grip force, dom. hand, heavy condition (log of mean position)</td>
<td>0.147</td>
<td>0.016</td>
<td>-0.332</td>
<td>0.586</td>
</tr>
<tr>
<td>Grip force, nondom. hand, heavy condition (log of coefficient of variation)</td>
<td>0.247</td>
<td>0.114</td>
<td>-0.399</td>
<td>0.451</td>
</tr>
<tr>
<td>Grip force, dom. hand, light condition (log of coefficient of variation)</td>
<td>0.516</td>
<td>0.213</td>
<td>0.108</td>
<td>0.003</td>
</tr>
<tr>
<td>Grip force, nondom. hand, light condition (log of coefficient of variation)</td>
<td>0.615</td>
<td>0.488</td>
<td>0.252</td>
<td>0.009</td>
</tr>
<tr>
<td>Grip force, dom. hand, light condition (log of coefficient of variation)</td>
<td>0.568</td>
<td>0.518</td>
<td>0.207</td>
<td>0.033</td>
</tr>
<tr>
<td>Grip force, nondom. hand, light condition (log of coefficient of variation)</td>
<td>0.647</td>
<td>0.311</td>
<td>0.250</td>
<td>0.034</td>
</tr>
<tr>
<td>Variance Explained (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### 3.3.2 The progression scores are correlated with change in more widely used clinical measures of Huntington’s disease

The cross-domain first principal component was used as a unified Huntington’s disease progression measure in the TRACK-HD cohort (Figure 3.5). To confirm that our progression measure correlated with commonly recognised measures of Huntington’s disease severity not included in the progression analysis, we examined the residual change relationships between the progression score and UHDRS TMS change and TFC change after controlling for the CPO. We found a correlation of \( r = 0.448 \) (\( p < 0.0001 \)) for the residual motor slope and \( r = 0.421 \) (\( p < 0.0001 \)) for the residual TFC slope. One unit increase in unified Huntington’s disease progression measure corresponded to an increase of 0.71 (95% CI=0.34,1.08) units per year in the rate of change of TMS, and an increase of approximately 0.2 (95% CI=0.12,0.30) units per
year in the rate of change of TFC. The 15 fastest progressing subjects in TRACK-HD showed a mean annual rate of decline in the UHDRS TMS of 2·52 more points than would be expected; the 15 slowest progressing subjects had an annual TMS decline of 0·45 points less than predicted by age and CAG length.

3.3.3 Cross-sectional severity score used as the progression measure in REGISTRY

The longitudinal unified HD progression measure developed in TRACK-HD could not be transferred directly to REGISTRY subjects due to more limited data. Individual clinical measures in REGISTRY showed correlations across the motor, cognitive, and functional domains (see Table 3.5 for loading onto PCs), consistent with our finding in TRACK-HD. The first principal component, PC1, in the REGISTRY analysis accounted for 75·6% of the variance in severity; no other principal components explained any substantial amount of the common variance within the measures used (Table 3.5). The dominance of the first principal component is shown in Figure 3.7.

![Scree Plot and Variance Explained](image)

**Figure 3.7**: The first principal component accounts for a high proportion of the variance in the REGISTRY progression analysis. (A) Scree Plot and (B) Plot showing proportion of variance explained in the REGISTRY progression principal component analysis: the dominance of the first PC is illustrated. Figure prepared by Antonio Pardiñas.

Therefore this first principal component was chosen as a measure of severity in the REGISTRY cohort: this is referred to as the unified REGISTRY progression measure. The distribution of REGISTRY progression scores in the cohort analysed is given in Figure 3.8. Higher values of this measure mean greater severity than expected at a given time: we infer that this is the result of faster progression (Figure 3.9) and we used this as the unified Registry progression measure.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable explanation</th>
<th>Factor1</th>
<th>Factor2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sqrtmotor</td>
<td>Square root of the UHDRS total motor score</td>
<td>-0.84233</td>
<td>0.30062</td>
</tr>
<tr>
<td>verfl</td>
<td>UHDRS verbal fluency</td>
<td>0.79108</td>
<td>0.24136</td>
</tr>
<tr>
<td>sdmt</td>
<td>UHDRS symbol digit score</td>
<td>0.89833</td>
<td>0.1522</td>
</tr>
<tr>
<td>scnt</td>
<td>UHDRS Stroop colour naming</td>
<td>0.89596</td>
<td>0.25872</td>
</tr>
<tr>
<td>swrt</td>
<td>UHDRS Stroop word reading</td>
<td>0.88978</td>
<td>0.2109</td>
</tr>
<tr>
<td>sit1</td>
<td>UHDRS Stroop interference score</td>
<td>0.87684</td>
<td>0.21789</td>
</tr>
<tr>
<td>tfc</td>
<td>UHDRS total functional capacity</td>
<td>0.8746</td>
<td>-0.39367</td>
</tr>
<tr>
<td>fasscore</td>
<td>UHDRS functional assessment scale</td>
<td>0.88355</td>
<td>-0.38555</td>
</tr>
</tbody>
</table>

Table 3.5: Factor pattern of the first two principal component analysis of the REGISTRY severity score which was used as a progression score for the Registry data.

Factor 1 = 1st PC; Factor 2 = 2nd PC.

Figure 3.8: Distribution of atypical severity (compared to predicted severity at final visit) in 1835 members of the REGISTRY cohort. The curve is the normal distribution approximations of the severity score distributions. (Figure prepared by Prof Langbehn, edited by me, version of this figure used in Figure 2 in (Hensman Moss et al., 2017b)).
**Figure 3.9:** Assessing progression in Huntington’s disease. Graphical illustration of the trajectory of HD symptoms and signs over time, annotated to show what time period the different measures of onset and progression discussed in this paper cover. The TRACK-HD progression score uses longitudinal data over 3 years. Given limited longitudinal data in REGISTRY, cross-sectional severity at last visit compared to predicted severity was used as a proxy for progression. Age at onset occurs when a subject has unequivocal motor signs of Huntington’s disease. (Figure prepared by me, version of this figure used in Figure 2 in (Hensman Moss et al., 2017b)).

A figurative summary of what time period the TRACK-HD and REGISTRY progression scores encapsulate, and the relationship between them and age at onset is given in Figure 3.9 above.

### 3.3.4 The TRACK-HD and REGISTRY progression measures are correlated

To ensure that the unified TRACK-HD progression measure and the unified REGISTRY progression measure encapsulated very similar clinically relevant information we explored the relationship between them. Within the TRACK data, the last-visit severity scores had a correlation of 0.674 ($p<0.0001$) with the previously calculated longitudinal progression scores. We used a Pearson correlation since it can be shown that the predicted values obtained from the TRACK-HD and REGISTRY formulas are nearly linear (Figure 3.10).

We were therefore satisfied that our progression measures for TRACK and REGISTRY reflected substantially related elements of phenotype. Further support for this conclusion was given by 14 subjects present in both studies: we found that there was a correlation of 0.631 ($p = 0.0156$) between the progression scores generated by the TRACK-HD longitudinal analysis.
method, and the REGISTRY cross sectional progression analysis method in these 14 overlapping subjects.

![Figure 3.10: TRACK-HD and REGISTRY progression scores are correlated. Linear relationship between the longitudinal atypical severity scores used for the TRACK-HD analysis and cross-sectional atypical severity scores at the last TRACK visit when calculated using the method employed for the REGISTRY data (r = .674).]

3.3.5 Progression scores are associated with AAO

In the TRACK-HD cohort, Huntington’s disease subjects in the early stages of the disease were significantly faster progressors on the unified HD progression measure than those still in the premanifest phase (p < 0.0001). Amongst the 96 subjects who had experienced onset, the rater AAO showed the expected relation with predicted AAO based on CAG length (Figure 3.11), and earlier than predicted AAO was correlated with faster progression on our unified HD progression measure (r=0.315; p = 0.002).
Figure 3.11: Observed versus Expected Age of Onset among those who have Experienced Onset in the TRACK-HD analysis: amongst these 96 subjects who had experienced onset, the rater AAO showed the expected relation with predicted AAO based on CAG length. Earlier than predicted onset age was correlated with faster progression (using the unified HD progression measure) ($r = -0.315; p = 0.002$). rater_aao_age - rater AAO; expbirth - the AAO predicted from birth based on HTT CAG repeat.

The unified REGISTRY progression measure and AAO were modestly, but significantly, correlated ($r = 0.2338; p<0.0001$) (Figure 3.12). Interestingly, atypically rapidly or slowly progressing subjects tend to become more atypical over time: correlation between time since disease onset and REGISTRY progression ($-0.3074; p<0.0001$) is greater than that between AAO and REGISTRY progression.
Figure 3.12: REGISTRY progression measure (Residual severity score) and atypical onset age (Standardised onset) are modestly correlated in REGISTRY. Note bias for very late expected onset for those with low CAG repeats. SD = Standard deviation. (Figure from Prof Langbehn)

3.3.6 Genome wide association analysis highlights a locus associated with HD progression on chromosome 5 in TRACK-HD

We performed a genome-wide association analysis using the unified TRACK-HD progression measure as a quantitative trait, which yielded a significantly associated locus on chromosome 5 spanning DHFR, MSH3 and MTRNR2LR in the TRACK-HD study. The index SNP rs557874766 is a coding missense variant in MSH3 (p =5·8x10⁻⁸; G=0·2179/1091 (1000 Genomes); Figure 3.13A, Figure 3.14 and Table 3.6), and classed as of moderate-impact, which arguably reduces the genome-wide significance threshold to P = 1·2 x 10⁻⁷ (Sveinbjornsson et al., 2016). The genes in this locus were the only ones to reach genome-wide gene-wide significance (i.e. p<2·5x10⁻⁶ (Kiezun et al., 2012)) in a MAGMA analysis (de Leeuw et al., 2015) (MTRNR2L2 p=2·14x10⁻⁹, MSH3 p=2·94x10⁻⁸, DHFR p=8·37x10⁻⁷; Table 3.7).
**Figure 3.13:** Genome-wide Association Analysis of Progression Score. Green line in A-C: $5.8 \times 10^{-8}$. (A) Manhattan plot of TRACK-HD GWA analysis yielding a locus on chromosome 5. Significance of SNPs (log10[p value], y axis) is plotted against genomic location (x axis). (B) Manhattan plot of REGISTRY GWA analysis showing suggestive trails on chromosome 15 in the same area as the GeM GWAS significant locus, and also on chromosome 5 in the same area as the TRACK progression GWAS. (C) Manhattan plot of Meta-analysis of TRACK and REGISTRY progression analysis showing that the meta-analysis strengthens the association at the chromosome 5 locus. (Manhattan plots produced by Dr Pardiñas, figure prepared by me, these plots were also adapted for publication as Figure 3 in (Hensman Moss et al., 2017b))
Figure 3.14: Locus zoom plot of the TRACK-HD (top), REGISTRY (middle) and meta-analysis (bottom) data showing the structure of linkage disequilibrium (LD) and $-\log^{10}(p\text{-value})$ of the significant locus on chromosome. The top image shows the chromosome; the red square shows the region which is zoomed in on in the other panels. The colours of the circles are based on $r^2$ with the lead SNP in TRACK-HD as shown in the bottom of the plot; intensity of colour reflects multiple overlying SNPs. Dashed lines: $5\times10^{-8}$.

(Plots produced by Dr Pardiñas, these plots were also adapted for publication as Figure 3 in (Hensman Moss et al., 2017b))
<table>
<thead>
<tr>
<th>Chr</th>
<th>Start (BP)</th>
<th>End (BP)</th>
<th>Index SNP</th>
<th>A1</th>
<th>A2</th>
<th>MAF</th>
<th>INFO score</th>
<th>Beta</th>
<th>Standard Error</th>
<th>P-value</th>
<th>No. of SNPs</th>
<th>Length (KB)</th>
<th>Gene(s) tagged (+/- 20 KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>79895438</td>
<td>80196258</td>
<td>rs557874766</td>
<td>G</td>
<td>C</td>
<td>0.238</td>
<td>1.000</td>
<td>-0.581</td>
<td>0.107</td>
<td>5.80E-08</td>
<td>380</td>
<td>300.82</td>
<td>DHFR, MSH3, MTRNR2L2</td>
</tr>
<tr>
<td>4</td>
<td>74064920</td>
<td>74362359</td>
<td>rs16849472</td>
<td>T</td>
<td>C</td>
<td>0.019</td>
<td>1.000</td>
<td>1.677</td>
<td>0.318</td>
<td>1.34E-07</td>
<td>10</td>
<td>297.44</td>
<td>AFM, AFP, ALB, ANKRD17, LOC728040</td>
</tr>
<tr>
<td>3</td>
<td>20860340</td>
<td>20919615</td>
<td>rs111902872</td>
<td>T</td>
<td>C</td>
<td>0.012</td>
<td>0.920</td>
<td>2.419</td>
<td>0.460</td>
<td>1.47E-07</td>
<td>2</td>
<td>59.276</td>
<td>none</td>
</tr>
<tr>
<td>1</td>
<td>239493679</td>
<td>239917976</td>
<td>rs115206404</td>
<td>A</td>
<td>G</td>
<td>0.009</td>
<td>0.805</td>
<td>2.598</td>
<td>0.503</td>
<td>2.46E-07</td>
<td>2</td>
<td>424.3</td>
<td>CHRM3, CHRM3-AS2</td>
</tr>
<tr>
<td>13</td>
<td>89829918</td>
<td>89856005</td>
<td>rs546753686</td>
<td>A</td>
<td>G</td>
<td>0.009</td>
<td>0.949</td>
<td>2.610</td>
<td>0.506</td>
<td>2.50E-07</td>
<td>2</td>
<td>26.088</td>
<td>none</td>
</tr>
<tr>
<td>6</td>
<td>31892827</td>
<td>31895971</td>
<td>rs188144048</td>
<td>G</td>
<td>C</td>
<td>0.016</td>
<td>1.000</td>
<td>-1.923</td>
<td>0.380</td>
<td>4.30E-07</td>
<td>2</td>
<td>3.145</td>
<td>C2, CFB, LOC102060414</td>
</tr>
<tr>
<td>4</td>
<td>52815077</td>
<td>52815077</td>
<td>rs151302971</td>
<td>C</td>
<td>T</td>
<td>0.060</td>
<td>0.998</td>
<td>0.963</td>
<td>0.192</td>
<td>4.98E-07</td>
<td>1</td>
<td>0.001</td>
<td>none</td>
</tr>
<tr>
<td>10</td>
<td>132818509</td>
<td>132818509</td>
<td>rs150136271</td>
<td>T</td>
<td>C</td>
<td>0.007</td>
<td>0.845</td>
<td>2.881</td>
<td>0.582</td>
<td>7.38E-07</td>
<td>3</td>
<td>62.805</td>
<td>TCERG1L</td>
</tr>
<tr>
<td>8</td>
<td>128074135</td>
<td>128092501</td>
<td>rs76712904</td>
<td>T</td>
<td>A</td>
<td>0.009</td>
<td>1.000</td>
<td>2.532</td>
<td>0.512</td>
<td>7.68E-07</td>
<td>13</td>
<td>18.367</td>
<td>PCAT2, PRNCR1</td>
</tr>
<tr>
<td>6</td>
<td>147033320</td>
<td>147049507</td>
<td>rs76605780</td>
<td>G</td>
<td>A</td>
<td>0.009</td>
<td>1.000</td>
<td>2.524</td>
<td>0.512</td>
<td>8.42E-07</td>
<td>4</td>
<td>16.188</td>
<td>ADGB</td>
</tr>
</tbody>
</table>

**Table 3.6:** Independent association signals from the TRACK-HD Progression GWAS (at p-value < 10^-5).

Chr - chromosome; MAF - minor allele frequency; index SNP according to dbSNP b146 build; A1: Reference Allele; A2: Alternate Allele. (Top 10 signals shown, full data available at http://hdresearch.ucl.ac.uk/data-resources/)
<table>
<thead>
<tr>
<th>Entrez</th>
<th>Gene Symbol</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>p(TRACK)</th>
<th>p(REGISTRY)</th>
<th>p(META)</th>
<th>p(GeM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4437</td>
<td>MSH3</td>
<td>5</td>
<td>79950467</td>
<td>80172634</td>
<td>2.94E-08</td>
<td>9.52E-04</td>
<td>8.89E-11</td>
<td>1.98E-02</td>
</tr>
<tr>
<td>1719</td>
<td>DHFR</td>
<td>5</td>
<td>79922045</td>
<td>79950800</td>
<td>8.37E-07</td>
<td>8.45E-04</td>
<td>1.04E-09</td>
<td>6.46E-02</td>
</tr>
<tr>
<td>100462981</td>
<td>MTRNR2L2</td>
<td>5</td>
<td>79945819</td>
<td>79946854</td>
<td>2.15E-09</td>
<td>1.20E-03</td>
<td>1.88E-09</td>
<td>N/A</td>
</tr>
<tr>
<td>7852</td>
<td>CXCR4</td>
<td>2</td>
<td>136871919</td>
<td>136875725</td>
<td>3.96E-04</td>
<td>6.46E-03</td>
<td>4.40E-06</td>
<td>8.22E-02</td>
</tr>
<tr>
<td>54893</td>
<td>MTMR10</td>
<td>15</td>
<td>31231144</td>
<td>31283807</td>
<td>3.01E-09</td>
<td>3.49E-07</td>
<td>1.42E-05</td>
<td>2.74E-11</td>
</tr>
<tr>
<td>10873</td>
<td>ME3</td>
<td>11</td>
<td>86152150</td>
<td>86383678</td>
<td>5.07E-03</td>
<td>5.81E-02</td>
<td>2.19E-05</td>
<td>4.78E-01</td>
</tr>
<tr>
<td>118</td>
<td>ADD1</td>
<td>4</td>
<td>2845584</td>
<td>2931803</td>
<td>5.84E-02</td>
<td>2.82E-03</td>
<td>2.95E-05</td>
<td>1.16E-03</td>
</tr>
<tr>
<td>8605</td>
<td>PLA2G4C</td>
<td>19</td>
<td>48551100</td>
<td>48614109</td>
<td>3.53E-03</td>
<td>1.90E-01</td>
<td>6.73E-05</td>
<td>5.82E-02</td>
</tr>
<tr>
<td>9209</td>
<td>LRRFIP2</td>
<td>3</td>
<td>37094117</td>
<td>37217851</td>
<td>5.37E-02</td>
<td>3.16E-04</td>
<td>6.98E-05</td>
<td>3.19E-04</td>
</tr>
<tr>
<td>8690</td>
<td>JRLK</td>
<td>11</td>
<td>96123158</td>
<td>96126727</td>
<td>4.37E-05</td>
<td>5.29E-02</td>
<td>8.39E-05</td>
<td>8.91E-01</td>
</tr>
<tr>
<td>2788</td>
<td>GNG7</td>
<td>19</td>
<td>25112128</td>
<td>2702746</td>
<td>1.02E-01</td>
<td>3.62E-03</td>
<td>1.11E-04</td>
<td>7.83E-02</td>
</tr>
<tr>
<td>22909</td>
<td>FAN1</td>
<td>15</td>
<td>31196055</td>
<td>31235311</td>
<td>5.30E-01</td>
<td>2.16E-06</td>
<td>1.15E-04</td>
<td>1.68E-09</td>
</tr>
<tr>
<td>4292</td>
<td>MLH1</td>
<td>3</td>
<td>37034841</td>
<td>37092337</td>
<td>6.98E-02</td>
<td>3.97E-04</td>
<td>1.28E-04</td>
<td>3.91E-04</td>
</tr>
<tr>
<td>79780</td>
<td>CCDC82</td>
<td>11</td>
<td>96085929</td>
<td>96123083</td>
<td>2.34E-03</td>
<td>5.99E-02</td>
<td>1.30E-04</td>
<td>7.39E-02</td>
</tr>
<tr>
<td>9852</td>
<td>EPM2AIP1</td>
<td>3</td>
<td>37027357</td>
<td>37034795</td>
<td>7.94E-02</td>
<td>4.29E-04</td>
<td>1.53E-04</td>
<td>1.39E-03</td>
</tr>
<tr>
<td>4308</td>
<td>TRPM1</td>
<td>15</td>
<td>31293264</td>
<td>31453476</td>
<td>4.78E-01</td>
<td>1.77E-05</td>
<td>1.83E-04</td>
<td>8.33E-04</td>
</tr>
<tr>
<td>115509</td>
<td>ZNF689</td>
<td>16</td>
<td>30614686</td>
<td>30621682</td>
<td>1.86E-02</td>
<td>7.52E-03</td>
<td>1.85E-04</td>
<td>9.53E-01</td>
</tr>
<tr>
<td>23167</td>
<td>EFR3A</td>
<td>8</td>
<td>132916356</td>
<td>133025889</td>
<td>3.90E-02</td>
<td>1.26E-02</td>
<td>2.32E-04</td>
<td>2.19E-01</td>
</tr>
<tr>
<td>146540</td>
<td>ZNF785</td>
<td>16</td>
<td>30591994</td>
<td>30597092</td>
<td>1.17E-01</td>
<td>1.53E-02</td>
<td>2.45E-04</td>
<td>9.39E-01</td>
</tr>
<tr>
<td>Rank</td>
<td>Gene</td>
<td>Chromosome</td>
<td>Start</td>
<td>End</td>
<td>p-value</td>
<td>Bonferroni-corrected p-value</td>
<td>p(META)</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------------</td>
<td>-------</td>
<td>-----</td>
<td>-----------</td>
<td>-------------------------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>909</td>
<td>CD1A</td>
<td>158223927</td>
<td>158228059</td>
<td>1.60E-01</td>
<td>5.96E-04</td>
<td>3.27E-04</td>
<td>4.85E-01</td>
<td></td>
</tr>
</tbody>
</table>

*Table 3.7: Gene-wide p-values for top genes in TRACK-HD, REGISTRY, the TRACK-REGISTRY meta analysis (p(META)), and GeM from the MAGMA analysis. (Top 20 genes only; full data available at [http://hdresearch.ucl.ac.uk/data-resources/](http://hdresearch.ucl.ac.uk/data-resources))
Analyses conditioning on the most significant SNP (rs557874766) failed to show evidence for a second independent signal in the chromosome 5 region in TRACK-HD (Figure 3.15).

**Figure 3.15:** Regional plot of TRACK-HD and REGISTRY meta-analysis GWAS signal in the MSH3-DHFR region before (top) and after (bottom) conditioning on the most significant SNP in TRACK-HD (rs557874766). The lack of significant association after conditioning on this SNP is consistent with there being only one association signal in the region. (Figure by Dr Pardiñas)

### 3.3.7 The chromosome 5 signal is replicated in a genome wide association study in REGISTRY, and strengthened in meta-analysis

Performing a genome-wide association analysis in REGISTRY using the progression score replicated the signal identified in TRACK-HD \((p = 1.39 \times 10^{-5})\) on a narrower locus \((\text{chr}5:79902336-79950781)\), but still tagging the same three genes (Figure 3.13B, 3.14). No genes reach genome-wide significant gene-wide association in the MAGMA analysis, though
DHFR and MSH3 were still in the top 50 most associated genes (DHFR $p=8.45 \times 10^{-4}$, MSH3 $p=9.36 \times 10^{-4}$, MTRNR2L2 $p=1.20 \times 10^{-3}$, Table 3.11).

Co-localisation analyses between TRACK-HD and REGISTRY showed this locus was likely influenced by the same SNPs (posterior probability 74.33%), although conditioning REGISTRY on rs55787466 did not remove the association signal entirely (Figure 3.16).

**Figure 3.16**: Regional plot of REGISTRY GWAS signal in the MSH3-DHFR region before (top) and after (bottom) conditioning on the most significant SNP in TRACK-HD (rs557874766). The significance of association is largely unaffected by conditioning on this SNP. This indicates that rs557874766 does not explain the REGISTRY association signal in this region.

Meta-analysis of TRACK-HD and REGISTRY strengthened the signal of both individual SNPs in this region, encompassing the first three exons of MSH3 along with DHFR and MTRNR2L2 (Figures 3.13 and 3.14, Table 3.8), and also genic associations over MSH3, DHFR, and MTRNR2L2 (Table 3.7). The most significant SNP in the meta-analysis is rs1232027, which reaches genome-wide significance ($p=1.12 \times 10^{-10}$), with the p-value of rs557874766 being
No other regions attained genome-wide significance (Table 3.8). Rs557874766 is nominally significant in REGISTRY (p=0.010), with a direction of effect consistent with that in TRACK-HD. Analyses conditional on rs1232027 largely removed the association in this region (Figure 3.17A), suggesting that there is only one signal. Conditioning on rs557874766 has a similar effect (Figure 3.17B), so this SNP remains a plausible causal variant.

<table>
<thead>
<tr>
<th>Index SNP</th>
<th>P-value</th>
<th>Clump coordinates</th>
<th>Clump size (KB)</th>
<th>Gene(s) tagged</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1232027</td>
<td>1.12E-10</td>
<td>chr5:79895438..80198404</td>
<td>302.967</td>
<td>DHFR, MSH3, MTRNR2L2</td>
</tr>
<tr>
<td>rs73786719</td>
<td>8.53E-07</td>
<td>chr6:147034576..147037984</td>
<td>3.409</td>
<td>ADGB</td>
</tr>
<tr>
<td>rs114688092</td>
<td>1.51E-06</td>
<td>chr3:47026101..47315538</td>
<td>289.438</td>
<td>CCDC12, KIF9, KIF9-AS1, KLHL18, NBEAL2, NRADDP, SETD2</td>
</tr>
<tr>
<td>rs79029191</td>
<td>1.67E-06</td>
<td>chr18:8053863..8080538</td>
<td>26.676</td>
<td>PTPRM</td>
</tr>
<tr>
<td>rs932428</td>
<td>1.79E-06</td>
<td>chr20:37518361..37876772</td>
<td>358.412</td>
<td>DHX35, FAM83D, LOC39568, PPP1R16B</td>
</tr>
<tr>
<td>rs3889139</td>
<td>2.13E-06</td>
<td>chr11:6885429..6917038</td>
<td>31.61</td>
<td>OR2D2, OR10A2, OR10A4, OR10A5</td>
</tr>
<tr>
<td>rs114643193</td>
<td>2.65E-06</td>
<td>chr4:2844682..2939191</td>
<td>94.51</td>
<td>ADD1, MFSD10, NOP14, NOP14-AS1, SH3BP2</td>
</tr>
<tr>
<td>rs6882169</td>
<td>2.72E-06</td>
<td>chr5:167668230..167668230</td>
<td>0.001</td>
<td>CTB-178M22.2, TENM2</td>
</tr>
<tr>
<td>rs80260687</td>
<td>2.92E-06</td>
<td>chr8:97232364..97304966</td>
<td>72.603</td>
<td>MTERFD1, PTDSS1, UQCRB</td>
</tr>
<tr>
<td>rs28406206</td>
<td>3.13E-06</td>
<td>chr14:105680474..105688082</td>
<td>7.609</td>
<td>BRF1</td>
</tr>
<tr>
<td>rs4736525</td>
<td>3.37E-06</td>
<td>chr8:132924474..133030989</td>
<td>106.516</td>
<td>EFR3A, OC90</td>
</tr>
<tr>
<td>rs78621558</td>
<td>4.44E-06</td>
<td>chr5:80012735..80012735</td>
<td>0.001</td>
<td>MSH3</td>
</tr>
<tr>
<td>rs72715653</td>
<td>4.80E-06</td>
<td>chr4:178641337..178730329</td>
<td>88.993</td>
<td>LINCO1098, LINCO1099</td>
</tr>
<tr>
<td>rs4720024</td>
<td>4.94E-06</td>
<td>chr7:39094155..39042312</td>
<td>1.058</td>
<td>AQP1, FAM188B, INMT-FAM188B</td>
</tr>
<tr>
<td>rs117933444</td>
<td>5.75E-06</td>
<td>chr6:167362873..167410443</td>
<td>47.571</td>
<td>FGFR1OP, MIR3939, RNASET2</td>
</tr>
<tr>
<td>rs116220136</td>
<td>5.82E-06</td>
<td>chr5:23353255..23436446</td>
<td>83.192</td>
<td>none</td>
</tr>
<tr>
<td>rs8031584</td>
<td>8.15E-06</td>
<td>chr15:31185616..31292023</td>
<td>106.408</td>
<td>FAN1, MTMR10, TRPM1</td>
</tr>
<tr>
<td>rs3013648</td>
<td>9.1E-06</td>
<td>chr13:85296644..85374146</td>
<td>77.503</td>
<td>none</td>
</tr>
<tr>
<td>rs11197481</td>
<td>9.12E-06</td>
<td>chr10:117708803..117708803</td>
<td>0.001</td>
<td>ATRNL1</td>
</tr>
<tr>
<td>rs117440785</td>
<td>9.15E-06</td>
<td>chr10:17411451..17531334</td>
<td>119.884</td>
<td>ST8SIA6, ST8SIA6-AS1</td>
</tr>
</tbody>
</table>

Table 3.8: Independent association signals from the meta-analysis of TRACK-HD and REGISTRY Progression GWAS (at p-value < 10-5).
Figure 3.17: Conditional analysis. (A) Regional plot of TRACK-HD and REGISTRY meta-analysis GWAS signal in the MSH3-DHFR region before (top) and after (bottom) conditioning on the most significant SNP in the meta-analysis (rs1232027). The lack of significant association after conditioning on this SNP is consistent with there being only one association signal in the region. (B) Regional plot of TRACK-HD and REGISTRY meta-analysis GWAS signal in the MSH3-DHFR region before (top) and after (bottom) conditioning on the most significant SNP in TRACK-HD (rs557874766). The lack of significant association after conditioning on this SNP is consistent with there being only one association signal in the region.

3.3.8 Variants associated with slower HD progression are associated with decreased MSH3 expression

One of the ways in which genetic variants may result in phenotypic variation is via effects on the transcript levels of genes. Loci which are responsible for the genetic control of expression levels and patterns are known as expression quantitative trait loci (eQTLs) (Majewski and Pastinen, 2011, Nica and Dermitzakis, 2013). Co-localisation analyses with the GTex expression data (Consortium, 2015b) showed strong evidence (posterior probability 96-99%) that SNPs influencing progression in TRACK-HD were also eQTLs for DHFR in brain and peripheral tissues (Figure 3.18A).
Figure 3.18: Expression analysis. A: Regional plot of TRACK-HD GWAS signal in the MSH3-DHFR region (top, red), along with GTEx eQTL associations with DHFR expression in (top-bottom) whole blood, skeletal muscle, cerebellum, cortex. B: Regional plot of REGISTRY GWAS signal in the MSH3-DHFR region (top, blue), along with GTEx eQTL associations with MSH3 expression in (top-bottom) whole blood, transformed fibroblasts.

Conversely, there was strong evidence (posterior probability=97.8%) that progression SNPs in REGISTRY were eQTLs for MSH3 in blood and fibroblasts (Table 3.9, Figure 3.18B). Despite the lack of co-localisation between the TRACK GWAS and MSH3 expression signal, several of the most significant GWAS SNPs were associated with decreased MSH3 expression and slower progression (Table 3.9). Thus, the signal on chromosome 5 could be due to the coding change in MSH3, or to expression changes in MSH3, DHFR or both, and both effects may operate in disease.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>rs863215</td>
<td>79948005</td>
<td>T</td>
<td>C</td>
<td>8.29E-08</td>
<td>-0.5441</td>
<td>1.12E-15</td>
<td>-0.4645</td>
<td>1.43E-13</td>
<td>-0.5248</td>
<td>0.0487</td>
<td>-0.1887</td>
<td>0.00857</td>
<td>-0.2471</td>
<td>0.00042</td>
<td>-0.4224</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>rs147834</td>
<td>79949575</td>
<td>A</td>
<td>C</td>
<td>8.29E-08</td>
<td>-0.5441</td>
<td>1.12E-15</td>
<td>-0.4645</td>
<td>1.43E-13</td>
<td>-0.5248</td>
<td>0.0487</td>
<td>-0.1887</td>
<td>0.00857</td>
<td>-0.2471</td>
<td>0.00042</td>
<td>-0.4224</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>rs1382539</td>
<td>79952154</td>
<td>A</td>
<td>G</td>
<td>8.70E-08</td>
<td>-0.5432</td>
<td>1.12E-15</td>
<td>-0.4645</td>
<td>1.43E-13</td>
<td>-0.5248</td>
<td>0.0487</td>
<td>-0.1887</td>
<td>0.00857</td>
<td>-0.2471</td>
<td>0.00042</td>
<td>-0.4224</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>rs1677703</td>
<td>79957737</td>
<td>T</td>
<td>C</td>
<td>1.13E-07</td>
<td>-0.5342</td>
<td>6.12E-17</td>
<td>-0.4823</td>
<td>4.06E-15</td>
<td>-0.5518</td>
<td>0.0487</td>
<td>-0.1887</td>
<td>0.00857</td>
<td>-0.2471</td>
<td>0.00042</td>
<td>-0.4224</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>rs1650667</td>
<td>79962226</td>
<td>T</td>
<td>C</td>
<td>1.13E-07</td>
<td>-0.5342</td>
<td>3.55E-16</td>
<td>-0.4642</td>
<td>3.58E-15</td>
<td>-0.543</td>
<td>0.0487</td>
<td>-0.1887</td>
<td>0.00705</td>
<td>-0.247</td>
<td>0.00063</td>
<td>-0.3989</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>rs1650666</td>
<td>79962439</td>
<td>A</td>
<td>G</td>
<td>1.13E-07</td>
<td>-0.5342</td>
<td>3.55E-16</td>
<td>-0.4642</td>
<td>3.58E-15</td>
<td>-0.543</td>
<td>0.0487</td>
<td>-0.1887</td>
<td>0.00705</td>
<td>-0.247</td>
<td>0.00063</td>
<td>-0.3989</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>rs857287</td>
<td>80098957</td>
<td>C</td>
<td>G</td>
<td>2.63E-07</td>
<td>-0.5296</td>
<td>1.92E-11</td>
<td>0.40096</td>
<td>8.39E-10</td>
<td>0.443</td>
<td>0.6642</td>
<td>0.0396</td>
<td>0.05648</td>
<td>0.18319</td>
<td>0.01774</td>
<td>0.28212</td>
<td>G</td>
<td>C</td>
</tr>
<tr>
<td>rs863214</td>
<td>79984714</td>
<td>G</td>
<td>A</td>
<td>3.09E-07</td>
<td>-0.5302</td>
<td>6.00E-14</td>
<td>-0.44</td>
<td>6.09E-13</td>
<td>-0.5053</td>
<td>0.1665</td>
<td>-0.1406</td>
<td>0.01627</td>
<td>-0.2326</td>
<td>0.00142</td>
<td>-0.4023</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>rs1222809</td>
<td>79917517</td>
<td>G</td>
<td>A</td>
<td>3.14E-07</td>
<td>-0.5355</td>
<td>3.02E-16</td>
<td>-0.4728</td>
<td>1.08E-12</td>
<td>-0.5022</td>
<td>0.1065</td>
<td>-0.16</td>
<td>0.01615</td>
<td>-0.2273</td>
<td>0.00052</td>
<td>-0.4288</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>rs836794</td>
<td>80012251</td>
<td>A</td>
<td>C</td>
<td>4.01E-07</td>
<td>-0.5367</td>
<td>2.63E-14</td>
<td>-0.4445</td>
<td>7.40E-13</td>
<td>-0.509</td>
<td>0.1628</td>
<td>-0.1404</td>
<td>0.03621</td>
<td>-0.1998</td>
<td>0.00132</td>
<td>-0.3996</td>
<td>A</td>
<td>C</td>
</tr>
</tbody>
</table>

**Table 3.9:** Significant (p<0.001) SNPs from TRACK-HD GWAS chromosome 5 region showing direction of effect (beta) on progression (GWAS) and expression (eQTL). Negative beta means the reference allele associated with reduced progression or expression. Only 10 most significant SNPs are shown here, full data available at [http://hdresearch.ucl.ac.uk/data-resources/](http://hdresearch.ucl.ac.uk/data-resources/)
3.3.9 REGISTRY association analysis highlights locus on chromosome 15

The second most significant association region in REGISTRY (Figure 3.13, Table 3.8) tags a locus on chromosome 15 which has been previously associated to HD AAO (Consortium, 2015a). Five genes were highlighted, two of which reached genome-wide genic significance (MTMR10 p=2·51x10⁻⁷; FAN1 p=2·35x10⁻⁶; Table 3.7).

Interestingly, another DNA repair gene, MLH1 on chromosome 3 contains SNPs approaching genome-wide significance (p = 2.2 x 10⁻⁷) in GeM-HD (8), and also shows some association in the REGISTRY progression gene-wide analysis (p = 3·97x10⁻⁴; p = 1.28x10⁻⁴ in the meta-analysis).

3.3.10 The observed associations with progression are not all driven by age at onset

As noted above, both TRACK-HD and REGISTRY progression measures are correlated with AAO. Thus, to test whether there is an association with progression independent of AAO, we repeated the REGISTRY progression GWAS conditioning for the AAO measure previously associated with this locus in GeM in the individuals (N=1,314) for whom we had measures of both progression and AAO. Both MTMR10 (p=1·33x10⁻⁵) and FAN1 (p=1·68x10⁻⁴) remained significant. Furthermore, the most significant SNP (rs10611148, p=2·84x10⁻⁷) was still significant after conditioning on AAO (p=2·40x10⁻⁵).

Notably, the genic associations at the MSH3 locus in the TRACK-HD sample also remain significant after correcting for AAO (Table 3.10), as does the association with rs557874766 (p=6·30x10⁻⁶). A similar pattern is observed at the MSH3 locus in the meta-analysis. Thus, the associations reported here are mainly due to disease progression, rather than AAO.
<table>
<thead>
<tr>
<th>Entrez</th>
<th>Gene Symbol</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>p(TRACK)</th>
<th>p(TRACKcond)</th>
<th>p(REG)</th>
<th>p(REGcond)</th>
<th>p(META)</th>
<th>p(METAcond)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100462981</td>
<td>MTRNR2L2</td>
<td>5</td>
<td>79945819</td>
<td>79946854</td>
<td>2.15E-09</td>
<td>6.97E-07</td>
<td>1.20E-03</td>
<td>5.51E-02</td>
<td>1.88E-09</td>
<td>3.24E-06</td>
</tr>
<tr>
<td>4437</td>
<td>MSH3</td>
<td>5</td>
<td>79950467</td>
<td>80172634</td>
<td>2.94E-08</td>
<td>4.98E-06</td>
<td>9.52E-04</td>
<td>6.24E-02</td>
<td>8.89E-11</td>
<td>6.86E-07</td>
</tr>
<tr>
<td>1719</td>
<td>DHFR</td>
<td>5</td>
<td>79922045</td>
<td>79950800</td>
<td>8.37E-07</td>
<td>3.74E-05</td>
<td>8.45E-04</td>
<td>4.42E-02</td>
<td>1.04E-09</td>
<td>2.20E-06</td>
</tr>
<tr>
<td>8339</td>
<td>HIST1H2BG</td>
<td>6</td>
<td>26216428</td>
<td>26216872</td>
<td>1.56E-05</td>
<td>4.40E-03</td>
<td>5.10E-01</td>
<td>7.88E-01</td>
<td>2.31E-02</td>
<td>1.14E-01</td>
</tr>
<tr>
<td>387638</td>
<td>C10orf113</td>
<td>10</td>
<td>21414692</td>
<td>21435488</td>
<td>2.45E-05</td>
<td>1.46E-05</td>
<td>6.00E-01</td>
<td>3.42E-01</td>
<td>1.65E-01</td>
<td>2.22E-01</td>
</tr>
<tr>
<td>8690</td>
<td>JRKL</td>
<td>11</td>
<td>96123158</td>
<td>96126727</td>
<td>4.37E-05</td>
<td>1.30E-03</td>
<td>5.29E-02</td>
<td>4.40E-02</td>
<td>8.39E-05</td>
<td>1.06E-03</td>
</tr>
<tr>
<td>55269</td>
<td>PSPC1</td>
<td>13</td>
<td>20248892</td>
<td>20357159</td>
<td>4.80E-05</td>
<td>4.79E-04</td>
<td>5.80E-01</td>
<td>8.84E-01</td>
<td>7.95E-02</td>
<td>4.28E-02</td>
</tr>
<tr>
<td>3007</td>
<td>HIST1H1D</td>
<td>6</td>
<td>26234440</td>
<td>26235216</td>
<td>6.67E-05</td>
<td>2.82E-03</td>
<td>5.51E-01</td>
<td>8.13E-01</td>
<td>9.04E-03</td>
<td>6.40E-02</td>
</tr>
<tr>
<td>1553</td>
<td>CYP2A13</td>
<td>19</td>
<td>41594356</td>
<td>41602100</td>
<td>7.81E-05</td>
<td>4.06E-05</td>
<td>6.62E-01</td>
<td>6.74E-01</td>
<td>1.96E-03</td>
<td>4.41E-03</td>
</tr>
<tr>
<td>8369</td>
<td>HIST1H4G</td>
<td>6</td>
<td>26246839</td>
<td>26247205</td>
<td>8.49E-05</td>
<td>2.36E-03</td>
<td>5.83E-01</td>
<td>8.08E-01</td>
<td>1.18E-02</td>
<td>7.36E-02</td>
</tr>
</tbody>
</table>

Table 3.10: Gene-wide p-values for all genes in TRACK-HD, REGISTRY and the TRACK-REGISTRY meta-analysis after conditioning on AAO [p(TRACKcond); p(REGcond), p(METAcond) respectively], compared to their values without conditioning. Only 10 most significant SNPs are shown here, full data available at [http://hdresearch.ucl.ac.uk/data-resources/](http://hdresearch.ucl.ac.uk/data-resources/).
3.3.11 Effect of index MSH3 SNP on clinical measures

We found that the top MSH3 SNP in TRACK (rs557874766) is associated with a difference in the rate of change of widely used clinical measures: the Total Motor Score (TMS) and Total Functional Capacity (TFC) after controlling for the CPO.

The effect size at the top MSH3 SNP in TRACK (rs557874766) is -0.58 (S.E. =0.087) units of progression per copy of the minor allele G – this corresponds to a change of -0.33 (95% CI =0.10, 0.56) to -0.41 (0.16,0.66) units in TMS rate compared to the major allele C, which can be interpreted as a reduction in the rate of TMS increase by 0.33-0.41 units per year for each copy of the G allele. Similarly, this corresponds to a reduction in the rate of TFC change of 0.12 (0.06,0.18) units per year per G allele.

3.3.12 Pathway analysis shows association between HD progression and genes involved in DNA repair

Gene set analysis of the 14 pathways highlighted by the GeM-HD paper (Consortium, 2015a) show that the four most significant pathways in the TRACK-HD progression GWAS are related to mismatch repair, and all show significant enrichment of signal in REGISTRY (Table 3.11). This enrichment is strengthened in the meta-analysis (Table 3.11). Notably, the top two pathways in TRACK-HD are also significant in the MAGMA competitive gene-set analysis (GO:32300 p=0.010, KEGG:3430 p=0.00697). MSH3 (2.94x10^{-8}) and POLD2 (7.21x10^{-4}) show association in TRACK, with MSH3 (9.52x10^{-4}) and MLH1 (3.97x10^{-4}) showing association in REGISTRY (http://hdresearch.ucl.ac.uk/data-resources/).

<table>
<thead>
<tr>
<th>Pathway</th>
<th>p(TRACK)</th>
<th>p(REGISTRY)</th>
<th>P(META)</th>
<th>p(GeM)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO: 32300</td>
<td>3.46E-09</td>
<td>8.34E-04</td>
<td>1.14E-11</td>
<td>3.82E-05</td>
<td>mismatch repair complex</td>
</tr>
<tr>
<td>KEGG 3430</td>
<td>2.79E-07</td>
<td>4.80E-02</td>
<td>1.34E-16</td>
<td>6.65E-06</td>
<td>KEGG_MISMATCH_REPAIR</td>
</tr>
<tr>
<td>GO: 30983</td>
<td>6.66E-07</td>
<td>4.20E-04</td>
<td>3.17E-11</td>
<td>7.43E-06</td>
<td>mismatched DNA binding</td>
</tr>
<tr>
<td>GO: 6298</td>
<td>3.53E-06</td>
<td>4.59E-02</td>
<td>6.54E-09</td>
<td>3.25E-06</td>
<td>mismatch repair</td>
</tr>
<tr>
<td>GO: 1.82E-02</td>
<td>1.10E-01</td>
<td>6.40E-04</td>
<td>5.74E-06</td>
<td></td>
<td>MutSalpha complex binding</td>
</tr>
</tbody>
</table>
Table 3.11: Setscreen enrichment p-values for the 14 pathways highlighted in GeM-HD (8).

<table>
<thead>
<tr>
<th>GO:</th>
<th>32407</th>
<th>32389</th>
<th>32389</th>
<th>MutLalpha complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.25E-02</td>
<td>4.69E-02</td>
<td>5.23E-04</td>
<td>1.66E-05</td>
<td>nucleotide-excision repair, DNA incision</td>
</tr>
<tr>
<td>8.01E-02</td>
<td>5.87E-04</td>
<td>6.74E-03</td>
<td>1.69E-06</td>
<td></td>
</tr>
<tr>
<td>3.32E-01</td>
<td>5.93E-02</td>
<td>7.87E-01</td>
<td>2.30E-06</td>
<td>positive regulation of mitochondrial fission</td>
</tr>
<tr>
<td>4.10E-01</td>
<td>7.29E-01</td>
<td>6.93E-01</td>
<td>8.39E-05</td>
<td>regulation of peroxisome organization</td>
</tr>
<tr>
<td>4.58E-01</td>
<td>5.44E-01</td>
<td>5.28E-01</td>
<td>8.89E-08</td>
<td>positive regulation of release of cytochrome c from mitochondria</td>
</tr>
<tr>
<td>5.39E-01</td>
<td>3.32E-01</td>
<td>8.10E-01</td>
<td>1.57E-05</td>
<td>regulation of mitochondrial fission</td>
</tr>
<tr>
<td>6.21E-01</td>
<td>6.28E-01</td>
<td>8.53E-01</td>
<td>7.63E-05</td>
<td>positive regulation of mitochondrion organization</td>
</tr>
<tr>
<td>9.64E-01</td>
<td>6.97E-01</td>
<td>9.79E-01</td>
<td>2.66E-05</td>
<td>ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor</td>
</tr>
<tr>
<td>9.64E-01</td>
<td>6.97E-01</td>
<td>9.79E-01</td>
<td>2.66E-05</td>
<td>oxidoreductase activity, acting on CH or CH2 groups, disulfide as acceptor</td>
</tr>
</tbody>
</table>

The GO and KEGG terms in the first column refer to pathways of biologically related genes in the Gene Ontology Consortium (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) databases respectively. The p-values in columns 2 – 4 refer to the association between the pathway indicated and rate of progression described in this paper (TRACK- TRACK-HD study; REGISTRY- REGISTRY study; META- meta-analysis).

P(GeM) refers to the association between the indicated pathway and age at motor onset in the GeM-HD study (8).

These findings are supported by analysis of DNA damage response pathways derived from Pearl et al. (Pearl et al., 2015) (Figure 3.19A, Table 3.12) where two mismatch repair pathways are significantly associated with the unified TRACK-HD progression measure after correction for multiple testing of pathways. Again, the meta-analysis strengthens the enrichment (Figure 3.19B, Table 3.12). Genes from the two significant pathways in TRACK-HD
are shown in Table 3.13, with the significant genes being very similar to those from the GeM pathways in Table 3.12. A complete list of genes in the Pearl et al. (Pearl et al., 2015) pathways is given in http://hdresearch.ucl.ac.uk/data-resources/.
<table>
<thead>
<tr>
<th>Gene Set</th>
<th>p(TRACK)</th>
<th>p(REGISTRY)</th>
<th>p(META)</th>
<th>p (GeM)</th>
<th>Description 1</th>
<th>Description 2</th>
<th>Description 3</th>
<th>Description 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2071015</td>
<td>9.05E-07</td>
<td>4.43E-03</td>
<td>2.93E-11</td>
<td>2.01E-02</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>MMR</td>
<td>Mismatch &amp; loop recognition factors</td>
</tr>
<tr>
<td>2071000</td>
<td>2.43E-06</td>
<td>6.85E-02</td>
<td>1.49E-14</td>
<td>5.15E-04</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>MMR</td>
<td></td>
</tr>
<tr>
<td>2070000</td>
<td>5.77E-03</td>
<td>4.76E-02</td>
<td>3.32E-07</td>
<td>1.42E-02</td>
<td>Repair pathway</td>
<td>SSR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2071017</td>
<td>1.95E-02</td>
<td>2.44E-02</td>
<td>5.84E-05</td>
<td>8.92E-08</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>MMR</td>
<td>MutL homologs</td>
</tr>
<tr>
<td>2111513</td>
<td>4.71E-02</td>
<td>2.55E-01</td>
<td>8.12E-01</td>
<td>2.86E-03</td>
<td>Repair pathway</td>
<td>Associated process</td>
<td>TLS</td>
<td>DNA polymerases</td>
</tr>
<tr>
<td>2070600</td>
<td>5.02E-02</td>
<td>7.99E-01</td>
<td>1.10E-01</td>
<td>2.92E-01</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>NER</td>
<td></td>
</tr>
<tr>
<td>2070607</td>
<td>5.18E-02</td>
<td>7.61E-01</td>
<td>3.02E-02</td>
<td>2.26E-01</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>NER</td>
<td>TCR (Transcription coupled repair)</td>
</tr>
<tr>
<td>2071104</td>
<td>5.35E-02</td>
<td>3.90E-01</td>
<td>2.07E-02</td>
<td>5.37E-02</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>BER</td>
<td>LONG PATCH-BER factors</td>
</tr>
<tr>
<td>2022100</td>
<td>6.69E-02</td>
<td>3.19E-02</td>
<td>7.21E-04</td>
<td>7.29E-02</td>
<td>Repair pathway</td>
<td>DSR</td>
<td>Alt-NHEJ</td>
<td></td>
</tr>
<tr>
<td>1100000</td>
<td>7.52E-02</td>
<td>6.14E-01</td>
<td>1.94E-01</td>
<td>6.13E-01</td>
<td>Associated process</td>
<td>DNA replication</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1080700</td>
<td>8.99E-02</td>
<td>8.35E-01</td>
<td>2.82E-01</td>
<td>4.92E-01</td>
<td>Associated process</td>
<td>Checkpoint factors</td>
<td>S-CC phase</td>
<td></td>
</tr>
<tr>
<td>1051930</td>
<td>1.02E-01</td>
<td>5.68E-01</td>
<td>1.30E-01</td>
<td>7.62E-01</td>
<td>Associated process</td>
<td>Ubiquitin response</td>
<td>Ubiquitin- conjugating enzymes (E2)</td>
<td>UBL-conjugating enzymes</td>
</tr>
<tr>
<td>2000000</td>
<td>1.13E-01</td>
<td>2.60E-01</td>
<td>1.03E-03</td>
<td>1.11E-02</td>
<td>Repair pathway</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2070605</td>
<td>1.14E-01</td>
<td>5.00E-01</td>
<td>8.14E-01</td>
<td>4.64E-01</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>NER</td>
<td>DNA polymerase epsilon</td>
</tr>
<tr>
<td>1030000</td>
<td>1.59E-01</td>
<td>1.90E-01</td>
<td>3.59E-01</td>
<td>2.63E-01</td>
<td>Associated process</td>
<td>Telomere maintenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2070606</td>
<td>1.60E-01</td>
<td>9.56E-01</td>
<td>6.55E-01</td>
<td>5.49E-01</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>NER</td>
<td>DNA polymerase kappa</td>
</tr>
<tr>
<td>2071020</td>
<td>1.73E-01</td>
<td>3.14E-01</td>
<td>9.86E-03</td>
<td>7.97E-02</td>
<td>Repair pathway</td>
<td>SSR</td>
<td>MMR</td>
<td>Other MMR factors</td>
</tr>
</tbody>
</table>

104
Table 3.12: Setscreen enrichment p-values for the Pearl et al. (2015) pathways in TRACK-HD, REGISTRY, the TRACK-HD meta-analysis and GeM.

<table>
<thead>
<tr>
<th>Entrez</th>
<th>Gene Symbol</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>p(TRACK)</th>
<th>p(REG)</th>
<th>p(META)</th>
<th>p(GeM)</th>
<th>Pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>4437</td>
<td>MSH3</td>
<td>5</td>
<td>79950467</td>
<td>80172634</td>
<td>2.94E-08</td>
<td>9.52E-04</td>
<td>8.88E-11</td>
<td>1.98E-02</td>
<td>Repair pathway/SSR/MMR/Mismatch and loop recognition factors</td>
</tr>
<tr>
<td>5425</td>
<td>POLD2</td>
<td>7</td>
<td>44154279</td>
<td>44163169</td>
<td>7.21E-04</td>
<td>3.12E-01</td>
<td>2.75E-03</td>
<td>5.17E-01</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
<tr>
<td>3978</td>
<td>LIG1</td>
<td>19</td>
<td>48618703</td>
<td>48673560</td>
<td>1.65E-02</td>
<td>8.28E-02</td>
<td>5.35E-04</td>
<td>6.39E-02</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
<tr>
<td>27030</td>
<td>MLH3</td>
<td>14</td>
<td>75480467</td>
<td>75518235</td>
<td>1.69E-02</td>
<td>6.69E-01</td>
<td>1.47E-01</td>
<td>6.39E-03</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
<tr>
<td>5395</td>
<td>PMS2</td>
<td>7</td>
<td>6012870</td>
<td>6048737</td>
<td>2.58E-02</td>
<td>3.66E-01</td>
<td>8.84E-03</td>
<td>1.76E-05</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
<tr>
<td>4439</td>
<td>MSH5</td>
<td>6</td>
<td>31707725</td>
<td>31730455</td>
<td>4.35E-02</td>
<td>8.54E-01</td>
<td>7.73E-01</td>
<td>5.11E-01</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
<tr>
<td>5982</td>
<td>RFC2</td>
<td>7</td>
<td>73645832</td>
<td>73668738</td>
<td>4.80E-02</td>
<td>5.91E-01</td>
<td>2.02E-02</td>
<td>4.44E-01</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
<tr>
<td>6119</td>
<td>RPA3</td>
<td>7</td>
<td>7676575</td>
<td>7758238</td>
<td>6.55E-02</td>
<td>7.22E-01</td>
<td>9.17E-02</td>
<td>4.37E-01</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
<tr>
<td>4292</td>
<td>MLH1</td>
<td>3</td>
<td>37034841</td>
<td>37092337</td>
<td>6.98E-02</td>
<td>3.97E-04</td>
<td>1.28E-04</td>
<td>3.91E-04</td>
<td>Repair pathway/SSR/MMR</td>
</tr>
</tbody>
</table>

Table 3.13: Gene-wide p-values for the most significant genes in the two Pearl et al. pathways showing significant enrichment in TRACK (Pearl et al., 2015).
Figure 3.19: Significant genes are functionally linked and may cause somatic expansion of the HTT CAG repeat tract. STRING diagram showing all proteins from the Pearl et al (20) dataset with gene-wide p-values for association with Huntington’s disease progression < 0.02 in A: the TRACK-HD dataset and B, the meta-analysis of TRACK-HD and REGISTRY (Table 3.8). Genes with p<0.02 coloured; 10 further interactors in grey, confidence of interaction is shown in the ‘Edge confidence’ box, homo sapiens protein data used: http://string-db.org/cgi/ accessed October 2016 and January 2017 (36). (Figure prepared by me)

Since MSH3 is a member of all the most significantly enriched pathways, we tested whether MSH3 was individually responsible for the pathway enrichments by removing it and repeating the analyses. GO:32300 and KEGG:3430 are still nominally significant in TRACK (p=0.0413, p=0.0452 respectively) but not in REGISTRY. Neither of the two Pearl pathways is significant in TRACK or REGISTRY. The only pathways nominally significant both in TRACK and REGISTRY are GO:32389 (MutLalpha complex) and Pearl pathway “Repair_pathway/SSR/MMR/MutL_homologs”, neither of which contain MSH3. Thus, it appears that the mismatch repair pathway enrichments are mainly driven by MSH3. However, in the TRACK-REGISTRY meta-analysis, the Pearl et al. MMR pathway (p=1.27x10⁻⁴), GO:32300 (p=1.02x10⁻³), KEGG 3430 (1.07x10⁻⁴) and GO:30983 are at least nominally significant without MSH3. Pathway enrichments without MSH3 are shown in Table 3.14 for the 14 GeM pathways and can be found at http://hdresearch.ucl.ac.uk/data-resources/ for the Pearl et al. pathways.
<table>
<thead>
<tr>
<th>Pathway</th>
<th>p(TRACK)</th>
<th>p(TRACK no MSH3)</th>
<th>p(REGISTRY)</th>
<th>p(REGISTRY no MSH3)</th>
<th>p(META)</th>
<th>p(META no MSH3)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO: 32300</td>
<td>3.455E-09</td>
<td>0.04127</td>
<td>0.0008336</td>
<td>0.07162</td>
<td>1.13E-11</td>
<td>0.001024</td>
<td>mismatch repair complex</td>
</tr>
<tr>
<td>KEGG 3430</td>
<td>2.794E-07</td>
<td>0.04521</td>
<td>0.04795</td>
<td>0.1471</td>
<td>1.34E-16</td>
<td>0.000107</td>
<td>KEGG_MISMATCH_REPAIR</td>
</tr>
<tr>
<td>GO: 30983</td>
<td>6.661E-07</td>
<td>0.1001</td>
<td>0.0004195</td>
<td>0.009264</td>
<td>3.17E-11</td>
<td>0.000274</td>
<td>mismatched DNA binding</td>
</tr>
<tr>
<td>GO: 6298</td>
<td>0.000003533</td>
<td>0.2446</td>
<td>0.04589</td>
<td>0.1839</td>
<td>6.54E-09</td>
<td>0.0729</td>
<td>mismatch repair</td>
</tr>
<tr>
<td>GO: 32407</td>
<td>0.01818</td>
<td>0.01818</td>
<td>0.1101</td>
<td>0.1101</td>
<td>0.000640</td>
<td>0.000640</td>
<td>MutSalpha complex binding</td>
</tr>
<tr>
<td>GO: 32389</td>
<td>0.02249</td>
<td>0.02249</td>
<td>0.04688</td>
<td>0.04688</td>
<td>0.000523</td>
<td>0.000523</td>
<td>MutLalpha complex</td>
</tr>
<tr>
<td>GO: 33683</td>
<td>0.08014</td>
<td>0.08014</td>
<td>0.0005874</td>
<td>0.0005874</td>
<td>0.00675</td>
<td>0.00675</td>
<td>nucleotide-excision repair, DNA incision</td>
</tr>
<tr>
<td>GO: 90141</td>
<td>0.3318</td>
<td>0.3318</td>
<td>0.05934</td>
<td>0.05934</td>
<td>0.7872</td>
<td>0.7872</td>
<td>positive regulation of mitochondrial fission</td>
</tr>
<tr>
<td>GO: 1900063</td>
<td>0.4103</td>
<td>0.4103</td>
<td>0.7287</td>
<td>0.7287</td>
<td>0.6926</td>
<td>0.6926</td>
<td>regulation of peroxisome organization</td>
</tr>
<tr>
<td>GO: 90200</td>
<td>0.4582</td>
<td>0.4582</td>
<td>0.544</td>
<td>0.544</td>
<td>0.5280</td>
<td>0.5280</td>
<td>positive regulation of release of cytochrome c from mitochondria</td>
</tr>
</tbody>
</table>

*Table 3.14: Effect of removing MSH3 on the Setscreen enrichment \( p \)-values for the top 14 GeM pathways in TRACK-HD, REGISTRY and the TRACK-REGISTRY meta-analysis. (Only top 10 pathways shown, full table can be found at http://hdresearch.ucl.ac.uk/data-resources/)*
Setscreen gene set analysis of the large set of pathways analysed by the GeM-HD Consortium (2015) is shown in Table 3.15. There were 26 pathways showing significant ($q<0.05$) enrichment in TRACK after correction for multiple testing of pathways. These pathways mainly relate to DNA repair and binding, and none is more significant than GO:32300 (mismatch repair complex). The genes in these 26 pathways are shown in http://hdresearch.ucl.ac.uk/data-resources/, and are similar to those in Table 3.7. Thus, analysis of the large set of pathways does not appear to throw up any novel areas of biology outside those indicated by the GeM paper.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>p(TRACK)</th>
<th>q(TRACK)</th>
<th>p(REGISTRY)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO: 32300</td>
<td>3.46E-09</td>
<td>1.22E-05</td>
<td>8.34E-04</td>
<td>mismatch repair complex</td>
</tr>
<tr>
<td>GO: 43570</td>
<td>8.02E-09</td>
<td>1.41E-05</td>
<td>3.20E-03</td>
<td>maintenance of DNA repeat elements</td>
</tr>
<tr>
<td>GO: 32135</td>
<td>2.13E-08</td>
<td>2.50E-05</td>
<td>6.99E-03</td>
<td>DNA insertion or deletion binding</td>
</tr>
<tr>
<td>GO: 710</td>
<td>4.59E-08</td>
<td>4.05E-05</td>
<td>1.19E-02</td>
<td>meiotic mismatch repair</td>
</tr>
<tr>
<td>GO: 51095</td>
<td>9.01E-08</td>
<td>6.35E-05</td>
<td>1.52E-02</td>
<td>regulation of helicase activity</td>
</tr>
<tr>
<td>GO: 404</td>
<td>1.14E-07</td>
<td>6.70E-05</td>
<td>3.38E-03</td>
<td>loop DNA binding</td>
</tr>
<tr>
<td>KEGG 3430</td>
<td>4.04E-07</td>
<td>2.03E-04</td>
<td>4.80E-02</td>
<td>KEGG MISMATCH REPAIR</td>
</tr>
<tr>
<td>GO: 32138</td>
<td>5.69E-07</td>
<td>2.31E-04</td>
<td>1.28E-02</td>
<td>single base insertion or deletion binding</td>
</tr>
<tr>
<td>GO: 19237</td>
<td>5.91E-07</td>
<td>2.31E-04</td>
<td>5.97E-03</td>
<td>centromeric DNA binding</td>
</tr>
<tr>
<td>GO: 30983</td>
<td>6.66E-07</td>
<td>2.35E-04</td>
<td>4.20E-04</td>
<td>mismatched DNA binding</td>
</tr>
<tr>
<td>GO: 32142</td>
<td>9.05E-07</td>
<td>2.90E-04</td>
<td>4.43E-03</td>
<td>single guanine insertion binding</td>
</tr>
<tr>
<td>GO: 403</td>
<td>1.87E-06</td>
<td>5.14E-04</td>
<td>7.03E-03</td>
<td>Y-form DNA binding</td>
</tr>
<tr>
<td>REACTOME 1234</td>
<td>1.90E-06</td>
<td>5.14E-04</td>
<td>2.66E-02</td>
<td>REACT:TETRAHYDROBIOPTERIN (BH4) SYNTHESIS</td>
</tr>
<tr>
<td>GO: 32139</td>
<td>2.85E-06</td>
<td>7.16E-04</td>
<td>2.45E-04</td>
<td>dinucleotide insertion or deletion binding</td>
</tr>
<tr>
<td>GO: 6298</td>
<td>3.53E-06</td>
<td>8.30E-04</td>
<td>6.74E-03</td>
<td>mismatch repair</td>
</tr>
<tr>
<td>REACTOME 656</td>
<td>4.76E-06</td>
<td>1.05E-03</td>
<td>8.13E-02</td>
<td>REACT:METABOLISM OF FOLATE AND PTERINES</td>
</tr>
<tr>
<td>GO: 217</td>
<td>7.69E-06</td>
<td>1.59E-03</td>
<td>2.91E-02</td>
<td>DNA secondary structure binding</td>
</tr>
<tr>
<td>GO:</td>
<td>p-value</td>
<td>p-value</td>
<td>p-value</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>GO: 16447</td>
<td>1.13E-05</td>
<td>2.18E-03</td>
<td>3.66E-02</td>
<td>somatic recombination of immunoglobulin gene segments</td>
</tr>
<tr>
<td>GO: 51096</td>
<td>1.18E-05</td>
<td>2.18E-03</td>
<td>5.93E-03</td>
<td>positive regulation of helicase activity</td>
</tr>
<tr>
<td>GO: 16445</td>
<td>1.36E-05</td>
<td>2.39E-03</td>
<td>4.55E-02</td>
<td>somatic diversification of immunoglobulins</td>
</tr>
<tr>
<td>REACTOME 452</td>
<td>3.99E-05</td>
<td>6.70E-03</td>
<td>9.11E-02</td>
<td>REACT: G1 S- SPECIFIC TRANSCRIPTION</td>
</tr>
<tr>
<td>GO: 45910</td>
<td>5.44E-05</td>
<td>8.71E-03</td>
<td>5.43E-03</td>
<td>negative regulation of DNA recombination</td>
</tr>
<tr>
<td>KEGG 790</td>
<td>6.00E-05</td>
<td>9.19E-03</td>
<td>1.93E-02</td>
<td>KEGG FOLATE BIOSYNTHESIS</td>
</tr>
<tr>
<td>GO: 16444</td>
<td>7.07E-05</td>
<td>9.96E-03</td>
<td>8.28E-02</td>
<td>somatic cell DNA recombination</td>
</tr>
<tr>
<td>GO: 2562</td>
<td>7.07E-05</td>
<td>9.96E-03</td>
<td>8.28E-02</td>
<td>somatic diversification of immune receptors via germline recombination within a single locus</td>
</tr>
<tr>
<td>GO: 2200</td>
<td>1.15E-04</td>
<td>1.55E-02</td>
<td>1.02E-01</td>
<td>somatic diversification of immune receptors</td>
</tr>
<tr>
<td>REACTOME 659</td>
<td>7.20E-04</td>
<td>9.07E-02</td>
<td>1.16E-01</td>
<td>REACT: METABOLISM OF NITRIC OXIDE</td>
</tr>
<tr>
<td>REACTOME 367</td>
<td>7.20E-04</td>
<td>9.07E-02</td>
<td>1.16E-01</td>
<td>REACT: ENOS ACTIVATION AND REGULATION</td>
</tr>
<tr>
<td>GO: 35825</td>
<td>8.81E-04</td>
<td>1.03E-01</td>
<td>7.04E-02</td>
<td>reciprocal DNA recombination</td>
</tr>
<tr>
<td>GO: 7131</td>
<td>8.81E-04</td>
<td>1.03E-01</td>
<td>7.04E-02</td>
<td>reciprocal meiotic recombination</td>
</tr>
</tbody>
</table>

**Table 3.15:** Setscreen enrichment p-values for the large set of GeM pathways in TRACK-HD and REGISTRY.

Top 30 pathways shown; the full table can be found at http://hdresearch.ucl.ac.uk/data-resources/.

### 3.4 Discussion

The evidence from the work presented in this chapter suggests that *MSH3* is likely to be a modifier of disease progression in Huntington’s disease. With collaborators, I undertook an unbiased genetic screen using a novel disease progression measure in the TRACK-HD study, and identified a significant locus on chromosome 5, which encompasses three genes:
**MTRNR2L2**, **MSH3** and **DHFR** (Hensman Moss et al., 2017b). This locus replicated in an independent group of subjects from the European Disease Huntington’s Disease Network REGISTRY study using a parallel disease progression measure, and was genome-wide significant in a meta-analysis of the two studies \( (p=1.12\times10^{-10}) \) (Hensman Moss et al., 2017b). The lead SNP in TRACK-HD, rs557874766, is a coding variant in MSH3 \((5.80 \times 10^{-08})\), and it is classed of moderate impact. Furthermore, eQTL analyses show association of lower MSH3 expression with slower disease progression.

Genetic modifiers of disease in people highlight pathways for therapeutic development; any pathway containing genetic variation that ameliorates or exacerbates disease forms a pre-validated relevant target. The proportion of drug mechanisms with direct genetic support increases significantly across the drug development pipeline from 2.0% at the preclinical stage, to 8.2% among mechanisms for approved drugs (Nelson et al., 2015), suggesting that genetic data may be valuable in highlighting drugs that will successful. The classic example is the target for statins, **HMGCR**, which has been associated with serum cholesterol level (Kathiresan et al., 2009), though there are increasing examples, particularly in musculoskeletal and metabolic disease of therapeutic targets being identified through genetic analysis (Nelson et al., 2015). The correlation between successful drug targets and underlying genetic evidence may be because genes that result in notable phenotypic changes when altered genetically are also the most responsive to drug-induced alterations.

The classical case-control design to examine complex disease has yielded multiple genetic associations highlighting relevant biology for novel treatment design (Plenge et al., 2013), however studies of potential genetic modifiers in genetically simple Mendelian diseases have been difficult to conduct. The diseases are rare and show gene and locus heterogeneity, thus finding genuine modifying associations in such a noisy background is inherently difficult. However, variants that modify disease in the context of a Mendelian causative gene may not be under negative selection pressure in the general population, thus may be relatively common. Recent successful identifications of modifiers have been made in specific genetic subtypes of disease (Trinh et al., 2016) or in relatively large samples with consistent clinical data (Corvol et al., 2015, GeM-HD-Consortium, 2015).

One way to increase the power of genetic studies is to obtain a more accurate measure of phenotype (Sham and Purcell, 2014). Prospective multivariate longitudinal measures such as those collected in TRACK-HD are ideal (Sham and Purcell, 2014). Our analysis of Huntington’s disease progression showed that motor, cognitive and brain imaging variables typically
progress in parallel and that patterns of loss are not sufficiently distinct to be considered sub-phenotypes for genetic analysis. The first psychiatric PC has notably lower correlation with motor and cognitive domains and CPO variables, suggesting that psychiatric symptoms showed a different trajectory. This may be because the data were less quantitative, and that psychiatric symptoms of Huntington’s disease are relatively amenable to treatment which may be started during the course of the study, making progression analysis problematic. We therefore developed a single progression measure excluding the psychiatric data. It would be interesting to explore whether genes and pathways linked to psychiatric diseases more broadly also influence the psychiatric manifestation of HD.

We found that AAO was correlated with the unified progression measure but did not explain the genetic associations observed with progression (Hensman Moss et al., 2017b). Thus, progression seems to be measuring a different aspect of disease to AAO, or a similar aspect of disease, but with greater precision. The latter option seems more plausible given that AAO itself reflects disease trajectory over a subject’s lifespan up till disease onset (Figure 3.9). The data available in REGISTRY are less comprehensive; therefore we used a different approach by comparing cross-sectional severity at the most recent visit with that expected based on age and CAG. The unified progression measures in TRACK-HD and REGISTRY are correlated and again, the genetic associations in REGISTRY are not completely driven by AAO, demonstrating the utility of retrospective composite progression scores in genetic analysis. Prognostic indices for motor onset have been developed (Long et al., 2017), and the development of progression scores for prospective use, for example to empower drug trials by stratifying patients by predicted rate of progression warrants further attention.

The work described in this chapter has a number of limitations. TRACK-HD has the same standardised detailed phenotypic information on nearly all participants, but in only 243 HD gene mutation carrying subjects. The REGISTRY study is much larger but the phenotypic data are less complete (Table 3.16), often not collected at regular intervals and not on everyone in the study, and in multiple centres which will inevitably lead to intrinsic variation. Medications, particularly starting or changing the dose of neuroleptics may impact subject ratings, however medication use was not included in the progression analysis models due to prior work which showed little convincing evidence of causal effects of medication on clinical performance once confounding factors were controlled for in the TRACK-HD cohort (Keogh et al., 2016). However this would be worth further analysis if larger cohorts become available in the future. Subjects with comorbid neurological disease were excluded from TRACK-HD, and the presence of other comorbidities was not included in the progression analysis.
Nevertheless, the progression measures show the expected relationship with change in TMS and TFC in both TRACK-HD and REGISTRY indicating their clinical relevance. However, future development of the progression statistic and confirmation of the genetic association in subjects from ongoing large studies such as ENROLL (Landwehrmeyer et al., 2016), with data collected more systematically than in REGISTRY but in less detail than TRACK-HD, would be ideal.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Missing Values</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Count</td>
<td>Percent</td>
</tr>
<tr>
<td>Motor</td>
<td>1744</td>
<td>91</td>
<td>4.96</td>
</tr>
<tr>
<td>Verbal Fluency</td>
<td>1145</td>
<td>690</td>
<td>37.6</td>
</tr>
<tr>
<td>Stroop Color</td>
<td>1052</td>
<td>783</td>
<td>42.67</td>
</tr>
<tr>
<td>Stroop Color</td>
<td>1116</td>
<td>719</td>
<td>39.18</td>
</tr>
<tr>
<td>Stroop Word</td>
<td>1104</td>
<td>731</td>
<td>39.84</td>
</tr>
<tr>
<td>Stroop Interference</td>
<td>1092</td>
<td>743</td>
<td>40.49</td>
</tr>
<tr>
<td>TFC</td>
<td>1758</td>
<td>77</td>
<td>4.2</td>
</tr>
<tr>
<td>FAS score</td>
<td>1616</td>
<td>219</td>
<td>11.93</td>
</tr>
</tbody>
</table>

*Table 3.16: Summary of missing data in REGISTRY*

The genetic locus identified by the unified TRACK-HD progression measure association includes three genes, but *MSH3* is the likeliest candidate. Firstly, the lead SNP is a coding variant in exon 1 of *MSH3*, *MSH3* Pro67Ala, with the potential to affect function (SNiPA (Arnold et al., 2015) accessed 10/11/2016). Clinically, each copy of the minor allele (G) at this SNP corresponds to a decrease of approximately 0.4 (95% Cl=0.16,0.66) units per year in the rate of change of TMS, and a reduction of approximately 0.12 (95% Cl=0.06,0.18) units per year in the rate of change of TFC. Secondly, *MSH3* has been extensively implicated in the pathogenesis of HD in both mouse and cell studies, though this is the first human study to link *MSH3* to HD.

*MSH3* is a neuronally expressed member of a family of DNA mismatch repair (MMR) proteins (Gonitel et al., 2008); the proteins that mediate the MMR pathway are highly conserved from bacteria through to humans, though there are also some features which are unique to higher eukaryotes (Modrich, 2006, Larrea et al., 2010). While DNA repair more broadly was highlighted by our pathway analysis, MMR in particular which was associated with progression in HD (see also General Introduction, Chapter 1) (*Tables 3.14 and 3.15*).
Much of the work on mismatch repair has been done in prokaryotes. In the prokaryotic model system *E. coli*, MutS binds as an asymmetric clamp to DNA containing the mismatch, then the MutL homodimer couples MutS recognition to distinguishing the template and nascent DNA strands (Larrea et al., 2010). In eukaryotes there are several different MutS and MutL homologs with different specificities. MSH3 is a MutS homolog which forms a heteromeric complex with MSH2 to form MutSβ, this recognises insertion-deletion loops of up to 13 nucleotides (Figure 3.20). MSH2 also forms a complex with MSH6 to form the MutSα complex which repairs mispaired bases and smaller mispaired loops (Figure 3.20). The MutL heterodimer is also present in a number of forms, including the MutLα complex, which is made up of MLH1 and PMS2 proteins, the MutLβ heterodimer (MLH1 and PMS1), and MutLγ (MLH1 and MLH3). Of these MutLα has the primary role in mismatch correction (Martin et al., 2010).

**Figure 3.20:** Schematic of DNA damage recognized by the MMR pathway. **A**, the MutSα (MSH2/MSH6) heterodimer recognizes base-base mismatches and **B**, small insertion-deletion loops (IDL). The MutSβ (MSH2/MSH3) heterodimer recognizes single nucleotide IDLs and **C**, longer IDLs (10-nucleotide loops). In association with the MutL heterodimer and other associated proteins, these mismatches are excised and repaired. Figure from (Martin et al., 2010), image reproduced with permission of the rights holder, AACR Journals.
As discussed in the General Introduction, Chapter 1, somatic instability of CAG repeats occurs in various repeat expansion disorders. Variants in DNA repair pathways contribute to age of onset modification of multiple CAG repeat expansion diseases as will be discussed in Chapter 4 (Bettencourt et al., 2016) implicating the CAG repeat itself as the source of modification in these diseases rather than a factor specific to huntingtin.

The reason for the tissue specificity of somatic expansion is not clear. While several groups found that stoichiometric levels of repair proteins are associated with variable levels of CAG instability between the striatum and cortex of HD mice, another study of 14 different mouse tissue types revealed widely varying levels of MMR proteins between tissues and no clear correlation with CAG expansion levels (Tome et al., 2013b). In human embryonic stem cell lines derived from oocytes and sperm of DM1 and HD patients, somatic instability is seen and correlates with expression of MMR proteins (Du et al., 2013). In this stem cell system, the overall tendency of triplet repeats to expand ceased on differentiation into differentiated embryoid body or neurospheres (Du et al., 2013). It will be critical to determine whether somatic instability occurs predominantly during development, or throughout the lifespan as suggested by murine model systems (Gonitel et al., 2008), as this has implications for the likely mechanism and whether the pathway would be amenable to therapeutic manipulation. Abnormal secondary structures including hairpins and G-quadruplexes have also been linked to G-rich sequences, and are associated with stalled replication forks (Mirkin, 2013). It has been proposed that formation of unusual and non-B-form DNA structures by CAG trinucleotide repeats underlies the phenomenon of repeat expansion, but the molecular basis for expansion, either through the germ-line or in somatic cells remains poorly understood (Liu and Wilson, 2012, Mirkin, 2007). Liu and Wilson (Liu and Wilson, 2012) suggest a role for oxidative damage in the base excision repair pathway in TNR expansion through the generation of gaps and hairpin structures. TNRs hairpin structures are stabilized by mismatch repair MSH2/MSH3 complexes (Owen et al., 2005), and MSH2/MSH3 interfere with flap processing to produce small incremental expansion events (Kantartzis et al., 2012).

The importance of Msh3 was shown in 2002 in a myotonic dystrophy mouse model where Msh 3 deficient background abolished is CTG repeat instability (van den Broek et al., 2002). It has subsequently been demonstrated that Msh3 is required for both somatic expansion of HTT CAG repeats and for enhancing an early disease phenotype in mouse striatum (Dragileva et al., 2009), and expansion of CAG and CTG repeats is prevented by msh3Δ in Saccharomyces cerevisiae (Williams and Surtees, 2015a). These data suggest a plausible mechanism, via effect on CAG somatic expansion, through which variation in MSH3 could operate in HD.
(Figure 3.21A and B). In patients with DM1 an *MSH3* variant was recently associated with somatic instability in blood DNA of patients (Morales et al., 2016).
Figure 3.21: A Schematic diagram showing how DNA mismatch repair proteins may be involved in somatic expansion of the CAG tract. Proteins with p<0.01 in the meta-analysed progression GWAS are coloured red. (i) The CAG repeat DNA is partly unwound by lesions, constraints of the CAG tract structure or by transcription. (ii) This unwound DNA is recognised by MutSbeta (MSH2/MSH3) which recruits the endonuclease MutLalpha (PMS2/MLH1) and cleaves the DNA. (iii) Repair of the strand break leads to erroneous expansion of the CAG repeat. In neurones of the striatum somatic expansion is an ongoing process that occurs throughout life and variants in MSH3 may promote or inhibit repeat recognition, binding or repair. B Potential link between degree of somatic expansion over a patient’s lifespan and rate of Huntington’s disease progression. (Figures from (Hensman Moss et al., 2017b), and made by me).
As touched upon above, other proteins have been linked to trinucleotide repeat instability.\textit{MSH2} in particular has been shown to be essential for somatic expansion of the CAG repeat in HTT to occur in model systems (Manley et al., 1999, Wheeler et al., 2003, Kovalenko et al., 2012). \textit{MSH2} forms a heterodimeric complex with \textit{MSH3} or \textit{MSH6} (Figure 3.21). We did not find any association between \textit{MSH2} variants and progression in HD (gene-wide p-value 0.7034 in the meta-analysis (http://hdresearch.ucl.ac.uk/data-resources/)). It may be that \textit{MSH2}'s association with both colonic and extra-colonic malignancy (Martin et al., 2010) mean that variants within it that would influence trinucleotide repeat instability are also selected against due to their oncogenic potential.

In the study described in this Chapter the proteins of the MutL complex were also highlighted as nominally significant, MLH1 had a gene-wide p-value of $1.28 \times 10^{-4}$ in the meta-analysis; PMS2 had a gene-wide p-value of $8.84 \times 10^{-3}$ in the meta-analysis; by contrast MLH3 had a gene-wide p-value of $1.47 \times 10^{-2}$ in the meta-analysis (Table 3.7) (Hensman Moss et al., 2017b). As noted above, MLH1 and PMS2 are MutL homologs, together they form a MutL\textalpha heterodimer which associates with the MutS\textalpha or MutS\textbeta complex after the MutS complex has bound the mismatched DNA (Cannavo et al., 2007). MLH1 can bind two other human MutL homologues, PMS1 and MLH3, to form the heterodimers MutL\textbeta and MutL\textgamma, respectively (Cannavo et al., 2007). Mutations in MLH1 and PMS2 predispose to a range of tumorigenic conditions, including hereditary nonpolyposis colon cancer (Martin et al., 2010). MLH1 was previously implicated in modifying onset in HD: in the GeM GWAS a SNP tagged to MLH1 approached significance (p= $2.2 \times 10^{-7}$) (GeM-HD-Consortium, 2015), and was significant in replication by Lee \textit{et al} (Lee et al., 2017). It has also been implicated in model systems: the mouse homolog, Mlh1, was highlighted in a genome-wide genetic screen to modify somatic instability of the CAG repeat and the timing of CAG length-dependent phenotypes in the striatum of genetic HD replica CAG knock-in mice (Pinto et al., 2013).

Other proteins which play a part in mismatch repair are also implicated in the study described here. LIG1, which had a gene-wide p-value of $5.35 \times 10^{-4}$ in the meta-analysis, ligates nicked DNA fragments following replication and/or repair (Schmidt and Pearson, 2016) and is involved in the repair of slipped strand DNA intermediates (Mason et al., 2014). CAG/CTG repeat instability is modulated by the levels of \textit{LIG1}, and its interaction with \textit{PCNA} (Lopez Castel et al., 2009). \textit{POLD2}, which encodes the catalytic subunit of DNA polymerase delta, had a gene-wide p-value of $2.75 \times 10^{-4}$ in the meta-analysis (Hensman Moss et al., 2017b).
Interestingly, FAN1, which was associated with the two most significant signals in the GeM AAO GWAS (GeM-HD-Consortium, 2015), was just highlighted by our REGISTRY GWAS (p=2.35x10^-6). The absence of signal in TRACK-HD may be due to the lower sample size: the MAF of the index SNP in GeM is 1.1% in European populations which would be hard to pick up in a sample of 216. The second signal on chromosome 15 in GeM had a MAF of 30.2% in European populations, though the effect size at this locus was lower (GeM-HD-Consortium, 2015). A more extensive discussion about the role of FAN1 is in Chapter 5.

This chapter describes the first study to use a measure of progression to look for modifiers of a neurodegenerative Mendelian disorder. We detected association with a coding variant on chromosome 5, reaching genome-wide significance given its annotation (Sveinbjornsson et al., 2016) in just 216 subjects, which replicated and strengthened in a larger independent sample and strengthened on meta-analysis. This indicates that either our progression measure developed in TRACK-HD is an excellent reflection of disease pathophysiological progression or that this is a locus with a very large effect size, or, most likely, both. While there are three genes at the locus, the most significant variant gives a coding change in MSH3, which together with the prior biological evidence makes it the most likely candidate. Somatic expansion of the CAG repeat through alterations in MSH3 is a plausible mechanism for pathogenesis in HD which can be followed up in functional experiments in HD models. These data provide additional support for the therapeutic targeting of Huntingtin and the stability of its CAG repeat. While variants in or loss of the mismatch repair proteins MSH2, MLH1, PMS2 and MSH6, predispose to a range of tumorigenic conditions, including hereditary nonpolyposis colon cancer, also known as Lynch syndrome, MSH3 has not been generally linked to malignancy. Furthermore MSH3 is not essential given that it can tolerate loss of function variation (Lek et al., 2016) (suggesting that it is not constrained by selection pressures). These factors make it an attractive therapeutic target in HD.
Chapter 4: DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases

4.1 Introduction

The polyglutamine repeat disorders include the SCAs and HD as discussed in Chapter 1, other neurodegenerative disorders also caused by CAG repeat expansions are Dentatorubral-pallidoluysian atrophy (DRPLA) and spinal and bulbar muscular atrophy (SBMA). These repeat disorders, whose clinical features are summarised in Table 4.1 form the focus of this chapter.
<table>
<thead>
<tr>
<th>Repeat disorder</th>
<th>Gene</th>
<th>Prevalence (per 100,000 European population)</th>
<th>Phenotype</th>
<th>Variance in AAO explained by repeat length (heritability of residual variability)</th>
<th>Normal range</th>
<th>Pathogenic range</th>
<th>Somatic instability</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD</td>
<td>HTT</td>
<td>3-10 (Wood, 2012, Rawlins, 2010b, Bates et al., 2014, Warby et al., 2014)</td>
<td>Involuntary movements, cognitive impairment</td>
<td>50-60% (Gusella et al., 2014, Persichetti et al., 1994, Snell et al., 1993, Andrew et al., 1993, Duyao et al., 1993) (40-60%) (Djousse et al., 2003, Wexler et al., 2004a)</td>
<td>6-35</td>
<td>40-121</td>
<td>Yes</td>
</tr>
<tr>
<td>SCA1</td>
<td>ATXN1</td>
<td>0.16 (Durr, 2010)</td>
<td>Ataxia, ophthalmoplegia, pyramidal and extrapyramidal features (Subramony, 2012)</td>
<td>64-76% (Ranum et al., 1994, Tezenas du Montcel et al., 2014b, Globas et al., 2008, van de Warrenburg et al., 2002, van de Warrenburg et al., 2005) (no significant heritable component) (van de Warrenburg et al., 2005))</td>
<td>6-38</td>
<td>45-83</td>
<td>Yes (Hashida et al., 1997)</td>
</tr>
<tr>
<td>SCA</td>
<td>Gene</td>
<td>Mutation Size</td>
<td>Literature Support</td>
<td>Clinical Manifestations</td>
<td>Frequency</td>
<td>Age at Onset</td>
<td>Age at Death</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>---------------</td>
<td>-------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>SCA3</td>
<td>ATXN3</td>
<td>0.4 (Durr, 2010)</td>
<td>Ataxia, pyramidal signs, neuropathy, extrapyramidal features, ophthalmoplegia (Durr et al., 1996)</td>
<td>45-80% (Bettencourt and Lima, 2011, Durr et al., 1996) (46%) (van de Warrenburg et al., 2005)</td>
<td>12-44</td>
<td>52-87</td>
<td>Yes (Hashida et al., 1997)</td>
</tr>
<tr>
<td>SCA6</td>
<td>CACNA1A</td>
<td>0.04 (Durr, 2010)</td>
<td>Ataxia (Rub et al., 2013)</td>
<td>26-52% (Tezenas du Montcel et al., 2014a, van de Warrenburg et al., 2005) (no significant heritable component) (van de Warrenburg et al., 2005)</td>
<td>4-18</td>
<td>20-33</td>
<td>Unknown</td>
</tr>
<tr>
<td>Disease</td>
<td>Gene</td>
<td>Repeat Range</td>
<td>Clinical Features</td>
<td>CAG Repeat Ranges</td>
<td>Epidemiology</td>
<td>Clinical Features</td>
<td>CAG Repeat Ranges</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>--------------</td>
<td>------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>SCA7</td>
<td>ATXN7</td>
<td>0.12 (Durr, 2010)</td>
<td>Ataxia, macular degeneration, ophthalmoplegia, pyramidal and extrapyramidal features (Sokolovsky et al., 2010)</td>
<td>71-88% (Rub et al., 2013, Tezenas du Montcel et al., 2014a) (no significant heritable component) (van de Warrenburg et al., 2005)</td>
<td>3-19</td>
<td>37-460</td>
<td>Yes (Trang et al., 2015)</td>
</tr>
<tr>
<td>SCA17</td>
<td>TBP</td>
<td>&lt;0.02 (Durr, 2010)</td>
<td>Ataxia, pyramidal signs, dementia, seizures, extrapyramidal features (Schneider et al., 2006, Rolfs et al., 2003)</td>
<td>Unknown</td>
<td>25-40</td>
<td>49-66</td>
<td>Unknown</td>
</tr>
<tr>
<td>SCA12</td>
<td>PPP2R2</td>
<td>&lt;0.02</td>
<td>Ataxia, tremor, neuropathy</td>
<td>Unknown</td>
<td>4-32</td>
<td>40-78</td>
<td>Yes</td>
</tr>
<tr>
<td>DRPLA</td>
<td>ATN1</td>
<td>0.005-0.04</td>
<td>Myoclonus, epilepsy, ataxia, dementia (Tsuji, 1999)</td>
<td>50-68% (Wardle et al., 2009, Potter, 1996)</td>
<td>6-35</td>
<td>48-93</td>
<td>Yes (Hashida et al., 1997)</td>
</tr>
<tr>
<td>SBMA</td>
<td>AR</td>
<td>0.65-2 (Udd et al., 1998, Spada, 2014)</td>
<td>Limb and bulbar weakness, neuropathy, endocrine features (Atsuta et al., 2006, Kennedy et al., 1968, Rhodes et al., 2009)</td>
<td>29% (Sinnreich et al., 2004)</td>
<td>9-34</td>
<td>38-72</td>
<td>Yes (Tanaka et al., 1999)</td>
</tr>
</tbody>
</table>

Table 4.1: Characteristics of the polyglutamine diseases showing epidemiology, clinical features, and CAG repeat ranges of polyglutamine diseases. HD: Huntington’s disease (MIM #143100), SCA1: spinocerebellar ataxia 1 (MIM #164400), SCA2: spinocerebellar ataxia 2(SCA2, MIM #183090), SCA3: spinocerebellar ataxia 3 (MIM #109150; also known as Machado-Joseph disease (MJD)), SCA6: spinocerebellar ataxia 6 (MIM#183086), SCA7:
spinocerebellar ataxia 7 (MIM #164500), SCA12: spinocerebellar ataxia 12 (MIM #604326), SCA17: spinocerebellar ataxia 17 (MIM #607136), dentatorubral-pallidoluysian atrophy (DRPLA, MIM #125370) and spinal and bulbar muscular atrophy (SBMA, MIM #313200).
Longer CAG repeat tracts lead to earlier age at onset (AAO) in the polyglutamine diseases though the exact relationship between repeat length and AAO varies between diseases (Table 4.1) (Tezenas du Montcel et al., 2014b, Wexler et al., 2004a). Not all of the difference in age at onset is accounted for by CAG repeat length, and in Huntington’s disease (HD) (Wexler et al., 2004a) and spinocerebellar ataxia (SCA) types 2 and 3 (van de Warrenburg et al., 2005) it has been established that a substantial portion of this residual variance is heritable, suggesting the existence of additional modifying factors within the genome. It is likely, though remains to be established, whether there is a residual heritability for the other conditions. The Genetic Modifiers of Huntington’s Disease (GeM-HD) genome-wide association study (GWAS) (discussed in Chapter 1)(GeM-HD-Consortium, 2015) found two genome-wide loci associated with age of motor onset on chromosomes 15 and 8, with two independent signals at the same locus on chromosome 15 and a significant association with variants in DNA repair pathways. There are few known candidate modifiers of the spinocerebellar ataxias (Bettencourt et al., 2011, van de Warrenburg et al., 2005, Tezenas du Montcel et al., 2014b), and no GWAS have been reported.

Genetic anticipation, whereby successive generations become symptomatic at a younger age, occurs in the polyglutamine repeat diseases because the repeats are meiotically unstable, and tend to expand over successive generations (Hughes and Jones, 2014). Most of these conditions also show tissue-specific instability of repeat length in the somatic tissues (somatic instability) (Lopez Castel et al., 2010) (Table 4.1). In HD somatic instability is expansion-biased and age-dependent, with larger tracts more susceptible to expansion (Iyer et al., 2015, Gomes-Pereira and Monckton, 2006). It occurs in post-mitotic neurons and is prominent in striatum and cortex, tissues which are particularly affected in HD (Gonitel et al., 2008). Expansion of the repeat is ameliorated if the repeated sequence is interrupted by other codons (Jones et al., 2017, Choudhry et al., 2001, Calabresi et al., 2001). Somatic instability has been linked to disease onset and progression in both human (Swami et al., 2009) and mouse HD-studies (Dragileva et al., 2009) and decreasing somatic expansion in HD model mice delays phenotype progression (Budworth et al., 2015). Many of the principles of somatic instability in HD extend to SCAs (McMurray, 2010, Lopez Castel et al., 2010). Somatic instability (Mason et al., 2014, Pearson et al., 2005, Gomes-Pereira and Monckton, 2006) has been attributed to the actions of DNA repair proteins, as discussed in the General Introduction, in addition to individually significant variants, the GeM-HD GWAS found significant association between age at motor onset and several DNA repair pathways overall (GeM-HD-Consortium, 2015). These GeM-HD GWAS findings, along with evidence for somatic
instability in other polyglutamine diseases (Table 4.1), led to the hypothesis that variants in DNA repair genes have a universal effect modifying AAO in all polyglutamine diseases.

There are currently no disease-modifying treatments for these devastating conditions, and particularly given that many are extremely rare making their study challenging, a pharmacological approach which could be applied across the diseases is very desirable. The work described in this chapter was a collaborative project led by researchers from Cardiff and UCL, with important collaborators elsewhere. I was involved in the project from inception, taking part in discussions around study design, obtaining phenotypic data from the clinical notes from University College London Hospital/ National Hospital for Neurology and Neurosurgery subjects, collating spreadsheets and writing the manuscript of the resultant paper which was published in Annals of Neurology (Bettencourt et al., 2016). I am a co-first author on this paper.

4.2 Materials and Methods

4.2.1 Cohort
We collaborated with a multinational group of investigators to assemble an independent cohort of subjects with Huntington’s disease (HD) and the spinocerebellar ataxias (SCAs) types 1, 2, 3, 6, 7, and 17 (Table 4.2). Subject cohorts were gathered from the Neurogenetics Unit of the National Hospital for Neurology and Neurosurgery (NHNN) (London, UK), TRACK-HD (Europe)(Tabrizi et al., 2013b), SPATAK network (France), the University of Athens Medical School/Eginition Hospital (Athens, Greece), the National Institute of Neurology and Neurosurgery, Manuel Velasco Suarez (Mexico), and the University of Azores (Ponta Delgada, Portugal)(Table 4.2). Of these, I helped clinically phenotype the NHNN samples from their patient records, and with Professor Sarah Tabrizi contributed TRACK-HD samples to the analysis. While we also collected subjects with DRPLA and SBMA, very few samples were available to us so these diseases were not included in the analysis. 182 subjects with C9orf72 expansion mutations were collected, but were excluded on the basis of them not having sizing data of the expansion.

1699 subjects with HD, and SCA1, 2, 3, 6, 7, 17 and DRPLA were genotyped, of which age at onset (AAO) and CAG repeat size was available for 1462 who were used in the analysis (Table 4.2). Given the varied phenotypes of polyglutamine diseases, motor onset (HD) or the onset of the first progressive symptom as reported by the patient was used to determine AAO.
throughout all cohorts. Given the small number of patients, SCA17 was only considered in the combined SCA analysis.
<table>
<thead>
<tr>
<th>Cohort</th>
<th>Disease</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Gender</th>
<th></th>
<th></th>
<th>Ethnicity</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HD</td>
<td>SCA1</td>
<td>SCA2</td>
<td>SCA3</td>
<td>SCA6</td>
<td>SCA7</td>
<td>SCA17</td>
<td>Total</td>
<td>Male</td>
<td>Female</td>
<td>Unavailable</td>
<td>Caucasian</td>
<td>Other</td>
<td>Unavailable</td>
<td></td>
</tr>
<tr>
<td>Athens, Greece</td>
<td>351</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>351</td>
<td>174</td>
<td>177</td>
<td>0</td>
<td>351</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Azores, Portugal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>91</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>91</td>
<td>48</td>
<td>43</td>
<td>0</td>
<td>91</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>London, UK</td>
<td>0</td>
<td>30</td>
<td>66</td>
<td>45</td>
<td>69</td>
<td>7</td>
<td>1</td>
<td>218</td>
<td>103</td>
<td>82</td>
<td>33</td>
<td>109</td>
<td>72</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>Mexico</td>
<td>0</td>
<td>0</td>
<td>113</td>
<td>0</td>
<td>0</td>
<td>66</td>
<td>6</td>
<td>185</td>
<td>91</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>185</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Paris, France</td>
<td>0</td>
<td>147</td>
<td>115</td>
<td>261</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>523</td>
<td>279</td>
<td>244</td>
<td>0</td>
<td>463</td>
<td>42</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>TRACK-HD, Europe</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>94</td>
<td>46</td>
<td>48</td>
<td>0</td>
<td>90</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>445</td>
<td>177</td>
<td>294</td>
<td>397</td>
<td>69</td>
<td>73</td>
<td>7</td>
<td>1462</td>
<td>741</td>
<td>688</td>
<td>33</td>
<td>1104</td>
<td>303</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

| Mean AAO ± SD (range) | 45 ± 12.1 (6-82) | 37 ± 10.5 (16-65) | 33 ± 12.9 (8-73) | 39 ± 11.6 (9-74) | 57 ± 10.5 (18-76) | 35 ± 17.6 (5-84) | 30 ± 13.4 (8-44) |
| Mean (CAG)n length ± SD (range) | 44 ± 5.0 (37-92) | 48 ± 5.3 (39-66) | 42 ± 4.5 (33-64) | 71 ± 4.4 (50-82) | 22 ± 0.9 (21-26) | 48 ± 11.1 (36-100) | 51 ± 6.4 (42-58) |

Table 4.2: Cohort characteristics: HD – Huntington’s disease; SCA – spinocerebellar ataxia; AAO – age at onset; SD – standard deviation.
4.2.2 Selection of SNPs

SNPs for genotyping were selected from the most significant genes (gene-wide p<0.1) in the “DNA repair pathway cluster” from the GeM-HD analysis (GeM-HD-Consortium, 2015). We also included SNPs from the genome-wide significant chromosome 8 locus comprising \textit{RRM2B} and \textit{UBR5}, both members of GO:6281 “DNA Repair”. These were nominally significant in GeM, but did not reach q<0.05 and were therefore not used to create the pathway cluster, but both lie within a genome-wide significant association peak in GeM-HD, and both have significant gene-wide p-values (see Table S5 of the GeM-HD paper (GeM-HD-Consortium, 2015)).

Specifically, within the DNA repair cluster, we genotyped SNPs from members of the DNA mismatch repair pathway GO: 32300, along with LIG1 which was included in the KEGG MISMATCH REPAIR pathway and which had a low p-value in GeM (best p= 0.00559) (Table 4.3). We were unable to design a successful assay for MSH2, a member of the mismatch repair pathway. SNPs were also selected from the two most significant genes in the GO:33683 (nucleotide-excision repair, DNA incision) pathway: FAN1 and ERCC3. For each gene, the most significant SNP was selected, along with a small number of proxy SNPs in close LD ($r^2>0.8$) with the most significant SNP that also showed association in GeM-HD. Where possible, these proxy SNPs were chosen to have functional annotation (http://browser.1000genomes.org/index.html: accessed 12/6/14). If a gene contained two independent significant signals in GeM-HD (for example, \textit{FAN1}), then the lead SNP for the second signal was included. Note that this selection procedure is not intended to give comprehensive coverage of the genes in question, but instead to highlight SNPs likely to be disease relevant in the context of finite resources. To guard against the effects of population stratification, SNPs were removed from the analysis if they had a Hardy-Weinberg p-value <0.001 in the whole dataset. These procedures yielded 22 genotyped SNPs with success rates ranging from 94.2-98%, as described in Table 4.3.
<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Chr: position (bp) (GRCh37/hg19)</th>
<th>Gene symbol</th>
<th>Functional annotation</th>
<th>P (GeM-HD)</th>
<th>MAF*</th>
<th>Genotype call rate*</th>
<th>P (HWE)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1800937</td>
<td>2:48025764</td>
<td>MSH6</td>
<td>Stop gained</td>
<td>4.30E-03</td>
<td>0.074</td>
<td>0.973</td>
<td>0.840</td>
</tr>
<tr>
<td>rs4150407</td>
<td>2:128049631</td>
<td>ERCC3</td>
<td>Intron variant</td>
<td>4.60E-04</td>
<td>0.479</td>
<td>0.964</td>
<td>0.003</td>
</tr>
<tr>
<td>rs5742933</td>
<td>2:190649316</td>
<td>PMS1</td>
<td>NMD transcript variant</td>
<td>9.49E-04</td>
<td>0.205</td>
<td>0.972</td>
<td>1.000</td>
</tr>
<tr>
<td>rs1799977</td>
<td>3:37053568</td>
<td>MLH1</td>
<td>Missense variant</td>
<td>7.16E-07</td>
<td>0.28</td>
<td>0.966</td>
<td>0.354</td>
</tr>
<tr>
<td>rs6151792</td>
<td>5:80056961</td>
<td>MSH3</td>
<td>Intron variant</td>
<td>2.09E-04</td>
<td>0.117</td>
<td>0.978</td>
<td>0.706</td>
</tr>
<tr>
<td>rs115109737</td>
<td>5:80102444</td>
<td>MSH3</td>
<td>Intron variant</td>
<td>4.50E-04</td>
<td>0.041</td>
<td>0.980</td>
<td>0.489</td>
</tr>
<tr>
<td>rs71636247</td>
<td>5:80118976</td>
<td>MSH3</td>
<td>Intron variant</td>
<td>2.55E-04</td>
<td>0.034</td>
<td>0.976</td>
<td>1.000</td>
</tr>
<tr>
<td>rs1805323</td>
<td>7:6026942</td>
<td>PMS2</td>
<td>Missense variant</td>
<td>3.04E-02</td>
<td>0.043</td>
<td>0.975</td>
<td>0.736</td>
</tr>
<tr>
<td>rs12531179</td>
<td>7:6028687</td>
<td>PMS2</td>
<td>Intron variant</td>
<td>3.84E-05</td>
<td>0.169</td>
<td>0.971</td>
<td>0.925</td>
</tr>
<tr>
<td>rs3735721</td>
<td>8:103217695</td>
<td>RRM2B</td>
<td>3' _UTR_variant</td>
<td>5.68E-07</td>
<td>0.083</td>
<td>0.971</td>
<td>0.058</td>
</tr>
<tr>
<td>rs1037700</td>
<td>8:103250775</td>
<td>RRM2B</td>
<td>Intron variant</td>
<td>5.03E-08</td>
<td>0.094</td>
<td>0.973</td>
<td>0.002</td>
</tr>
<tr>
<td>rs5893603</td>
<td>8:103250839</td>
<td>RRM2B</td>
<td>Frameshift variant</td>
<td>4.28E-08</td>
<td>0.093</td>
<td>0.973</td>
<td>0.007</td>
</tr>
<tr>
<td>rs1037699</td>
<td>8:103250930</td>
<td>RRM2B</td>
<td>Missense variant</td>
<td>2.70E-08</td>
<td>0.094</td>
<td>0.976</td>
<td>0.002</td>
</tr>
<tr>
<td>rs16869352</td>
<td>8:103306033</td>
<td>UBR5</td>
<td>Synonymous variant</td>
<td>4.01E-07</td>
<td>0.08</td>
<td>0.975</td>
<td>0.030</td>
</tr>
<tr>
<td>rs61752302</td>
<td>8:103311153</td>
<td>UBR5</td>
<td>Synonymous variant</td>
<td>3.03E-03</td>
<td>0.026</td>
<td>0.977</td>
<td>0.621</td>
</tr>
<tr>
<td>rs72734283</td>
<td>14:75495059</td>
<td>MLH3</td>
<td>Intron variant</td>
<td>4.32E-03</td>
<td>0.089</td>
<td>0.971</td>
<td>0.623</td>
</tr>
<tr>
<td>rs175080</td>
<td>14:75513828</td>
<td>MLH3</td>
<td>Missense variant</td>
<td>7.72E-03</td>
<td>0.435</td>
<td>0.971</td>
<td>0.447</td>
</tr>
<tr>
<td>rs146353869</td>
<td>15:31126401</td>
<td>FAN1</td>
<td>Intron variant</td>
<td>4.30E-20</td>
<td>0.017</td>
<td>0.973</td>
<td>1.000</td>
</tr>
<tr>
<td>SNP</td>
<td>Chromosome</td>
<td>Gene (ID)</td>
<td>Type</td>
<td>p-value</td>
<td>MAF</td>
<td>HWE</td>
<td>HWE (p-value)</td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>---------</td>
<td>------</td>
<td>---------</td>
<td>---------------</td>
</tr>
<tr>
<td>rs114136100</td>
<td>15:31197976</td>
<td>FAN1</td>
<td>Synonymous variant</td>
<td>8.49E-16</td>
<td>0.019</td>
<td>0.976</td>
<td>0.423</td>
</tr>
<tr>
<td>rs150393409</td>
<td>15:31202961</td>
<td>FAN1</td>
<td>Missense variant</td>
<td>9.34E-18</td>
<td>0.013</td>
<td>0.975</td>
<td>1.000</td>
</tr>
<tr>
<td>rs3512</td>
<td>15:31235005</td>
<td>FAN1</td>
<td>3' UTR_variant</td>
<td>5.28E-13</td>
<td>0.283</td>
<td>0.973</td>
<td>1.000</td>
</tr>
<tr>
<td>rs20579</td>
<td>19:48668830</td>
<td>LIG1</td>
<td>NMD transcript variant</td>
<td>6.65E-03</td>
<td>0.134</td>
<td>0.942</td>
<td>0.732</td>
</tr>
</tbody>
</table>

Table 4.3: Characteristics of single nucleotide polymorphisms (SNPs) used in this study.

SNPs were selected from the most significant genes (gene-wide p < 0.1) in the “DNA repair pathway cluster” from the GeM-HD analysis (Consortium, 2015c) (listed in Table S4 of GeM-HD). Genes annotated by the SNPs are indicated. *Refers to the current study. Chr = chromosome; MAF = minor allele frequency; HWE = Hardy-Weinberg equilibrium.
4.2.3 Genotyping

SNP genotyping was performed using custom KASP assays at LGC Genomics (Hertfordshire, UK). Gene level sense sequences were used to design SNP assays (Table 4.4). The assays for several SNPs were designed in reverse orientation to the chromosome (rs4150407, rs1805323, rs1037700, rs1037699, rs3512, and rs20579). For this reason, for all SNPs in reverse orientation to the chromosome (rs4150407, rs1805323, rs1037700, rs1037699, rs3512, and rs20579) genotypes resulting from these KASP assays will be complementary to those using HGVS nomenclature. This is reflected in Table 4.6, where the minor allele for these SNPs differs from GeM-HD which uses HGVS nomenclature (GeM-HD-Consortium, 2015), but corresponds to the same allele.
<table>
<thead>
<tr>
<th>SNPs</th>
<th>HGVS Names</th>
<th>SNP to Chromosome</th>
<th>Seed sense sequences for KASP assay design</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1800937</td>
<td>NC_000002.11:g.48025764C&gt;T</td>
<td>Forward</td>
<td>TTGCTGGGAGGTTAGGCACAATCTTA[C&gt;T]GTAACAGATAAGAGTGAAAGATA</td>
</tr>
<tr>
<td>rs4150407</td>
<td>NC_000002.11:g.128049631T&gt;C</td>
<td>Reverse</td>
<td>AGTACACAGTGAGGAAAGTTGGTTCCAT[A&gt;G]GACCAAGACCTTCCAGAAACTGA</td>
</tr>
<tr>
<td>rs5742933</td>
<td>NC_000002.11:g.190649316G&gt;C</td>
<td>Forward</td>
<td>GTAATTGGGCTGCTGGGCTAGGAG[&gt;G]AAGGTAGTGTTGGTGAGACTAAGGGG</td>
</tr>
<tr>
<td>rs1799977</td>
<td>NC_000003.11:g.37053568A&gt;G</td>
<td>Forward</td>
<td>CTTAACCGTGGGACAAATTTGAGGCAATCC[A&gt;G]CTTTTGGGAAATGGCTTTAGTCGGA</td>
</tr>
<tr>
<td>rs6151792</td>
<td>NC_000005.9:g.80056961C&gt;T</td>
<td>Forward</td>
<td>TCACACACGAGCCATTGAAATTAGGCC[C&gt;T]GACAGCAATTCAGAGAAGGAGAAA</td>
</tr>
<tr>
<td>rs115109737</td>
<td>NC_000005.9:g.80102444G&gt;A</td>
<td>Forward</td>
<td>GAATCAACAGCTTTATTTTGCTATA[G&gt;A]CATTATAAATACCTTTACATCTGT</td>
</tr>
<tr>
<td>rs71636247</td>
<td>NC_000005.9:g.80118976A&gt;G</td>
<td>Forward</td>
<td>TGTAAAAATATAATGAGGAAAGAAC[A&gt;G]CTTAGATAGAAGCCTTTATCCAAA</td>
</tr>
<tr>
<td>rs1805323</td>
<td>NC_000007.13:g.6026942G&gt;T</td>
<td>Reverse</td>
<td>TCCAGTCAAGCAGAGCCGTTGACCCCTA[C&gt;A]GAGACAGAGCGGAGGTGGAGAAGGAC</td>
</tr>
<tr>
<td>rs12531179</td>
<td>NC_000007.13:g.6028687C&gt;T</td>
<td>Forward</td>
<td>ATTTTTTAGAGAGACAGAGTTTGC[C&gt;T]GTTGTTAGATAGTCTGACTCTCTTA</td>
</tr>
<tr>
<td>rs3735721</td>
<td>NC_000008.10:g.103217695A&gt;G</td>
<td>Forward</td>
<td>GCTGGGGGAGGCTTTAGTTGAAGAA[A&gt;G]AACTATTATGTATAATAGCGAC</td>
</tr>
<tr>
<td>rs1037700</td>
<td>NC_000008.10:g.103250775G&gt;C</td>
<td>Reverse</td>
<td>GGCTCAGGCGCGGGTGGAGACTTAC[C&gt;G]CGCTGGTTATCCCGCTACGCCTCT</td>
</tr>
<tr>
<td>rs5893603</td>
<td>NC_000008.10:g.103250839_103250840insG</td>
<td>Forward</td>
<td>TTGGCTGGCCCCGGGCAGAGCAGCAGC[G&gt;C]GAGCGGGAGCGAAACCCAAAGTCAG</td>
</tr>
<tr>
<td>rs1037699</td>
<td>NC_000008.10:g.103250930C&gt;T</td>
<td>Reverse</td>
<td>AGGACAGGCGTGCCTCGGGCCGCCCTCT[C&gt;G]CCGCAGGCTGTTGCGTYTGGCA</td>
</tr>
<tr>
<td>rs16869352</td>
<td>NC_000008.10:g.103306033T&gt;C</td>
<td>Forward</td>
<td>CAGCGTAAGGCTCAATGCTGGGAA[T&gt;C]ACACGCTTGCATTTTCAATTGGCT</td>
</tr>
<tr>
<td>rs61752302</td>
<td>NC_000008.10:g.103311153C&gt;T</td>
<td>Forward</td>
<td>ACAATTTTTCATATTTAAATGAGCATATT[C&gt;T]GCTTTTGATTCCCTGATCTA</td>
</tr>
<tr>
<td>rs72734283</td>
<td>NC_000014.8:g.75495059A&gt;G</td>
<td>Forward</td>
<td>ATTTTTATGGATTGGACCTTTAGACA[A&gt;G]CCCCATTACAGCACTTCCCCATCCAGT</td>
</tr>
<tr>
<td>rs175080</td>
<td>NC_000014.8:g.75513828G&gt;A</td>
<td>Forward</td>
<td>GGTCATAGGACTTTCTCTCAAACTA[G&gt;A]GCATCTGTGGTTCTAAACAATCTTC</td>
</tr>
<tr>
<td>rs146353869</td>
<td>NC_000015.9:g.31126401C&gt;A</td>
<td>Forward</td>
<td>AATGGGTATGTATTTAATGTGAATC[C&gt;A]CAAGAGTGAGTGTCACTGTGCCT</td>
</tr>
<tr>
<td>rs114136100</td>
<td>NC_000015.9:g.31197976C&gt;T</td>
<td>Forward</td>
<td>GCTGCAATGGTCCTGGTCAAACAAC[C&gt;T]GGTCATCCTTACTACCTCCGGAGTT</td>
</tr>
<tr>
<td>rs150393409</td>
<td>NC_000015.9:g.31202961G&gt;A</td>
<td>Forward</td>
<td>GCCCTTCTCAAATTGGCAAAGAC[G&gt;A]TTCAATCTGCATGGGCGAAGAAT</td>
</tr>
<tr>
<td>rs3512</td>
<td>NC_000015.9:g.31235005G&gt;C</td>
<td>Reverse</td>
<td>ACAGAGAGCGTTAAAAAGTAAAGGCA[C&gt;G]GCTCAAGAGTAAACCTGCAATGCG</td>
</tr>
<tr>
<td>rs20579</td>
<td>NC_000019.9:g.48668830G&gt;A</td>
<td>Reverse</td>
<td>GCTGGACAGGAAGGAGAAATCTGA[C&gt;T]GCCACATGCGAAGTATCAGT</td>
</tr>
</tbody>
</table>

**Table 4.4:** Seed sense sequences for SNP KASP assay design.

*Note that genotypes for SNPs in reverse orientation to chromosome given by our KASP assays (highlighted in red) are complementary (reverse) to HGVS nomenclature.*
4.2.4 Statistical analysis

Given the major effect of CAG repeat length on AAO, it was important to remove this effect in order to look for secondary genetic modifiers. Ages of onset for all diseases were corrected for repeat length using a similar method to the GeM-HD GWAS (GeM-HD-Consortium, 2015). A linear regression was performed for each disease separately of ln(AAO) on expanded repeat length, this analysis was done by Professor Peter Holmans, Cardiff University. The regression parameters are given in Table 4.5. These parameters were used to construct an expected value of AAO for each individual, based on their repeat length, which was subtracted from their actual AAO to give a residual. The effect of gender on AAO (after accounting for CAG length) was also tested. Since this was nonsignificant for all disorders, gender was not included in the calculation of residuals.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Sample N</th>
<th>A</th>
<th>B</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD</td>
<td>445</td>
<td>6.119939</td>
<td>-0.052966</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>SCA1</td>
<td>177</td>
<td>5.682974</td>
<td>-0.043694</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>SCA2</td>
<td>294</td>
<td>5.799343</td>
<td>-0.056682</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>SCA3</td>
<td>397</td>
<td>7.137211</td>
<td>-0.049477</td>
<td>&lt;2e-16</td>
</tr>
<tr>
<td>SCA6</td>
<td>69</td>
<td>5.96740</td>
<td>-0.08686</td>
<td>0.00268</td>
</tr>
<tr>
<td>SCA7</td>
<td>73</td>
<td>4.643231</td>
<td>-0.026023</td>
<td>2.94e-5</td>
</tr>
<tr>
<td>SCA17</td>
<td>7</td>
<td>2.38659</td>
<td>0.01716</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 4.5: Effects of repeat length of the expanded allele on the age at onset.

Results of fitting a linear regression \( \ln(\text{AAO}) = A + B \times (\text{CAG})n \). P-value refers to the significance of the regression parameter (B) indexing the effect of repeat length.

Association of each SNP with AAO was tested by performing a linear regression of the residuals from the AAO analysis on the number of minor alleles in the genotype in PLINK (Purcell et al., 2007), this analysis was also done by Prof Holmans.

The primary analysis in this report tested whether there was an overall association of AAO across all 22 SNPs. This was done by combining the association p-values for each SNP using Brown’s method (Brown, 1975b), which is essentially Fisher’s method for combining p-values, corrected for linkage disequilibrium (LD) between SNPs. While Fisher’s method is a way of combining the information in the p-values from different statistical tests so as to form a single overall test, this requires that the individual test statistics should be statistically independent which is not the case if the SNPs are in linkage disequilibrium. Brown proposed the idea of
approximating $X$ using a scales $\chi^2$ distribution $cX^2(k')$ with $k'$ degrees of freedom (Brown, 1975b). The primary analysis used one-sided $p$-values for association in the same direction as that observed in GeM-HD. In order to assess the overall directionality of the associations, we compared the significance to that obtained from a similar analysis using two-sided $p$-values. The analyses were performed on eight disease groups: all (HD+SCAs), HD, all SCAs, SCA1, SCA2, SCA3, SCA6 and SCA7. P-values were Bonferroni corrected for eight tests – this is conservative since the disease groups are not independent. Individual SNPs significantly associated with AAO in each disease group were also noted.

Due to small sample size, SCA17 was not analysed independently, but was included in the analyses of all SCAs and HD+SCAs.

### 4.3 Results

#### 4.3.1 There is a combined effect of 22 DNA repair gene SNPs on Age at Onset

Significant associations (after Bonferroni correction for 8 tests) were observed for HD+SCAs ($p=1.43\times10^{-5}$), HD ($p=0.00194$), All SCAs ($p=0.00107$), SCA2 ($p=0.00350$), and SCA6 ($p=0.00162$) (Table 4.6). The increased significance of these associations compared to an undirected test using two-sided SNP $p$-values (see Table 4.6) indicates concordance in the direction of effects across SNPs between these samples and GeM-HD. Importantly, the observed association with HD is a convincing replication of the GeM-HD results in an independent sample.

<table>
<thead>
<tr>
<th>Disease Group</th>
<th>GeM-HD concordance?</th>
<th>$P$ (All SNPs)</th>
<th>$P$ (High LD SNPs removed)</th>
<th>$P$ (rs3512 removed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL (HD+SCAs)</td>
<td>non directional</td>
<td>$4.74\times10^{-4}$</td>
<td>$2.26\times10^{-4}$</td>
<td>0.00492</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>$1.43\times10^{-5}$</td>
<td>$6.98\times10^{-6}$</td>
<td>2.26$\times10^{-4}$</td>
</tr>
<tr>
<td>HD</td>
<td>non directional</td>
<td>0.0226</td>
<td>0.00775</td>
<td>0.0364</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>0.00194</td>
<td>$4.63\times10^{-4}$</td>
<td>0.00394</td>
</tr>
<tr>
<td>SCAs</td>
<td>non directional</td>
<td>0.0188</td>
<td>0.0236</td>
<td>0.0771</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>0.00107</td>
<td>0.00142</td>
<td>0.00667</td>
</tr>
<tr>
<td>SCA1</td>
<td>non directional</td>
<td>0.376</td>
<td>0.386</td>
<td>0.444</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>0.416</td>
<td>0.287</td>
<td>0.524</td>
</tr>
<tr>
<td>SCA2</td>
<td>non directional</td>
<td>0.0230</td>
<td>0.0629</td>
<td>0.0233</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>0.00350</td>
<td>0.0138</td>
<td>0.00442</td>
</tr>
<tr>
<td>SCA3</td>
<td>non directional</td>
<td>0.176</td>
<td>0.114</td>
<td>0.355</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>0.0809</td>
<td>0.0381</td>
<td>0.205</td>
</tr>
<tr>
<td>SCA6</td>
<td>non directional</td>
<td>0.00588</td>
<td>0.0735</td>
<td>0.00506</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------</td>
<td>---------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>0.00162</td>
<td>0.0340</td>
<td>0.00163</td>
</tr>
<tr>
<td>SCA7</td>
<td>non directional</td>
<td>0.155</td>
<td>0.217</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>Same as GeM-HD</td>
<td>0.0447</td>
<td>0.0885</td>
<td>0.113</td>
</tr>
</tbody>
</table>

**Table 4.6:** Results of combined analysis of SNPs.

P-values in this table obtained by combining single-SNP p-values using Brown’s method (Brown, 1975b), allowing for LD between SNPs. Non-directional analysis combines two-sided p-values. “Same as GeM-HD” analyses combine one-sided p-values in the same direction as the SNP effects observed in GeM-HD study (Consortium, 2015c). In the “High LD SNPs removed” analysis, rs1037700, rs5893603 and rs16869352 were removed due to high LD (r^2>0.8) with more significant SNPs in GeM-HD. P-values coloured red satisfy Bonferroni correction for 8 disease group tests. Note that SCA17 was included in the “HD+SCAs” and “All SCAs” grouped analyses, but was not tested independently due to small sample size.

### 4.3.2 Individual SNPs were also significantly associated with onset

Individual SNP associations were also examined. Three of these were significant after Bonferroni correction for 8 disease combinations and 22 SNPs (Table 4.3, Table 4.7): rs3512 in FAN1 with All SCAs and HD+SCAs and rs1805323 in PMS2 with HD+SCAs. Each association was in the same direction as in GeM-HD. We did not replicate the most significant signal in GeM-HD, rs146353869 (p = 4.30x10^{-20}, associated with 6.1 years earlier age of motor onset of HD). This is likely due to our sample being much smaller than GeM-HD and thus less well powered to find associations with SNPs with relatively low frequency MAF such as rs146353869 (MAF=0.017). However, rs3512, the most significant individual SNP in this study, indexes the second significant chromosome 15 signal in GeM-HD (p=5.28x10^{-11}, associated with 1.4 years later onset of HD), and is in the 3'UTR of FAN1.

Three SNPs (rs1037700, rs5893603, rs16869352) were found to be in high LD (r^2>0.8) in our sample with more significant SNPs from GeM-HD. Removing these SNPs reduced the significance of the multi-SNP associations with SCA2 and SCA6, although these remained nominally significant (see Table 4.6). Finally, all the significant multi-SNP associations from the primary analysis remained significant after removing the most significant single SNP (rs3512) (Table 4.6), suggesting that the signal enrichment is not being driven by a single SNP.
### Table 4.7: Single SNP associations.

Beta denotes the frequency of the minor allele in GeM-HD (3) (MAF [GeM HD]) and the present study (MAF [All]). P values highlighted green satisfy Bonferroni correction for 22 SNPs; those highlighted red satisfy Bonferroni correction for 8 disease groups and 22 SNPs. Note that for SNPs in reverse orientation to chromosome (rs4150407, rs1805323, rs1037700, rs1037699, rs3512, and rs20579) genotypes given by KASP assays (current study) are complementary to those obtained in GeM-HD, which uses HGVS nomenclature (Table 4.3), corresponding to the same allele.
4.3.3 Looking at the combined effect of the SNPs in a polygenic score

To visualise the combined effect of our SNPs on residual AAO a polygenic “age at onset score” was derived by Prof Holmans, defined as the sum of the number of minor alleles at each locus weighted by their effect size in GeM (note that negative scores here correspond to earlier AAO). The residual AAO for each quartile of this risk score was plotted in Figure 4.1. As expected, there was a positive correlation between residual AAO in our data and increasing age at onset score, although the effect was small – the score accounts for approximately 1% of the variance of residual AAO.

![Boxplot of residual AAO (across all samples) by quartiles of polygenic age at onset score. Polygenetic score calculated by summing the number of minor alleles (weighted by their effect on age at onset in the GeM GWAS) across the 22 SNPs. Note that lower scores correspond to earlier than expected AAO, and thus smaller residuals. Figure devised and produced by Professor Peter Holmans for publication in (Bettencourt et al., 2016).](image)

**Figure 4.1**

4.4 Discussion

In the study discussed in this chapter, we showed that DNA repair genes as a group significantly modify AAO in HD, in all SCAs as a group, and in SCA2 and SCA6 independently. Additionally, we have identified potential modifier SNPs in HD, SCA1 and SCA6.

The data suggest that polyglutamine diseases are modulated by a general mechanism which operates at the level of the CAG repeat tract rather than being a huntingtin specific.
phenomenon. As shown in Figure 4.2, the variants genotyped lie in a set of functionally related genes involved in DNA damage repair. In addition to supporting the findings of the GeM-GWAS linking DNA repair genes to HD onset, our data suggest a common mechanism by which genetic variation in DNA repair pathways underlies age at onset in the polyglutamine diseases as a group. Alterations in DNA repair pathways could predispose to earlier onset by interacting with polyglutamine aetiology at various levels (Massey and Jones, 2018, Bras et al., 2015). Repair pathways might operate directly on repeat sequences by licensing or inhibiting repeat expansion in neurons. Alternatively, or in addition, because intriguingly many of the genes containing pathogenic CAG repeats encode proteins that themselves have roles in the DNA damage response, it is possible that repeat expansions impair specific DNA repair pathways. DNA damage could then accrue in neurons, leading to further expansion at repeat loci, thus setting up a vicious cycle of pathology.

![Figure 4.2](https://example.com/figure42.png)

**Figure 4.2:** String diagram illustrating the functional connection between the proteins included in this study. Nodes represent proteins while edges represent protein-protein interactions and are the intensity of the lines reflect low (0.150), medium (0.400), high (0.700) and highest (0.900) confidence). Homo sapiens data used. String-db.org accessed 24/05/2017 (Szklarczyk et al., 2015). (Diagram devised and produced by me).

There are several ataxias caused by mutations in genes involved in the DNA damage response, the first noted being ataxia telangiectasia, a rare recessive childhood neurodegenerative disease caused by mutations in the ataxia-telangiectasia mutant serine/threonine kinase gene (ATM) (Jones et al., 2017). This gene controls cell-cycle arrest after DNA double-strand breaks, often leading to apoptosis and, thus, neurodegeneration (Paull, 2015). Mutations in other
genes that cause incorrect resolution of DNA double-strand breaks lead to severe developmental disorders of the nervous system, such as ataxia-telangiectasia-like disease (hMre11)(OMIM, 2017a), Seckel syndrome 1, involving the ataxia-telangiectasia and Rad3-related protein gene (ATR)(OMIM, 2017b), and Nijmegen breakage syndrome involving (NBN)(Pearl et al., 2015, McKinnon, 2009). These disorders also have widespread systemic effects, in contrast to those resulting from mutations in genes involved in the repair of DNA single-strand breaks, which usually have effects limited to the nervous system, although still with serious clinical outcomes (Paull, 2015). Spinocerebellar ataxia with axonal neuropathy is caused by mutations in the tyrosyl-DNA phosphodiesterase 1 gene (TDP1) and the recessive ataxias with oculomotor apraxia 1, 2, and 4 are caused by mutations in the aprataxin (APTX), senataxin (SETX), and polynucleotide kinase 3-phosphatase (PNKP) genes (Bras et al., 2015) respectively. TDP1 repairs stalled topoisomerase I–DNA complexes, APTX and PNKP46 operate on nucleotides, and SETX encodes a helicase involved in transcriptional termination (Yuce and West, 2013, Hatchi et al., 2015). The relationship between DNA damage and the nervous system, and particularly the cerebellum, is a fascinating outstanding question.

Considering the CAG-repeat disorders, we know that repetitive sequences can form unusual secondary DNA structures (Mirkin, 2007) such as hairpin loops, slipped strands, G-quadruplexes and R-loops. These structural perturbations of DNA have been implicated in both the normal regulation of cellular functions, such as chromatin organization and gene expression, and in the aberrant DNA processing that can lead to genomic instability (Massey and Jones, 2018). DNA mismatch repair proteins bind to these abnormal structures, and in the process of attempting repair cause somatic instability (often expansion) of the CAG repeats. We know that larger CAG repeats are associated with more severe pathology and earlier disease onset in affected patients, therefore somatic expansion of the repeat length provides a plausible mechanism by which the genetic variation we identify here can alter AAO of disease (Figure 4.3).
Figure 4.3: Potential mechanism by which variants in DNA repair could influence somatic expansion of CAG repeats. Hypothesised mechanism of somatic expansion of the CAG repeats in polyglutamine diseases due to variation in genes encoding DNA repair proteins. The accessibility of repetitive DNA sequences during replication, transcription, etc., allows the formation of secondary DNA structures: SNPs in genes encoding DNA repair proteins may alter the kinetics or activity of DNA repair complexes (pink bobble). After endonuclease activity on the opposite strand (nick indicated by the grey arrow), such impaired repair may lead to further expansion of the repeat tracts by consequent gap filling synthesis by DNA polymerase (blue bobble). Figure prepared by Dr Bettencourt, and is published in (Bettencourt et al., 2016).

The prospect of a common mechanism relating to DNA repair driving disease progression across a series of devastating diseases has exciting therapeutic implications since a treatment targeting this pathway in one disease may be transferrable to the other polyglutamine diseases. Some of these diseases are extremely rare, making studying them alone and conducting large scale clinical trials particularly challenging, and the repurposing of drugs from one disease to another is attractive. Using genetics to stratify patients by likely rate of progression also has the potential to improve clinical trial design by stratifying subject variability.

This study described in this chapter had various limitations which likely reduced the power to detect association, and indeed the effects of the studied SNPs on AAO are quite small (Figure 4.3). The small sample sizes for many of the SCAs reduces power both in terms of modelling the relationship between age at onset to CAG repeat length, and in determining the genetic associations themselves. There is likely to be heterogeneity in term of the effect of CAG on AAO for each disease which is why we modelled the effect separately for each disease- but
necessarily reducing the sample size in each disease group, but there may also be other aspects that we have not been able to consider.

The number of SNPs genotyped was limited primarily by financial considerations, thus not all genes in the DNA repair cluster were genotyped, and for most genes only one SNP was genotyped. This limited our ability to interrogate the effect of DNA repair gene variants on AAO.

Additionally, we could not account for interruptions of pure CAG repeat tracts, which may stabilize repeat instability (Menon et al., 2013, Wheeler et al., 2016), thus our power to detect effects mediated by somatic instability may have been reduced.

Notwithstanding these issues, it was demonstrated that DNA repair genes do modulate onset in multiple polyglutamine disease. The ongoing aim of several people who worked on this project is now to replicate the findings with more samples and to genotype more extensively to further explore this relationship. The shared mechanisms uncovered in this study may extend to diseases associated with non-CAG and non-translated repeats, most likely in those that show somatic instability. It would therefore be interesting to look in diseases such as myotonic dystrophy and C9orf72 associated ALS/FTD to see if there is a relationship between DNA repair protein variants and disease manifestation in these conditions. However to do this it will be necessary to adequately establish and control for the effect of repeat size: a considerable challenge in these disorders associated with large expansion mutations.
Chapter 5: Use of sequencing to look for rare variants of larger effect and identify sequence variants in loci highlighted by genetic analysis in Huntington’s disease

5.1 Introduction

As explained in Chapter 1, if sequence variation effecting phenotype is rare but having a large effect size it is unlikely to be picked up by GWAS, which are more commonly used to detect common variants of modest effect. When I started this study in 2013, little was known about the genetic architecture of modifiers of HD, so an exploratory analysis to look at this architecture was therefore of great interest. Furthermore, while genome wide association studies are able to highlight regions of the genome which are associated with a particular trait, they do not tell you what sequence variant is driving the signal: there are many strategies used to try and understand the important variants and their mechanism (Chapter 1). Advances in high throughput sequencing technologies now enable the efficient and cost-effective collection of vast amounts of fine-scale genomic data to complement genome wide association studies (GWAS), and localize causal variants accounting for GWAS hits.

I therefore had two main objectives to be explored by the targeted exome sequencing of the TRACK-HD cohort:

- To look for rare variants of large effect which modify HD
- To look for sequence variation underlying the signal in GWA studies of onset and progression in HD

Successfully sequencing candidates at phenotypic extremes to find rare alleles influencing a genetically complex quantitative trait was previously demonstrated with blood lipid levels (Cohen et al., 2004), and was used to identify a genetic modifier of cystic fibrosis progression (Drumm et al., 2005). Thus in the analysis to look for rare variants of large effect which modify HD I opted to focus on phenotypic extremes. I conducted whole exome sequencing (WES) of the fastest and slowest progressing subjects in TRACK-HD, and compared the variants in a case/control fashion, looking for variants that were enriched in either the fast or slow progressing group.

When the GeM GWAS data became available in 2015 (GeM-HD-Consortium, 2015), and with my own genome wide analysis of progression in 2016-7 (Chapter 3), I examined the WES data...
specifically in the regions highlighted in the genome wide analysis. Rare coding or structural
variants may modify HD onset or progression, and could potentially be identified by exome
sequencing (Majewski et al., 2011, Kiezun et al., 2012). The approaches are complimentary
since the GWAS, given their higher sample sizes, have greater power, while the WES data
provide valuable sequence data of patients with HD, and enable the possibility of locus
heterogeneity to be investigated.

Two loci are the focus of the work in this chapter: the locus on chromosome 5 overlying the
DNA mismatch repair (MMR) protein MutS Homolog 3 (MSH3) (Habraken et al., 1997, New et
al., 1993, Miret et al., 1993) (Chapter 3), and the recently identified DNA interstrand cross link
repair protein FANCD2 and FANCI associated nuclease 1 (FAN1) (Huang and D’Andrea, 2010,
MacKay et al., 2010, Smogorzewska et al., 2010, O’Donnell and Durocher, 2010, Kratz et al.,
2010, Liu et al., 2010) implicated by the chromosome 15 signals in the GeM GWAS (GeM-HD-
Consortium, 2015). FAN1 is a DNA endo/exonuclease involved in DNA repair that is highly
expressed in the brain (MacKay et al., 2010, Consortium, 2015b).

These two DNA damage response (DDR) proteins have also been implicated in other DNA
repair pathways (Jin and Cho, 2017, Brown et al., 2016, Cannavo et al., 2007, Schmutte et al.,
2001, Sugawara et al., 1997), and interactions between mismatch repair (MMR) and
interstrand cross link (ICL) DNA repair pathways have been reported (Goold et al., 2019), with
FAN1 capable of compensating for loss of EXO1 MMR activity under some circumstances
(Desai and Gerson, 2014). Therefore, FAN1 and MMR components may modulate HD AAO
through a shared mechanism. A stable physical interaction between FAN1 and MutLα
components MLH1 and PMS2 further supports this hypothesis (MacKay et al., 2010). The
putative role of MSH3 and FAN1 in Huntington’s disease and other repeat disorders is
discussed at greater length in Chapters 3 and 4.

5.2 Materials and Methods

5.2.1 Whole Exome Sequencing

5.2.1.1 Subject selection
My objective was to identify the most extreme subjects in terms of progression for focused
genetic analysis. The principal component analysis data (Chapter 2), performed in
collaboration with Prof Douglas Langbehn, was used to guide the choice of fast and slow
progressing subjects, with the age at onset data used as supporting data. I calculated expected AAOs from the Langbehn equation (Langbehn et al., 2004). Principal Component (PC) Score alone was used as the basis for the selection of the 25 fastest progressing subjects. The subject with the 21st lowest principal component score was not included on the basis of them having symptom onset later than would have been predicted at birth which is not consistent with them having atypically fast progression.

The selection of the slow progressors was based on not only the Principal Component progression score, but also the Age at Onset (AAO) data. This is because there is less sensitivity of the measures to differentiate change between individuals among this group because they were less clearly ‘atypical’ (Figure 5.1). Thus there was a less clear delineation of the slow progressors.

13 subjects (the very slowest), were chosen on the basis of having extremely slow progression based on their Progression score. A further 10 subjects were chosen on the basis of them having onset more than 3 years later than expected: expected AAOs were calculated from the Langbehn equation (Langbehn et al., 2004). This has been done in two groups, firstly, those subjects who given their age would be expected to have symptoms but do not (n=2), and secondly people who have symptoms but who developed them later than their expected AAO (n=8). Subjects were excluded if their progression score was not consistent with them being a slow progressor even though their AAO data suggested that they were.
5.2.1.2 Sample collection and DNA extraction

All biosamples were collected during TRACK-HD visits. I was responsible for collecting and processing biosamples from London site TrackON-HD subjects which was a follow-up study to TRACK-HD and followed very similar protocols to TRACK-HD study.

Blood for DNA was collected from the antecubital fossa in one ACD (Acid Citrate Dextrose) tube. After the blood draw, the tube was inverted 10 times then placed upright at room temperature. Samples were then shipped overnight at ambient temperature to Biorep, Milan, Italy. The cells were used to generate lymphoblastoid cell lines.

A manual salting-out procedure was used to extract DNA at Biorep, for the WES lymphoblastoid cell line DNA was used. The routine quality control tests performed consist of: spectrophotometric analysis (Nanodrop) to quantify and estimate the DNA quality by OD 260/OD 280 ratios, gel electrophoresis to establish the integrity of the DNA, and sample identity was confirmed by gender and microsatellite analysis.

5.2.1.3 Sequencing pipeline

I arranged for the DNA to be shipped on dry ice to DeCODE Genetics, Iceland. This sequencing was done through our membership of the European Commission Neuromics consortium for which DeCODE was a partner. DNA quality was assessed using picogreen measurements. Sequencing libraries were prepared using the Illumina Nextera Exome method which involves pooling of up to 12 samples for exome enrichment (see Chapter 2). The target region is approximately 62Mb of exons, untranslated regions and noncoding RNA. Pooled libraries were validated using the LibraryQC workflow of the MiSeq sequencing instruments. Validation includes assessment of cluster densities, insert size of each sample within the pool and the relative distribution of each sample within a pool. Finally, validated pooled libraries were sequenced on a HiSeq 2000 (paired-end, 2x100 cycles).

5.2.1.4 Bioinformatics pipeline

Raw fastq files were shipped to me and I transferred them to Dr Vincent Plagnol, UCL Genetics Institute, for bioinformatic analysis as a part of the UCL exomes consortium. Raw fastq files were aligned to the GRCh37 reference genome using novoalign version 2.08.03. Duplicate reads were marked using Picard tools MarkDuplicates. Calling was performed using the haplotype caller module of GATK (https://www.broadinstitute.org/gatk, version 3.3-0),
creating gVCF formatted files for each sample. The individual gVCF files for the exomes discussed in this study, in combination with ~ 3,000 clinical exomes (UCL-exomes consortium), were combined into merged VCF files for each chromosome containing on average 100 samples each. The final variant calling was performed using the GATK Genotype GVCFs module jointly for all samples (cases and controls). Variant quality scores were then recalibrated according to GATK best practices separately for indels and SNPs. Resulting variants were annotated using ANNOVAR based on Ensembl gene and transcript definitions. Candidate variants were filtered based on function (non-synonymous, presumed loss-of-function or splicing variants, defined as intronic sites within 5 bp of an exon-intron junction) and minor allele frequency (< 0.5% minor allele frequency in our internal control group, as well as the NHLBI exome sequencing dataset). Prediction tools including PolyPhen-2 (Adzhubei et al., 2010), SIFT(Kumar et al., 2009), Mutation Assessor(Reva et al., 2007), Mutation Taster(Schwarz et al., 2010) and PhyloP (Pollard et al., 2010) were used to help stratify variants of unknown significance (Ng, 2008), and OMIM used to investigate potential associated phenotypes.

The 25 fast Huntington’s disease progressors and 23 slow Huntington’s disease progressors were compared in a case control design. Fast progressors were defined as cases, and slow progressors as controls, and were compared using bi-allelic tests. Control population frequencies were taken into account: the primary analysis looked for an excess of rare variants in cases compared to controls. Both external and internal population frequencies were used: external to define a frequency filter (around 25% of controls), and the internal set was used in the case control analysis. The case control tests were done both in terms of single variant, and also at the gene based level.

For the biallelic test, samples were marked as “1” if they contain at least 2 somewhat rare (MAF < 0.5%) putatively functional variants. Note that this is done without specific knowledge of whether these variants are on the same, or on different, haplotypes. Homozygous individuals for such rare variants count as two alleles. There was a minimum read depth of 5 or more for the homozygous calls.

A disease modifier of a disease such as HD may only be deleterious in the presence of mutant huntingtin so while we focus on rare variants more common variants are also considered in the case control analysis.
5.2.2 Pathway analysis of WES data

Sequence kernel association test (SKAT) is a SNP-set (e.g., a gene or a region) level test for association between a set of rare (or common) variants and dichotomous or quantitative phenotypes. SKAT aggregates individual score test statistics of SNPs in a SNP set and efficiently computes SNP-set level p-values, e.g. a gene or a region level p-value, while adjusting for covariates, such as principal components to account for population stratification (Wu et al., 2011). Importantly, SKAT allows for a mixture of risk and protective rare alleles in the same gene (Sham and Purcell, 2014).

The SKAT gene-wide results from the TRACK-HD fast vs slow progressor WES analysis were put through GSEA (Chapter 2) using the pathways significant at p>0.05 in GeM GWAS (GeM-HD-Consortium, 2015).

5.2.3 eQTL analysis of MSH3 variant

Braineac (UKBEC, 2015, Ramasamy et al., 2014) was used to evaluate the effect on MSH3 expression of having the rs184967 variant using the stratify expression by SNP function, data accessed December 2014, rechecked 29/08/2018.

5.2.4 Sanger sequencing of MSH3 region of interest

MSH3 FASTA sequence data was obtained from ENSEMBLE, GRCh37. I used SNPmasker 1.1 to mask repeats to prevent excess primer binding, then used Primer3 (Untergasser et al.) to design primers around 100bp either side of the repeat; with melt temperature set to 60°C.
- Left primer: TTGCCCTGCCATGTCTCG
- Right primer: TCCCACCTTCCCCTCTTCA

I carried out the PCR and sequencing reactions using a standard protocol (Appendix 1). Specifically, 1μg genomic DNA was used as a template in a final volume of 25μl with MegaMix Blue (Clint Life Science) and 0.5μM stock primer. They were run with the following cycling conditions: (a) 95°C for 1 min (b) 95°C for 30 secs (c) 58°C for 30 secs (d) 72°C for 1 min (e) Go to step b) for an additional 34 cycles. PCR product was cleaned-up using microCLEAN (Clint Life Science).

Sequencing was conducted using 1μl BigDye (Thermo Fisher Scientific), 5μl BetterBuffer (Clint Life Science) and 7.25μl 18MΩ ddH2O, 0.75μl sequencing primer (at 5μM concentration) and 1μl of PCR product. They were run with the following cycling conditions a) 96°C for 1 min (b) 96°C for 10 secs (c) 50°C for 5 secs (d) 60°C for 3 mins (e) Go to step b) for an additional 24 cycles.
cycles. The sequencing product was cleaned using EDTA and ethanol. Samples were resuspended in Hi-Di formamide and heated to denature before electrophoresis on an ABI 3730xl DNA Analyzer (Thermo Fisher Scientific).

Given the difficulty analysing repetitive DNA sequences, I extracted the sequence data using Sequence Scanner, then used nucleotide BLAST to align the sequences to wild type Homo sapience MSH3 using the somewhat similar sequence option. The output showed the presence or absence of the deleted region, however it was challenging to differentiate homozygous from heterozygous sequences.

5.2.6 Interrogation of RD-Connect database
The RD-Connect (Thompson et al., 2014) database of 1280 WES and whole genome sequencing (WGS) samples was interrogated, looking for subjects who held the MSH3 SNPs rs557874766 and rs1382539 which were highlighted by the HD Progression GWAS (Hensman Moss et al., 2017b), and also deletions in this region, data accessed 19/06/2017. I had access to RD-Connect through my involvement in the Neuromics Project which was part of an allied European Commission FP7 Grant, but access is freely available via an application process.

5.2.7 MSH3 structural prediction
I used the FASTA MSH3 protein sequence, both with and without the AAAAAAPPA deletion to look at the predicted effect of the presence of this deletion on the protein. I inputted the sequences into Raptorx, a web portal for protein structure and function prediction (Källberg et al., 2012) (accessed May 2017). This predicts structure properties of a protein sequence without using templates, including 3-/8-state secondary structure, solvent accessibility, and disordered regions. The 3-state structures feature helix, sheet and coil whereas the 8-state structures feature α helix, 3-helix, 5-helix (π helix), extended strand β ladder, isolated β bridge, hydrogen bonded tern, bends and coils. Raptorx also does structural prediction: creating tertiary structures based on templates from the Protein Data Bank (PDB), and contact map prediction: which uses a deep learning model to create a contact map and tertiary structure, which doesn’t use template information.

5.2.8 Phylogenetic analysis
Using Uniprot (The Uniprot, 2017) I ran a BLAST (Basic Local Alignment Search Tool) of the wild type human MSH3 protein sequence, then investigated the alignment against other ape MSH3 or MutS homolog 3 sequences. Default parameters were used: the default transition
matrix is Gonnet, gap opening penalty is 6 bits, gap extension is 1 bit. Glustal-Omega uses the HHAlign algorithm and its default settings as its core alignment engine (Soding, 2005).

5.3 Results

5.3.1 Whole Exome Sequencing

25 fast and 23 slow progressing TRACK-HD subjects underwent WES. All samples were sequenced at deCODE Genetics, Iceland, with a mean sequence depth of 97-fold, an average uniformity of 89% of targets with ≥20 reads, and an average of 44% of reads on target.

5.3.2 Several DNA repair pathways nominally associated with HD progression in the WES fast vs slow analysis.

Pathways with a p<0.05 in the GSEA of the TRACK-HD WES fast vs control analysis are shown in Table 5.1. Although the sample size is too small to infer much from these data and no results are significant after correcting for multiple comparisons, it is notable that the mismatch repair complex and several other DNA repair pathways reach nominal significance (in bold in Table 5.1). Other pathways of potential interest include those related to the extracellular matrix and RNA capping.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Number of genes</th>
<th>GSEA p (TRACK)</th>
<th>p (AAO-meta)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO: 31012</td>
<td>366</td>
<td>0.0042</td>
<td>0.02214</td>
<td>extracellular matrix</td>
</tr>
<tr>
<td>GO: 32300</td>
<td>8</td>
<td>0.0052</td>
<td>0.00000</td>
<td>mismatch repair complex</td>
</tr>
<tr>
<td>GO: 6370</td>
<td>25</td>
<td>0.0078</td>
<td>0.01801</td>
<td>7-methylguanosine mRNA capping</td>
</tr>
<tr>
<td>GO: 35035</td>
<td>13</td>
<td>0.0104</td>
<td>0.07773</td>
<td>histone acetyltransferase binding</td>
</tr>
<tr>
<td>GO: 48742</td>
<td>43</td>
<td>0.0144</td>
<td>0.01939</td>
<td>regulation of skeletal muscle fibre development</td>
</tr>
<tr>
<td>MGI: 11073</td>
<td>11</td>
<td>0.0168</td>
<td>0.08094</td>
<td>Abnormal macrophage apoptosis</td>
</tr>
<tr>
<td>GO: 9452</td>
<td>28</td>
<td>0.0194</td>
<td>0.03728</td>
<td>7-methylguanosine RNA capping</td>
</tr>
<tr>
<td>GO: 36260</td>
<td>28</td>
<td>0.0194</td>
<td>0.03728</td>
<td>RNA capping</td>
</tr>
<tr>
<td>REACTOME 715</td>
<td>23</td>
<td>0.0196</td>
<td>0.01534</td>
<td>REACT:MRNA_CAPPING</td>
</tr>
<tr>
<td>REACTOME 1035</td>
<td>21</td>
<td>0.0200</td>
<td>0.01446</td>
<td>REACT:RNA POL II CTD PHOSPHORYLATION AND INTERACTION WITH CE</td>
</tr>
<tr>
<td>NCI: 126</td>
<td>26</td>
<td>0.0212</td>
<td>0.06995</td>
<td>NCI: PROTEOGLYCAN SYNDICAN-MEDIATED SIGNALING EVENTS</td>
</tr>
</tbody>
</table>

Table 5.1: DNA repair pathways nominally associated with HD progression in the WES fast vs slow analysis.
Table 5.1: Pathways with an association to age of onset in the GeM GWAS (p<0.05) that also are associated with HD progression (p<0.05) in the TRACK-HD WES analysis.

Several pathways from the DNA repair pathway cluster (highlighted in bold) are nominally significant in both studies.

5.3.3 Sequence variants in FAN1 were identified from the exome sequence data

A region of interest on chromosome 15 in the region of the FAN1 gene was previously highlighted (GeM-HD-Consortium, 2015).

I compared the number of variants identified in cases (fast progressors) in each of the genes near this chromosome 15 region of interest. Though the numbers are too small for robust analysis it is evident that there are more variants in FAN1 than in any of the surrounding genes or pseudogenes (Table 5.2).
<table>
<thead>
<tr>
<th>Gene/ pseudogene</th>
<th>Number of variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOLGA8J</td>
<td>0</td>
</tr>
<tr>
<td>GOLGA8T</td>
<td>0</td>
</tr>
<tr>
<td>KFZP434L187</td>
<td>0</td>
</tr>
<tr>
<td>CHRFAM7A</td>
<td>0</td>
</tr>
<tr>
<td>GOLGA8R</td>
<td>0</td>
</tr>
<tr>
<td>GOLGA8H</td>
<td>0</td>
</tr>
<tr>
<td>ARHGAP11B</td>
<td>0</td>
</tr>
<tr>
<td>OC100288637</td>
<td>0</td>
</tr>
<tr>
<td>HERC2P10</td>
<td>0</td>
</tr>
<tr>
<td>FAN1</td>
<td>5</td>
</tr>
<tr>
<td>MTMR10</td>
<td>1</td>
</tr>
<tr>
<td>MIR211</td>
<td>0</td>
</tr>
<tr>
<td>TRPM1</td>
<td>2</td>
</tr>
<tr>
<td>RP11-16E12.2</td>
<td>0</td>
</tr>
<tr>
<td>KLF13</td>
<td>0</td>
</tr>
<tr>
<td>OTUD7A</td>
<td>0</td>
</tr>
<tr>
<td>CHRNA7</td>
<td>0</td>
</tr>
<tr>
<td>GOLGA8K</td>
<td>0</td>
</tr>
<tr>
<td>ULK4P3</td>
<td>0</td>
</tr>
<tr>
<td>ULK4P1</td>
<td>0</td>
</tr>
<tr>
<td>ULK4P2</td>
<td>0</td>
</tr>
<tr>
<td>GOLGA8O</td>
<td>0</td>
</tr>
<tr>
<td>ARHGAP11A</td>
<td>0</td>
</tr>
<tr>
<td>SCG5</td>
<td>0</td>
</tr>
</tbody>
</table>

*Table 5.2: Number of variants identified in cases showing an excess of rare variants in FAN1 compared to other genes in the Ch15 region of interest highlighted by the GeM-GWAS (GeM-HD-Consortium, 2015).*

The gene based summary of the Case vs Control analysis showed several variants in FAN1 (Table 5.3). The variants are well covered and have good quality scores.
<table>
<thead>
<tr>
<th>Group</th>
<th>Variant</th>
<th>Protein domain</th>
<th>MAFF in GeM GWAS</th>
<th>Prediction of functional effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amino acid</td>
<td>cDNA</td>
<td>SNP ID</td>
<td>AAO residual (onset minus expected onset)</td>
</tr>
<tr>
<td>Fast</td>
<td>p.R145H</td>
<td>c.G434A</td>
<td>rs146408181</td>
<td>9</td>
</tr>
<tr>
<td>Fast</td>
<td>p.E240K</td>
<td>c.G718A</td>
<td>rs150748572</td>
<td>5</td>
</tr>
<tr>
<td>Fast</td>
<td>p.Q829H</td>
<td>c.G2487C</td>
<td>Novel</td>
<td>7</td>
</tr>
<tr>
<td>Fast</td>
<td>p.F762F</td>
<td>c.C2286T</td>
<td>rs200756403</td>
<td>16</td>
</tr>
<tr>
<td>Slow (n=2)</td>
<td>p.P894S</td>
<td>c.C2680T</td>
<td>rs80120912</td>
<td>170, 209</td>
</tr>
</tbody>
</table>
Table 5.3: FAN1 variants identified in fast (n=5) and slow (n=3) progressing subjects from the TRACK-HD cohort. Standard nomenclature is used for the amino acids: R = arginine; H = histidine; E = glutamic acid; K = lysine; Q = glutamine; F = phenylalanine; P = proline; S = serine; W = tryptophan. SIFT (Sorting Tolerant From Intolerant algorithm) (Kumar et al., 2009). Domains - UBZ: ubiquitin-binding zinc finger; SAP: SAF-A/B, Acinus and PIAS; TPR: tetratricopeptide repeat; VRR: virus type replication-repair nuclease.

<table>
<thead>
<tr>
<th>Slow</th>
<th>p.R377W</th>
<th>c.C1129T</th>
<th>rs151322829</th>
<th>204</th>
<th>No onset</th>
<th>UBZ-SAP</th>
<th>0.0014</th>
<th>&gt;1E-5 or not present</th>
<th>Damaging</th>
<th>Damaging</th>
<th>Medium</th>
</tr>
</thead>
</table>

Table 5.3: FAN1 variants identified in fast (n=5) and slow (n=3) progressing subjects from the TRACK-HD cohort. Standard nomenclature is used for the amino acids: R = arginine; H = histidine; E = glutamic acid; K = lysine; Q = glutamine; F = phenylalanine; P = proline; S = serine; W = tryptophan. SIFT (Sorting Tolerant From Intolerant algorithm) (Kumar et al., 2009). Domains - UBZ: ubiquitin-binding zinc finger; SAP: SAF-A/B, Acinus and PIAS; TPR: tetratricopeptide repeat; VRR: virus type replication-repair nuclease.
Assessing the potential functional impact of variants is performed by a variety of prediction tools as a part of the bioinformatic analysis pipeline, and various tools are available which aim to pinpoint phenotypically causal variants (Cooper and Shendure, 2011). I have summarised the results of these tools on the variants listed above in Table 5.3.

### 5.3.4 Two MSH3 variants were highlighted by the WES fast vs slow analysis

The case control analysis of fast vs slow HD progressors highlighted two variants in the DNA mismatch repair gene *MSH3* among the 15 most significant variants. I was initially interested in these variants given the DNA mismatch repair pathway is highlighted by the GWAS (and WES) pathway analysis and conducted some preliminary analysis. Following the results of the HD progression GWAS (Hensman Moss et al., 2017b) which highlighted the region around the N-terminus of *MSH3* as significantly associated with HD progression I went back to these data to explore further; the results of these analyses will be presented together here.

The first *MSH3* variant highlighted by the case control analysis was rs184967. pE949R in exon 21, changing an uncharged to a positive residue. Fisher P value = 0.000217 (Table 5.4). The Minor Allele Frequency (MAF) of A = 0.098/490 (dbSNP). The minor allele is associated with slower progression. The read depth is good and there are calls for all subjects. rs184967 is not significantly associated with AAO in HD according to the GeM GWAS (GeM-HD-Consortium, 2015) (p=0.169). However given that there was an overlap of samples used for my HD Progression GWAS it is not surprising that the p-value for rs184967 = 1.06x10⁻⁴ in this study (Hensman Moss et al., 2017b).

<table>
<thead>
<tr>
<th></th>
<th>Fast progressors</th>
<th>Slow progressors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of G allele</td>
<td>46</td>
<td>27</td>
</tr>
<tr>
<td>Frequency of A allele</td>
<td>4</td>
<td>19</td>
</tr>
</tbody>
</table>

*Table 5.4: Frequency of MSH3 variant rs184967 alleles in fast and slow progressors*

rs184967 is associated with variability in MSH3 expression in the thalamus according to BRAINEAC data (UkBEC, 2015). The minor allele (A), which is found in higher frequency in the slow progressors, is associated with lower MSH3 expression in the thalamus, raising the possibility that progression is slowed in those with the A allele via this eQTL (Figure 5.2).
Figure 5.2: Influence of rs184967 allele status on brain expression of MSH3. eQTL analysis of rs184967 using Braineac database (UKBEC, 2015). Genotype (AA/AC/CC) shown on x-axis; expression levels in log2 scale shown on y axis.

The second MSH3 variant highlighted by the case control analysis was rs201874762, a 27 base pair non-frameshift deletion variant in exon 1. The read depth in the WES is poor, and the variant hasn’t been called for all subjects (Table 5.5), but despite this there is association between the deletion and fast vs slow status, Fisher P value = 0.000528.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Fast progressors</th>
<th>Slow progressors</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCAGCGGCTGCAGCGGCC</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>- (deletion)</td>
<td>2</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 5.5: Frequency of rs201874762 in TRACK-HD fast and slow progressors

rs201874762 is not detailed in 1000 genomes according to SNAP. This variant was neither genotyped nor imputed in the GeM GWAS. I therefore looked at variants in a 10kb window around the transcript boundaries of MSH3 (as defined by NCBI: 79950467-80172634) in the GeM GWAS study (GeM-HD-Consortium, 2015). The SNP with the lowest P-value in MSH3 in the GeM GWAS is rs6151792, p=1.47x10^-4, MAF 0.1.
5.3.5 MSH3 coding variant rs557874766, the index SNP from TRACK-HD GWAS was not found in exome sequence data

The data from the HD Progression GWAS highlighted MSH3, and specifically rs557874766 in the N-terminal region of the protein (Chapter 3) as being associated with slower progression in HD. According to dbSNP rs557874766 encodes a Pro67Ala change in MSH3 and has a (reported) MAF of G=0.2179/1091 (1000 Genomes). Of note, rather than being directly genotyped it was a SNP which was imputed in the GWAS. I re-interrogated the WES data of 48 fast and slow progressing TRACK-HD subjects to look at this region of high association in MSH3. However, when I looked at the WES sequence data on the DeCODE browser is was clear that no subjects had this rs557874766 variant. Instead people either do/ don’t have a small deletion (rs144629981) over this exact location (Figure 5.3). rs557874766 is very close (24 bases) to the deletion that came out of our WES fast / slow analysis described above: rs201874762 (Figure 5.3). Neither rs557874766 nor rs144629981 are on the reference panels meaning that obtaining LD data from online resources was not possible, however in my analysis there were co-segregated. A colleague attempted to SNP genotype samples for the presence of rs557874766 but no samples had it (personal communication, data unpublished). Comparison of the WES data with the SNP genotyping used for the GWAS showed that 25/25 people who are homozygous wild type according to the GWAS data have no deletions at the locus, whereas 6/6 of those who are homozygous variants have a deletion. The data for the 17 heterozygotes was more difficult to interpret particularly as coverage of this region was low. rs144629981 was called in 16/17 expected heterozygotes, but is clearly evident in the sequence data from the other person. rs201874762 is less well covered and called, it is called in 8/17 expected heterozygotes.

5.3.6 SNP in high linkage disequilibrium with rs557874766 was identified

Based on the data above I hypothesised that the presence of the deletion(s) rather than the rs557874766 SNP is observed in slow progressors and driving the GWAS signal. I therefore identified a SNP in high linkage disequilibrium: rs1382539, to facilitate identification of subjects expected to have the haplotype associated with slow progression. rs1382539 has an r²=0.91 with rs557874766 according to SNiPa (Arnold et al., 2015), and P=8.7013e⁻⁰⁸ in the TRACK-HD progression GWAS (Hensman Moss et al., 2017b); P=5.278e⁻⁰⁹ in the TRACK-HD/REGISTRY meta-analysis (Hensman Moss et al., 2017b)(Chapter 3).

5.3.7 Sanger sequencing of TRACK-HD subjects provided further evidence for the presence of deletions in people expected to have rs557874766
To further explore this region of MSH3 I performed Sanger sequencing of 125 members of the TRACK-HD cohort of which 103 were successfully sequenced (Table 5.6). Not one subject had the rs557874766 variant. Interestingly most of those that failed sequencing had the full wild type sequence while none with deletions failed sequencing. As with the WES it was more difficult to call the sequences of the subjects expected to be heterozygotes.

<table>
<thead>
<tr>
<th>rs557874766 genotype according to the GWAS (Hensman Moss et al., 2017b)</th>
<th>TRACK-HD subject (n=125)</th>
<th>Number with net 27 bp deletion</th>
<th>Number without net 27bp deletion</th>
<th>Inconsistencies between forward &amp; reverse strand / split alignment</th>
<th>Number failed sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild type</td>
<td>79</td>
<td>0</td>
<td>60</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Heterozygote</td>
<td>33</td>
<td>10</td>
<td>20</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Homozygote deletion</td>
<td>13</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>125</td>
<td>27</td>
<td>70</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

*Table 5.6: Results from the Sanger sequencing of TRACK-HD cohort subjects, showing the expected genotypes based on the GWAS, and whether deletions were found.*

Together the Sanger and exome sequencing data provided sequence information on a total of 173 TRACK-HD subjects, who had all also been included in the GWAS. All 75 sequenced subjects who were wild type at rs557874766 had no deletion, while all 19 homozygote variant at rs557874766 had a deletion at the locus. These data suggest that subjects who were driving the GWAS signal and thought to have the rs557874766 variant instead had either one large or a pair of deletions at this locus making up a total of 27 deleted base pairs. A more extensive investigation of this region of MSH3 has been undertaken by colleagues as a follow-up to my findings above, but is beyond the scope of this thesis (Flower et al., 2019).

**5.3.8 rs557874766 was not found in sequence data of 1280 individuals**

I interrogated the RD-Connect database of 1280 WES and whole genome sequencing (WGS) samples for rs557874766, rs1382539 and the deletions, and found that there were no samples holding the rs557874766 SNP. Given its reported allele frequency (0.2179) one would have expected around 279 samples to hold the SNP, and indeed there were 386 subjects who had the SNP rs1382539 which was thought to be in high LD with rs557874766 (see above).
396 samples were found to have the 9bp deletion rs144629981, while 460 people had a 18bp deletion which starts 1bp in the N-terminal direction to rs201871762 (TGCAGCGGCTGCAGCGGCC) which from the data appeared to be a data calling issue. This provided further evidence that there was likely to be an issue with the calling and alignment of this N-terminal region of MSH3 in published databases.

5.3.9 Structural predictions show that slow progressors have lost an alpha-helical region in the N-terminus of MSH3

The mutated region of interest within MSH3 maps to an intrinsically disordered N-terminal region of the protein. Secondary structure predictions show that in the wild type protein there is an alpha-helical region between residues 49-60, however, in the subjects who have the deletion associated with slower progression, (and genotyped as having rs557874766 in the GWAS) this alpha-helical structure is lost (Figure 5.3).

Figure 5.3: Secondary structure predictions for MSH3 in the wild type form (A) and with the deletion (B) which is seen in the slow progressing subjects who drive the progression GWAS
signal showing that an alpha-helical region is lost in the deleted form at residues 49-50. The rest of the protein appeared to be the same between the two forms. The 8 state version of Raptorx was used (Källberg et al., 2012) and was accessed 22/06/17. Only residues 1 – 400 are shown however the whole protein sequence was inputted. The Legend for the secondary structures is shown above within the figure.

Solvent exposure prediction analysis suggests that the region of interest is exposed in both wild type and deleted forms (Figure 5.4). However in the wild type form with (Ala)_{12} the residues have a 60-85% odds of being exposed whereas in the deleted form with (Ala)_{6} the residues have a 55-85% odds of being exposed.

Figure 5.4: Predicted solvent exposure for MSH3 wild type (A) and with deletion (B) using Raptor (Källberg et al., 2012) (accessed 22/06/17) showing that the region containing the deletion is predicted to be exposed both with and without the deletion present. Only residues 1 – 320 are shown however the whole protein sequence was inputted. The Legend for the solvent accessibility is shown above within the figure.
The published crystal structure does not include the N-terminal domain that we are particularly interested in: Gupta et al used residues 219-1134 complexed to MSH2 and a 4 base loop of DNA to produce their structures (Gupta et al., 2012). There are no templates on the Protein Data Bank which include the disordered N-terminal region that has been highlighted by our genetic analysis. However the structure of the protein from residue 211 to 1128 according to RaptorX is shown in Figure 5.5.

**Figure 5.5**: Tertiary structure predictions of MSH3 from residue 211 to 1128 generated by RaptorX (Källberg et al., 2012) accessed 22/06/17.

**5.3.10 Phylogenetic data suggest that the polyalanine can be viewed as a recent insertion**

Data was available for Pan troglodytes (Chimpanzee) (A0A2J8NN13), Pongo abelii (Sumatran orangutan) (A0A2J8UWT4), Pan paniscus (Bonobo) (A0A2R9B289), Gorilla gorilla gorilla (Western lowland gorilla)(G3R048) and Nomascus leucogenys (Northern white-cheeked gibbon) (G1RRE8). Overall there was 91.9% identity between the sequences. The cladogram produced is shown in Figure 5.6.
Figure 5.6: Cladogram of the apes showing the MSH3 protein sequence at the repetitive region of interest in six different ape species, showing that the presence of 12 alanines is novel to the human lineage. Gibbons, lower apes which diverged earlier from the lineage have just three consecutive alanines and also the most divergent sequence overall; orangutans also just have three alanines while gorillas and both types of chimpanzee have six. Sequence data is from UniProt (The UniProt, 2017). The sequence is shown from residue 51 which is an alanine conserved among the apes, to a phenylalanine which is also conserved. A: alanine, P: proline, F: phenylalanine. (A)3 delineates three consecutive alanine residues. Approximate times of lineage divergence (Locke et al., 2011) are shown on the right hand side. Myr: million years.

5.4 Discussion

In this chapter I have discussed the use of sequencing technologies to follow up on findings from large scale association studies which have implicated the DNA damage response (DDR) proteins FAN1 and MSH3 as modifiers of Huntington’s disease. The work presented in this chapter is limited in its small sample sizes and largely exploratory nature, however it provided important information about several variants which had been highlighted by previous association studies which has subsequently been followed up by other members of the Tabrizi group and collaborators.

While case control analysis of the WES data of the fastest and slowest progressing subjects in TRACK-HD did not yield any statistically significant variants or genes, one of the variants in
MSH3 highlighted by that analysis was subsequently given greater scrutiny following the results of the HD progression GWAS (Chapter 3) (Hensman Moss et al., 2017b), and several sequence variants within FAN1 were identified for functional follow-up in the laboratory. FAN1 was first described in four papers in 2010 (O'Donnell and Durocher, 2010): it was identified by virtue of its interaction with mismatch repair proteins (Cannavo et al., 2007, Kratz et al., 2010) domain homology (Liu et al., 2010, MacKay et al., 2010) and from a mitomycin C (MMC) sensitivity RNA interference screen (Smogorzewska et al., 2010). FAN1 exhibits domain architecture suggestive of a role in DNA repair, bearing a RAD18-like ubiquitin-binding (UBZ) domain, a putative DNA-binding (SAP) domain, a protein-protein interaction motif and a nuclease domain of the VRR_nuc family. All four studies uncovered a conserved role for FAN1 in DNA interstrand cross-link (ICL) repair, demonstrating that FAN1 deficiency sensitizes human cells and nematodes to crosslinking agents such as MMC and increases chromosome instability. FAN1 orthologs are members of the ancient restriction endonuclease-like superfamily. FAN1 cleaves DNA: it preferentially cleaves branched DNA structures that mimic intermediates of DNA repair, with a strong preference for the 5' DNA flap; it also possesses 5'-3' exonuclease activity. FAN1 is known to interact with both mismatch repair proteins and ICL repair proteins including MLH1, MLH3, PMS1, and PMS2, FANCD2, and FANCI. However, some evidence suggests that FAN1 may be pivotal to ICL repair but not mismatch repair (MacKay et al., 2010). Rather than trinucleotide repeat somatic expansion operating exactly via the pathways of DNA mismatch repair, or interstrand cross-link repair, it seems more likely that there is a pathway specific to trinucleotide repeat instability which employs DNA repair proteins from the mismatch and other repair pathways. This is a topic of ongoing investigation.

The work on FAN1 which is described in this Chapter fed into ongoing work by our group which I assisted with, investigating the mechanism through which FAN1 variants modulate HD. Given the association between somatic instability of the CAG repeat, and HD onset and progression which has been discussed elsewhere (Chapters 3 and 4) we hypothesized that FAN1 also has a role in the stability of the CAG repeat. In work that is beyond the scope of this thesis, we demonstrated that increased FAN1 expression is significantly associated with delayed AAO in HD (Goold et al., 2019). This finding was based firstly on a Transcription Wide Association Study (TWAS) in which gene expression values were imputed from 452 dorsolateral prefrontal cortex samples from the Common Mind Consortium into the GeM GWAS of AAO in HD and the TRACK-HD and Registry HD Progression GWAS which I describe in Chapter 3, and secondly on the finding that FAN1 trends towards significance in the TRACK-HD cohort such that decreased FAN1 expression is associated with faster progression and earlier
onset (Goold et al., 2019, Hensman Moss et al., 2017a). Recent evidence demonstrates Fan1 protects against expansion of the CGG repeat tract in the Fmr gene in a mouse model of Fragile X (Zhao and Usdin, 2018). A similar stabilization of the HTT CAG repeat tract would reduce somatic expansion of the HTT CAG repeat tract and could underlie the effect of FAN1 on HD course. Colleagues found that FAN1 expression stabilizes the CAG repeat in U20S HTT exon 1 cells and regulates the stability of the endogenous HTT repeat in patient derived induced pluripotent stem cells (Goold et al., 2019). Of the FAN1 variants described in this chapter, the p.R507H FAN variant which according to the genetics (GeM-HD-Consortium, 2015) is associated with earlier disease onset has received particular interest. Goold et al found that the p.R507H FAN variant does not affect HTT CAG repeat stability in U20S cells (Goold et al., 2019): although U20S HTT exon 1 (118 CAG) cells expressing p.R507H showed reduced CAG repeat expansion rates compared to those expressing WT forms, which trended toward significance, these changes were likely related to differences in FAN1 expression levels. They also found that FAN1 associates with CAG repeats in HTT and other proteins, both with and without the pR507H variant (Goold et al., 2019). It may be that the assay systems used are not sensitive to pick up the small changes in activity that pR507H may engender.

The pathway analysis of the WES data, while not statistically significant, supported the findings of the pathway analysis of the GWAS studies (GeM-HD-Consortium, 2015, Hensman Moss et al., 2017b), and provides further support that common genes and biological processes influence both HD age-at-onset and disease progression.

My work described in Chapter 3 highlighted MSH3 as a genetic modifier of disease progression in HD (Moss et al., 2017), and it was recently identified as a modifier of somatic instability in DM1 (Morales et al., 2016). The index SNP in the chromosome 5 region of high signal in the GWAS of HD progression was rs55787476, an imputed SNP, located within a 9 bp tandem repeat sequence in exon 1 of MSH3, which is also in the 5’UTR of dihydrofolate reductase (DHFR) on the opposite strand (Tome et al., 2013a). The MSH3 and DHFR genes are arranged in a head-to-head orientation and share a common promoter that divergently drives transcription (Tome et al., 2013a).

The WES and Sanger sequencing analysis presented here suggest that rs557874766 is an alignment artefact and corresponds to a deletion corresponding to 3 alanines in the protein sequence relative the wild type version. At the protein level, in silico modelling predicts that
the deletion allele results in the loss of a surface α-helix (Kallberg et al., 2012) at the N-terminus of MSH3, as compared to the reference sequence.

The region of interest in MSH3 exon 1 is GC rich and repetitive as shown in Figure 5.7. Although the configuration of deletions shown in Figure 5.7 is what the deCODE platform and our UCL WES pipeline generates, the deleted region(s) can actually be placed in various different positions to get the same net sequence results, with the effect of generating different SNPs in the surrounding region (Figure 5.8A), thus rs557874766 can be viewed as an alignment artefact. The net effect at the protein level is a variable number of alanine residues (Figure 5.8B).

Interestingly a 9-bp repeat polymorphism in exon 1 of the MSH3 gene was previously described in the Japanese population which corresponds to the deletion we identified (Nakajima et al., 1995). The Japanese data suggest that the region is highly polymorphic, with variable numbers of repeats corresponding to between 3 and 7 Alanines reported. The most
common allele in the East Asian populations was 6 alanines and the second most common 7 alanines, while the 3 alanine version associated with slower HD progression in my study is also relatively common (Table 5.7). Subsequent work from our group which is beyond the scope of this thesis shows that the region can be viewed as being composed of a variable number of different 9bp blocks (Flower et al., 2019).

<table>
<thead>
<tr>
<th>Repeats</th>
<th>Sizes (bp)</th>
<th>N (total 116)</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>171</td>
<td>18</td>
<td>0.155</td>
</tr>
<tr>
<td>4</td>
<td>180</td>
<td>5</td>
<td>0.043</td>
</tr>
<tr>
<td>5</td>
<td>189</td>
<td>1</td>
<td>0.009</td>
</tr>
<tr>
<td>6</td>
<td>198</td>
<td>70</td>
<td>0.603</td>
</tr>
<tr>
<td>7</td>
<td>207</td>
<td>22</td>
<td>0.190</td>
</tr>
</tbody>
</table>

Table 5.7: Allelic sizes and frequencies at exon 1 of the hMSH3 gene in 58 unrelated Japanese individuals, from Nakajima et al (Nakajima et al., 1995).

In a review of published binding sites, I found that the MSH3 N-terminal region of interest is close to the Exo1 binding site at 75–297 according to one study (Schmutte et al., 2001) (Figure 5.9). Exo 1 also binds to MSH2 and MLH1 (Schmutte et al., 2001), while MSH2 binds MSH3 in two places (Guerrette et al., 1998) to form the MutSβ heterodimer (Figure 5.9). PCNA, a DNA clamp which acts as a scaffold to recruit proteins involved in DNA replication, repair and epigenetics binds MSH3 at the N-terminus close to the region of interest (Kleczkowska et al., 2001, Clark et al., 2000, Flores-Rozas et al., 2000, Finn et al., 2016) (Figure 5.9).
Figure 5.9: Figure of MSH3 showing the putative binding domains for proteins with which it interacts, and the ATP binding site. The location on the deletion we identified is also shown, illustrating that the Exo1 and PCNA binding sites are very close to the deletion.
Both Exo1 and PCNA are involved in DNA mismatch repair (Kleczkowska et al., 2001). PCNA is a sliding clamp that participates in DNA replication, but in MMR it delivers MSH proteins to mismatches and increases binding specificity (Flores-Rozas et al., 2000). Exonuclease 1 (EXO1) excises the daughter strand after mismatch recognition, as well as being involved in end resection during homologous recombination (Goellner et al., 2015).

Given the proximity of the repeat region to MMR protein binding domains (Figure 5.9), the deletion-containing allele may change the secondary structure (Figure 5.3) and alter MSH3 function in the recognition and repair of insertion-deletion loops, double strand breaks or single strand annealing (Lyndaker and Alani, 2009, Schmidt and Pearson, 2016). It is also possible that the deletion variant alters the important quaternary structure in MSH3’s binding to MSH2. As discussed in Chapter 4, repetitive DNA sequences form unusual secondary structures such as slipped strands, hairpin loops, G-quadruplexes and R-loops (Mirkin, 2007, Neil et al., 2017), the stability of which correlates with expansion (Gacy et al., 1995). MSH3 may recognise these structures (Owen et al., 2005) and initiate repair, during which out of register synthesis could result in repeat expansion (Neil et al., 2017, Khan et al., 2015), the presence of the deletion may alter the kinetics, resulting in reduced somatic expansion of the CAG tracts and slower HD progression: this hypothesis is under ongoing investigation.

It is interesting that a further MSH3 variant was highlighted in the case control analysis of the TRACK-HD WES (rs184967, pE949R in exon 21): while this result may be spurious given the small sample size, it may also point to some locus heterogeneity in MSH3, with more than one variant modulating the onset/progression of HD. Tome et al used different mouse strains to show that MSH3 polymorphisms and protein expression levels affect CAG repeat instability in HD mice, and suggest that the T321I variant may be responsible (Tome et al., 2013a): a different MSH3 variant to those discussed here and one in a different domain of the protein. The region surrounding the exon 1 27bp deletion is poorly conserved between species (Figure 5.6). There are increased numbers of alanines in the polyalanine section in apes more closely related to humans, while the number of ‘PPA’ seems to vary in a way that differs from their phylogenetic relatedness, perhaps suggesting that this represents an old polymorphism. The deletion variant which is associated with slower HD progression, (A)6 PPA PA, is the version described in Chimpanzees and Gorillas thus may be an ancestral version of the protein.

However, the lack of evolutionary constraint observed suggests functional redundancy in the MMR pathway and a lack of effect of variation at the MSH3 N-terminus outside of the context of a repeat expansion disease. Unlike other MMR components, germline heterozygous MSH3
mutations and *MSH3* depletion are not particularly associated with increased risk of cancer, most likely because MutSα (MSH2/MSH6) can also initiate repair at replication errors (Haugen et al., 2008, Edelmann et al., 2000, Jiricny, 2006). Therefore, modulation of MSH3 has significant therapeutic potential in a range of neurodegenerative diseases.
Chapter 6- C9orf72 repeat expansion disease: examination of intergenerational repeat stability and expansion of the known phenotype to encompass HD phenocopy presentations

6.1 Introduction

Both FTD and ALS are neuropathologically characterized by the presence of neuronal inclusions containing TDP-43 protein (Tar DNA binding protein-43), and commonalities between the two diseases have been increasingly appreciated (Karch et al., 2018). As outlined in the General Introduction, an expanded hexanucleotide GGGGCC repeat in the C9orf72 gene has been established as a major cause of both FTD and ALS (DeJesus-Hernandez et al., 2011, Renton et al., 2011, Smith et al., 2012, Mahoney et al., 2012). The mutation is intronic, in a highly conserved gene (DeJesus-Hernandez et al., 2011, Morris et al., 2012) which has homology with the DENN-like superfamily suggesting a role as regulator of membrane traffic (Levine et al., 2013, Zhang et al., 2012, Morris et al., 2012), and which may be involved in other neurological conditions (Friedland et al., 2012). Several hundred-thousands of repeats have been documented in pathogenic expansions (Beck et al., 2013). Elucidating the pathogenic mechanism of this expansion has generated much interest; several non-mutually exclusive possibilities exist (Mori et al., 2013, Reddy et al., 2013, Fratta et al., 2012, Lashley et al., 2013, Ash et al., 2013, DeJesus-Hernandez et al., 2011): 1) C9orf72 haploinsufficiency-expanded repeats interfere with transcription or translation of the gene, leading to decreased expression of C9orf72 protein; 2) RNA gain of function- RNA foci formed by sense and antisense transcripts of expanded repeats interact and sequester essential RNA binding proteins, causing neurotoxicity; 3) Repeat associated non-ATG initiated (RAN) translation of GGGGCC repeat expansion- RAN translation of expanded sense and antisense repeats produces potential toxic dipeptide repeat protein (DPR) (Ash et al., 2013, Lashley et al., 2013).

As discussed in Chapter 4, many diseases associated with expansions in sequences of repetitive DNA are characterised by intergenerational and somatic mosaicism of the repeat size. In work described in this chapter and published in Beck et al (Beck et al., 2013) I look at the intergenerational stability of GGGGCC repeats in families without disease-associated expansions.
As discussed in Chapter 1, HD is an autosomal dominantly inherited neurodegenerative condition typically characterised by a triad of psychiatric, movement and cognitive impairment. In many cases where HD is suspected clinically, patients lack the CAG repeat expansion that causes HD (Andrew S. E., 1994, Persichetti F., 1994, Wild, 2007, Huntington's et al., 1993). Such individuals are said to have HD phenocopy syndromes or HD-like disorders (Moore R. C., 2001). Wild & Tabrizi (Wild, 2007) reviewed genes identified in different HD phenocopy cohorts to determine that Spino cerebellar ataxia 17 (TBP) accounts for 1.1%, Huntington’s Disease-Like 2 (HDL2) for 0.7%, Friedreich’s ataxia (JPH3) for 0.35% and inherited prion disease (PRNP) for 0.24% of HD phenocopies. Testing for these mutations is now routinely performed; however the majority of HD phenocopy patients still do not attain a formal genetic diagnosis.

Given the established phenotypic variability of C9orf72 associated disease it was my aim in the study described in this chapter to examine whether the C9orf72 expansion is also a cause of HD phenocopy clinical presentations, and hence whether testing for it should be considered in the routine genetic work-up of this patient group. The results of this work have been published as Hensman Moss et al (Hensman Moss et al., 2014).
6.2 Materials and Methods

6.2.1 Standard Protocol Approvals, Registrations, and Patient Consents
Ethical approval to undertake these analyses was given by the local NHNN/ION ethics committee. Informed consent for genetic studies was obtained from all participants.

6.2.2 Case ascertainment: Control samples for intergenerational stability analysis
DNA samples were obtained from the Fondation Jean Dausset-Centre d’Etude du Polymorphisme Humain (CEPH) (Dausset et al., 1990): 802 individuals from 61 families in the CEPH family series were analysed to determine the size of repeat at the C9orf72 locus.

6.2.3 Case ascertainment: HD phenocopy subjects
As previously described (Wild et al., 2008), subjects were classified as having HD phenocopy syndromes on the basis of a clinical presentation consistent with HD when assessed by an experienced neurologist or neurogeneticist, and a negative test for the expanded CAG repeat in the HTT gene which causes HD (<36 repeats). At the Neurogenetics Unit of the National Hospital for Neurology and Neurosurgery (NHNN), London, UK, 63.5% of diagnostic HD tests (those done on symptomatic patients) are negative for HD. A cohort of 514 HD phenocopy cases who underwent negative diagnostic genetic testing for HD at NHNN were identified.

6.2.4 Clinical phenotyping
I reviewed clinical summaries for all cases, and reviewed all available clinical case notes for cases positive for the C9orf72 expansion mutation. Demographic data, family history, examination findings, first symptoms and age of onset were recorded. Where available, neuropsychometry reports were reviewed, and additional investigations were documented including electrophysiological assessments, MRI, CSF and tissue biopsies. HTT CAG repeat length was recorded. I used Fisher’s exact test (Stata software) to examine the relationship between the presence of particular clinical signs and gene test outcome.

I gave all C9orf72-positive cases a modified Goldman score (Goldman et al., 2005, Beck et al., 2008) (Table 6.1), which was used to quantify the strength of the autosomal dominant family history. Scoring was modified to give a score of 0 for no data, 4 for definitely no family history, and 4.5 for unknown or undescribed family history.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description of family history structure</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Autosomal dominant FHx, 3 affected+ in 2 generations with 1 as first-degree relative</td>
</tr>
<tr>
<td>2</td>
<td>Familial aggregation of 3 or more family members with dementia not meeting 1</td>
</tr>
<tr>
<td>3</td>
<td>1 other affected family member (AAO &gt;65)</td>
</tr>
<tr>
<td>3.5</td>
<td>1 other affected family member (AAO 65+)</td>
</tr>
<tr>
<td>4</td>
<td>Definitely no FHx</td>
</tr>
<tr>
<td>4.5</td>
<td>Unknown/ undescribed FHx</td>
</tr>
<tr>
<td>0</td>
<td>No data</td>
</tr>
</tbody>
</table>

Table 6.1: Modified Goldman scoring system. FHx: Family History. AAO: Age At Onset of symptoms.

6.2.5 Repeat primed PCR

To test for the presence of an expansion at C9orf72, I carried out repeat primed PCR (rpPCR) using the previously described methods (Renton et al., 2011). Specifically, 100 ng of genomic DNA were used as template in a final volume of 28 ml containing 14 ml of FastStart PCR Master Mix (Roche Applied Science, Indianapolis, IN, USA), and a final concentration of 0.18 mM 7-deaza-dGTP (New England Biolabs, Ipswich, MA, USA), 13 Q-Solution (to facilitate amplification of GC-rich templates) (QIAGEN, Valencia, CA, USA), 7% DMSO (Sigma-Aldrich), 0.9 mM MgCl2 (QIAGEN), 0.7 mM reverse primer consisting of four GGGGCC repeats with an anchor tail, 1.4 mM 6FAM-fluorescent labelled forward primer located 280 bp telomeric to the repeat sequence, and 1.4 mM anchor primer corresponding to the anchor tail of the reverse primer. A touchdown PCR cycling program was used where the annealing temperature was gradually lowered from 70°C to 56°C in 2°C increments with a 3 min extension time for each cycle.

The repeat-primed PCR is designed so that the reverse primer binds at different points within the repeat expansion to produce multiple amplicons of incrementally larger size. The lower concentration of this primer in the reaction means that it is exhausted during the initial PCR cycles, after which the anchor primer is preferentially used as the reverse primer (Renton et al., 2011).

I undertook fragment length analysis on an ABI 3730xl automated sequencer. Analysis of repeat primed PCR electropherograms was performed using Peak Scanner v1.0 (ABI). Expansions with a characteristic ‘saw-tooth’ pattern were identified and put forward for Southern blotting.
To determine the size of the expansion in those identified as having it, fluorescent-labelled PCR was followed by fragment-length analysis on an ABI 3730xl automated sequencer (Beck et al., 2013). The PCR used 20 ng gDNA in FastStart PCR master mix (Roche Applied Science) supplemented with 13 Q solution (Roche Applied Science), 5% dimethyl sulphoxide, 0.2 mM 7-deaza-2-deoxy guanosine triphosphate, and 1 mM MgCl2 in a 20 ml final volume. Thermal cycling included initial denaturation for 5 min and 35 subsequent cycles of 30 s denaturation at 95°C, 30 s annealing at 60°C, and 1 min elongation at 72°C.

6.2.6 rs3849942 genotyping
DeJesus-Hernandez et al described a surrogate marker rs3849942 associated with an increased risk of mutation (DeJesus-Hernandez et al., 2011, Beck et al., 2013). Samples were genotyped for this SNP by allelic discrimination using the 5’ nuclease assay in conjunction with Minor Groove Binding (MGB) probes. The assay was performed on the SDS7500 Fast Real Time PCR system (ABI) and genotyping calls were made using software v2.0.6. An introduction to SNP genotyping is given in Chapter 2.

6.2.7 Microsatellite genotyping
Microsatellite analysis was performed using ten markers spanning approximately 13.1Mb of genomic DNA centred around the C9orf72 gene (Beck et al., 2013). PCR amplicons were generated using fluorescently end labelled primers at 500nM for microsatellite markers D9S1814(VIC), D9S976(FAM), D9S171(NED), D9S1121(VIC), D9S169(FAM), D9S263(HEX), D9S270(FAM), D9S104(FAM), D9S147E(NED) and D9S761(FAM) in MegaMix Royal hot start cocktail (Microzone). Thermal cycling conditions included an initial preheat at 95°C for 5 minutes, followed by 35 cycles of 95°C 30”, 58°C 40”, 72°C 1’. A loading mix of 1μl amplicon diluted 1:50 in ddH2O, 9.5μl HiDi formamide (ABI) and 0.5μl 500LIZ size standard was prepared and DNA products were electrophoresed on an ABI 3730xl automated sequencer. Data was analysed using ABI GeneMapper software v4.0 (Applied Biosystems (ABI)).

6.2.8 Southern hybridisation
A recently described Southern hybridisation protocol was used by my collaborator Mr Mark Poulter to estimate expansion size (Beck et al., 2013). This combined the use of an oligonucleotide (GGGGGC)5 probe which targets multiple sites within the expansion and genomic DNA (gDNA) digested with two frequently cutting restriction endonucleases whose sites closely flanked the repeat region. Hexanucleotide repeat number was estimated by interpolation of autoradiographs using a plot of log10 base pair number against migration distance which was created in Microsoft Excel.
6.3 Results

6.3.1 C9orf72 repeat intergenerational instability is seen in those with longer repeat lengths

I examined the stability of the size of the C9orf72 hexanucleotide repeat region in 802 individuals from 61 families in the CEPH family series, a panel of reference families which has proved an important resource for the characterization of DNA polymorphisms and the construction of the human genetic map (Dausset et al., 1990). No large expansions (>30 repeats) were identified via repeat primed PCR. In 1,046 transmissions, three changes in repeat size between generations were identified. In the CEPH families, the largest repeat (22 repeats) changed size twice in the same family: from 21 in the paternal grandparent to 22 in the father and from 22 in the father to 20 in the son (Figure 6.1). There were no unstable maternal transmissions. The overall intergenerational repeat change rate was 0.29%. Interestingly, all intergenerational changes occurred from a starting repeat length > 10. These changes were verified by repeat rpPCR and fluorescent-labeled PCR size fractionation (although alteration of flanking sequences cannot be excluded).
**Figure 6.1:** Fragment analysis of CEPH families with inter-generational repeat slippage. Data from fluorescent labelled PCR followed by fragment length analysis on an ABI 3730xl automated sequencer from 2 CEPH families showing evidence of inter-generational repeat slippage. For clarity, the numbers of base pairs of alleles demonstrating slippage are also shown with repeat size in red text. (a) CEPH family 1423 results showing slippage from paternal grandfather’s 21 repeats up to father’s 22 and then down to 20 repeats in his son. (b) CEPH family 1420 showing slippage from father’s 11 repeats to his daughter’s 12 repeats.

### 6.3.2 Identification of C9orf72 expansion in HD phenocopy cases

Of the 514 HD phenocopy cases screened, 10 probands (1.95%, 95% CI 1-4) were positive for the C9orf72 expansion, making this mutation the commonest identified cause of HD phenocopy syndromes in a UK cohort (Wild et al., 2008, Hensman Moss et al., 2014). No C9orf72-positive cases had intermediate sized HD CAG repeats in the Huntingtin gene, and there was no correlation between the larger HD normal allele and age of onset.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age at onset</th>
<th>Rs3849942 genotype</th>
<th>Expansion size estimated by southern hybridisation</th>
<th>Goldman score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60</td>
<td>AA</td>
<td>4010</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>GA</td>
<td>3441</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>55</td>
<td>AA</td>
<td>3682</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>AA</td>
<td>3180</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>GA</td>
<td>2939</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>GA</td>
<td>2939</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>GA</td>
<td>3186</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>GA</td>
<td>3518</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>19</td>
<td>AA</td>
<td>insufficient DNA</td>
<td>4.5</td>
</tr>
<tr>
<td>10</td>
<td>58</td>
<td>GA</td>
<td>insufficient DNA</td>
<td>3</td>
</tr>
</tbody>
</table>

*Table 6.2:* Age at onset and genetic results of C9orf72 expansion positive cases

Southern hybridisation (Table 6.2 and Figure 6.2) of 8/10 subjects for whom there was sufficient DNA demonstrated that the size of expansion in this HD Phenocopy case series was not significantly different from that found in series with other clinical presentations of the
C9orf72 expansion (Beck et al., 2013). There was no significant difference in expansion size between those with and without chorea/dystonia.

**Figure 6.2**: Southern Blot of eight HD phenocopy patient DNAs, blot produced by Mark Poulter. Southern Blot of eight HD phenocopy patient DNAs, showing that C9orf72 repeat expansions can be seen in all cases. The asterisk indicates a GGGGCC containing a short-tandem-repeat genome motif unrelated to C9orf72. The samples are ordered from 1 – 8 from left to right; there was insufficient DNA to blot samples 9 and 10.

### 6.3.3 Presence of risk haplotype in those with expansion mutations and with intergenerational repeat instability

Previous reports have linked the C9orf72 expansion mutation with the rs3849942 A allele: all individuals with the expansion were either heterozygous or homozygous for rs3849942 A (DeJesus-Hernandez et al., 2011, Beck et al., 2013). Genotyping of the C9orf72-positive HD phenocopy cases demonstrated that all were heterozygous or homozygous for the rs3849942 A allele, thus our data are consistent with previous reports (DeJesus-Hernandez et al., 2011) (Table 6.2).
In order to examine whether all expansion positive cases share an ancient common ancestor all cases in our lab were tested for 10 microsatellites over 13.1 Mb surrounding C9orf72. I examined these microsatellites on my HD phenocopy cases, these data supported the finding of a lack of association between the risk associated SNP rs3849942 and any microsatellite marker. In one case the microsatellite data supported findings from the sizing assay that the case was homozygous for the C9orf72 expansion. This case (case 4) is discussed in detail in Fratta et al, which I am a co-author on (Fratta et al., 2013).

6.3.4 Clinical data

The average age at onset in this cohort was 48.8 years in those with precise onset data (SD 19.3, N=176). 300 subjects were seen at NHNN, 214 at other hospitals. Of those seen at NHNN, 45.3% were seen by a Movement Disorders Consultant, 15.3% by a Cognitive Disorders Consultant, 14.3% by a Neurogenetics Consultant and 25% by other Consultant Neurologists. Of the entire cohort, 19.5% had a family history of similar neurodegenerative disease whereas 70% of C9orf72-positive cases had a positive family history (see Goldman scores, Table 6.1 & 6.2). These results suggest that there is a predominance of those with family history, but sporadic C9orf72-positive cases may be possible.

Of the C9orf72-positive cases the mean age of onset was 42.7 years, range 8-60. Early psychiatric and behavioural problems were common; they were the first recorded symptoms in six of the cohort. Depression occurred in four, obsessions in two, apathy in two and psychosis in two cases.

Movement disorders were a prominent feature - three exhibited chorea, four dystonia, four myoclonus and three tremor (Table 6.3). Six of the ten subjects had rigidity and five bradykinesia. Chorea was observed periorally in one, was generalised with predominant head and arm involvement in one, and in the left arm and leg in another. Of the four subjects with dystonia, three were observed to have torticollis. In four of the ten subjects upper motor neuron signs were noted; lower motor neuron signs were not observed in any. Cognitively, executive dysfunction was noted in six subjects, and memory impairment was present in six; in subject 6 for whom limited history was available, ‘cognitive impairment’ was noted.

Of eight cases with available MRI reports four had generalised atrophy.
Case 4 was found to be homozygous for the *C9orf72* expansion mutation and has been described in detail in Fratta *et al* (Fratta *et al.*, 2013).

<table>
<thead>
<tr>
<th>Case</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical feature</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chorea</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myoclonus</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dystonia</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tremor</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rigidity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradykinesia</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torticollis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UMN signs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apathy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Executive dysfunction</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Impaired memory</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Impaired face recognition</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Impaired verbal fluency</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Table 6.3: Summary of the clinical features of ten C9orf72 expansion-positive cases. UMN = upper motor neuron.*

### 6.3.5 Comparisons between C9orf72 positive cases and the rest of the HD phenocopy cohort

To examine whether there are particular HD phenocopy cases in whom *C9orf72* testing should be prioritized, I compared the frequencies of symptoms and signs between the whole cohort and those with the expansion (Table 6.4). Fisher’s exact test was performed to investigate association between each clinical feature and the outcome of the *C9orf72* genetic test. The presence of cognitive and psychiatric features, and some movement disorder features (dystonia, bradykinesia/rigidity, tremor, myoclonus and upper motor neuron features), were significantly associated with a positive *C9orf72* test (Table 6.4). Though there may be some degree of ascertainment bias as more clinical detail was recorded for positive cases, it remains clear that many symptoms characteristic of HD phenocopies are associated with a *C9orf72* gene expansion.
<table>
<thead>
<tr>
<th>Phenotypic feature</th>
<th>Number in C9orf72 negative cases (N=504) (Percentage)</th>
<th>Number in C9orf72 positive cases (N=10) (Percentage)</th>
<th>Number in whole HD phenocopy cohort (N=514) (Percentage)</th>
<th>P value (Fisher's exact test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All movement disorder features</td>
<td>394 (78%)</td>
<td>8 (80%)</td>
<td>402 (78%)</td>
<td>1</td>
</tr>
<tr>
<td>Chorea</td>
<td>154 (31%)</td>
<td>3 (30%)</td>
<td>157 (31%)</td>
<td>1</td>
</tr>
<tr>
<td>Dystonia</td>
<td>53 (11%)</td>
<td>4 (40%)</td>
<td>57 (11.1%)</td>
<td>0.017</td>
</tr>
<tr>
<td>Bradykinesia/ rigidity</td>
<td>78 (15%)</td>
<td>6 (60%)</td>
<td>84 (16%)</td>
<td>0.002</td>
</tr>
<tr>
<td>Tremor</td>
<td>39 (8%)</td>
<td>3 (30%)</td>
<td>42 (8%)</td>
<td>0.041</td>
</tr>
<tr>
<td>Ataxia</td>
<td>72 (14%)</td>
<td>1 (10%)</td>
<td>73 (14%)</td>
<td>1</td>
</tr>
<tr>
<td>Myoclonus</td>
<td>31 (6%)</td>
<td>4 (40%)</td>
<td>35 (7%)</td>
<td>0.003</td>
</tr>
<tr>
<td>UMN features</td>
<td>18 (4%)</td>
<td>4 (40%)</td>
<td>24 (5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>LMN features</td>
<td>8 (1.6%)</td>
<td>0 (0%)</td>
<td>8 (2%)</td>
<td>1</td>
</tr>
<tr>
<td>Psychiatric problems</td>
<td>53 (11%)</td>
<td>7 (70%)</td>
<td>60 (12%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Depression</td>
<td>17 (3%)</td>
<td>4 (40%)</td>
<td>21 (4%)</td>
<td>0.035</td>
</tr>
<tr>
<td>Anxiety</td>
<td>4 (0.8%)</td>
<td>2 (20%)</td>
<td>6 (1%)</td>
<td>0.005</td>
</tr>
<tr>
<td>Cognitive impairment</td>
<td>167 (33%)</td>
<td>9 (90%)</td>
<td>176 (34%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Executive dysfunction</td>
<td>19 (4%)</td>
<td>6 (60%)</td>
<td>25 (5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Memory problems</td>
<td>29 (6%)</td>
<td>9 (90%)</td>
<td>176 (34%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Family history</td>
<td>98 (19%)</td>
<td>7 (70%)</td>
<td>105 (20%)</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table 6.4: Phenotypic features of C9orf72 negative & positive cases within HD phenocopy cohort, and outcome of Fisher's exact test to test for association between clinical feature and genetic test outcome.

6.3.6 An illustrative case
Case 5, a right-handed Caucasian woman, had a normal birth and development and was university educated. She worked in a professional job and was well until a sudden bereavement when she was fifty after which she became depressed. At around 55y increasing fatigue was noted and she had her first falls, initially backwards. She stopped working, and developed a change in personality with decreased interest in her environment and child-like behaviour. She developed hypophonia and slurred speech. By 58y she was having difficulty mobilizing and within 12 months went from independent-living to being mute, profoundly bradykinetic and requiring a hoist to transfer. She developed dystonic posturing of her feet and hands, and involuntary movements and a tremor in her lower limbs.

In her family history, her father died of dementia without motor problems aged 69y. She was admitted to hospital for investigation aged 60y. On examination there was akinetic mutism with marked axial rigidity. There was left laterocollis, minor right torticollis, perioral movements and occasional right cheek movements. There was broken pursuit and slow broken saccades. There was moderate rigidity with spasticity in the upper limbs and severe rigidity in the lower limbs. Plantars were extensor. Palmomental and pout reflexes were present. There was perseveration and frontal features. MMSE (mini mental state examination) was 16/25.

There was no other significant medical history. Blood tests did not reveal any haematological, biochemical, endocrine, immunological or infective cause of the presentation. CSF was unremarkable; CSF specific proteins: 14-3-3 negative, S100 0.19, Tau 169, A-beta 1-42 313. MRI brain the year prior to admission showed small vessel disease only. CT brain: generalised volume loss of cerebrum and cerebellum, with no specific predilection and mild-moderate small vessel disease. Electroencephalography: normal background rhythm. Dopamine Transporter imaging/ DAT scan: suboptimal study.

6.3.7 An unusual case

Case 7, a right-handed Caucasian man, had a normal birth and early development. Aged three at nursery school, it was noted that he did not mix well with the other children. At primary school aged five he was found to have slight difficulties with writing; aged six he was unable to follow basic lessons. Soon thereafter he was seen by an educational psychologist and was diagnosed as having moderate learning difficulties and was transferred to special needs school.
By age 8y, he had abnormal movements under stress, particularly affecting his hands and head. These became a lot more prominent from 21y when they affected his walking. Occasionally his right leg was noted to jerk uncontrollably from under him, and he had some falls. The ‘fidgeting’ and jerking movements of hands and neck deteriorated. From 21y he had increased frustration and aggression.

His parents are non-consanguineous. His maternal grandmother died of motor neuron disease; both parents were well.

Aged 23y he was admitted to hospital for investigation. Gait was slightly broad based, with both arms tending to hold slightly dystonic postures, particularly on the right. There was decreased arm swing, nuchal more than axial rigidity, unsteadiness on heel-toe walking, and Romberg’s test was negative. Eye movements were abnormal, with poor gaze initiation, impaired pursuit, saccadic hypometria with head thrusts, and reduced vertical up-gaze. There was generalised chorea with mainly head and arm involvement, oro-buccal chorea, myoclonic movements of the head and neck, and some additional dystonic elements with mild bradykinesia. In the limbs there were prominent irregular myoclonic jerks, exacerbated by movement and stimuli. Reflexes and sensation were normal.

MMSE was 20/28. On Neuropsychological examination, the Wechsler Adult Intelligence Scale-Revised was within the defective range consistent with learning difficulties. There was evidence of memory impairment for visual and verbal memory.

MRI scan showed one small lacune. Nerve conduction studies and electromyography were normal. Electroencephalography revealed a diffuse and non-specific excess of theta activity with only a trace of alpha like activity. Although the bursts of high voltage slow activity had a bursting paroxysmal quality no definite epileptiform activity was seen. A very extensive set of blood tests including white cell enzymes, amino acid profiling did not reveal any haematological, biochemical, endocrine, immunological or infective cause of the presentation.

Genetic testing excluded mitochondrial mutations, DRPLA and HD, and karyotyping was normal. Cerebrospinal fluid, skeletal muscle biopsy, axillary skin biopsy, blood films, bone marrow aspirate and trephine analysis were all unremarkable.

6.3.8 A homozygous case
As described more extensively in (Fratta et al., 2013) clinically this patient developed early-onset fronto-temporal dementia, thus the presentation was severe, but not out of the normal range of presentations for C9orf72. Neuropathological analysis showed c9FTD/ALS characteristics, with abundant p62-positive inclusions in the frontal and temporal cortices, hippocampus and cerebellum, as well as less abundant TDP-43-positive inclusions.

6.4 Discussion

In this chapter I have presented a large case series which not only demonstrates that the C9orf72 expansion is the most frequent cause of HD phenocopy presentations in this UK-based population, but also that the phenotype of the C9orf72 encompasses a diversity of movement disorders, and a younger age of onset than previously recorded. Interestingly, and tying into the work above on somatic instability in repeat disorders, I also found that intergenerational repeat instability of the unexpanded C9orf72 repeat occurred on the same haplotype background, and in alleles with high normal numbers of repeats. The question of whether it is the haplotype, repeat or the interaction of the two which drives the instability warrants further investigation. It seems likely that it is the repeat size that is important here, given that in Huntington’s disease there is a relationship between CAG repeat size and propensity for intergenerational expansion in HD: sperm from individuals with CAG repeats of 34 and 35 were at higher risk of expansion than those with lower intermediate repeats (Semaka et al., 2013). Also that larger repeat tracts are known to be more susceptible to somatic expansion (Williams and Surtees, 2015b, Veitch et al., 2007, Pluciennik et al., 2013), and it is plausible that at least some of the mechanisms that lead to somatic instability and intergenerational stability are shared: this would be another interesting topic for future study.

In those in whom HD is suspected, but patients do not have a CAG repeat expansion in HTT, attaining genetic diagnosis has been rare (2.8% (Wild et al., 2008)). The data presented in this chapter demonstrate that the C9orf72 expansion is the commonest-identified genetic cause of HD phenocopy presentations in this UK cohort, with a prevalence of 1.95% (95% CI 1-4).

HD is an autosomal dominant condition, classically presenting with a triad of movement, cognitive and psychiatric symptoms. However there is clinical heterogeneity, particularly early in disease, and not all characteristic features may be apparent: 90% of adults with HD develop chorea, but the clinical spectrum is broad, including Parkinsonian akinetic-rigid syndromes and relatively pure dystonic, ataxic and psychiatric presentations (Bates, 2002). Around 8% of patients with HD present without an apparent family history of HD (Schneider et al., 2007). Because of this clinical diversity, it is accepted (Wild, 2007, Wild et al., 2008) that any definition of Huntington’s disease phenocopy syndromes need encompass not only the
classical triad of HD but also syndromes having a major degree of overlap with HD, and those without a known autosomal dominant family history. Those patients with a clear family history of HD and with classical manifest HD are more likely to have HD, however many patients seen by Neurologists do not present in such a clear cut manner. Our cohort is composed of patients seen by experienced neurologists in whom the diagnosis of HD was considered thus it reflects clinical reality. It is UK-based, and given that UK-based cohorts have similar ethnic descent to other European, Australian and North American cohorts, our findings are likely to be representative of cohorts from these areas. In patients of African origin (particularly Southern Africans), JPH3 expansion remains the commonest cause of HD-like presentations (Magazi et al., 2008).

Identifying the causes of HD phenocopy syndromes is of importance to the diagnosis and management of patients with these presentations, as well as the counselling of such individuals and their relatives in matters of genetic testing, life choices and reproduction (Wild, 2007). Diagnostic tests for the C9orf72 mutation are now available. Many symptoms characteristic of HD were associated with the subject being C9orf72 positive; given this, and the high frequency of C9orf72 expansion among HD phenocopies mean that we believe that it should be tested for in all HD phenocopy cases. In the future it is likely that multi-gene ‘disease panels’ will supersede the need for sequential genetic testing, however since C9orf72, like many other causes of HD phenocopies is an expansion mutation, it will remain important for the clinician to be aware of which tests are most appropriate for different patients and request them accordingly. In view of our findings I proposed a revised clinico-genetic algorithm for the investigation of HD phenocopy cases, shown here in Figure 6.3 (Hensman Moss et al., 2014).
Figure 6.3: Algorithm for the investigation of HD phenocopy cases. Proposed clinico-genetic algorithm for the work-up of Huntington’s disease phenocopy patients, highlighting key diagnoses to be considered. SCA, Spinocerebellar ataxia; HDL2, Huntington’s disease-like 2, DRPLA, dentatorubral-pallidoluysian atrophy; NBIA, Neurodegeneration with Brain Iron Accumulation. (Produced by me after (Wild, 2007), published in (Hensman Moss et al., 2014), image reproduced with permission of the rights holder, the American Academy of Neurology.

The effects of the C9orf72 expansion are known to be both clinically and pathologically varied (Murray et al., 2011) and it is the major cause of both familial and sporadic ALS and FTLD, which are themselves phenotypically heterogeneous conditions. Parkinsonism, particularly rigidity and bradykinesia, has been previously noted in C9orf72-positive individuals (O’Dowd et al., 2012, Dejesus-Hernandez et al., 2012, Boeve et al., 2012); the C9orf72 mutation has been found in some cohorts of patients with Parkinson’s disease (Lesage et al., 2013) and not others (Yeh et al., 2012, Akimoto et al., 2012, Dejesus-Hernandez et al., 2012). In this study we have demonstrated that the clinical phenotypes caused by C9orf72 expansion mutations are broader than previously noted to date. It can present with a movement disorder including chorea, dystonia, myoclonus and tremor. The combination of movement disorder, cognitive decline and psychiatric and behavioural problems, often with a family history of similar
problems, explains why \textit{C9orf72}-positive cases can have a presentation very similar to HD. It is notable that ALS-type symptoms were relatively infrequent in the HD phenocopy \textit{C9orf72} cases: none had lower motor neuron signs, while 40\% had upper motor neuron signs. By contrast, symptoms more characteristic of FTLD such as cognitive impairment were much more prevalent, suggesting that there is more overlap between the HD-like and FTLD-like cases.

The average age of onset for \textit{C9orf72} in published reports is around 57 years (Mahoney et al., 2012, Renton et al., 2011, Majounie et al., 2012, Boeve et al., 2012), in this study it is lower at 42.7 years, with range 8 – 60, suggesting that the condition should be considered in the differential diagnosis not only in a wider range of clinical presentations, but in a wider demographic group than previously identified.

We examined whether the difference in phenotype could be accounted for by a different size of expansion by Southern hybridisation: the size of expansion in our HD phenocopy cohort was not significantly different from that of other cohorts (Beck et al., 2013). Furthermore, among the 8 \textit{C9orf72}-positive subjects examined here, there is no statistically significant association between expansion size and age of onset. Case 7, who had motor onset at 8y, underwent whole-exome sequencing; no large-scale structural abnormalities were detected. An important caveat is that there is evidence of reduced penetrance of the \textit{C9orf72} expansion given that the population frequency of \textit{C9orf72} expansion is 1 in 691 (Beck et al., 2013) in the UK population, so there is a small possibility of false positives accounting for one or more of these unusual presentations of \textit{C9orf72} mutations.

Among the ten HD phenocopy \textit{C9orf72} cases, there was a tendency for those with chorea and dystonia to have younger ages of onset than those without them: the average age of onset of subjects with chorea/ dystonia in this cohort is 28.3, whereas the average age of onset of those without them is 54.8 (P=0.019, Independent samples Mann-Whitney U-test). This may reflect our ascertainment criteria, since HD-phenocopy cases are more likely to be young and have movement disorders than FTLD or ALS cases. However, it is possible that the \textit{C9orf72} expansion with these motor symptoms manifests with earlier onset.

Incomplete penetrance has been previously suggested in \textit{C9orf72} expanded individuals (Pamphlett et al., 2012, Friedland et al., 2012, Boeve et al., 2012) which has important implications for genetic testing. In this case series there was no reported family history in
three cases, and case 7’s family history is compatible with incomplete penetrance – the subject’s maternal grandmother had MND, but the mother was well.

One of the subjects discussed in this study was found to be homozygous for the \textit{C9orf72} expansion mutation. Given that the clinical phenotype of this subject was within the range of other presentations of the disease, rather than much more severe or completely different this supports a gain-of-function mechanism being operational in \textit{C9orf72} expansion disease.
Chapter 7: Investigations of the effect of disease status, stage and rate of progression on the transcriptome in Huntington’s disease

7.1 Introduction

As discussed in Chapter 1, HD manifests itself clinically with various changes within an individual which develop over time including motor, cognitive and psychiatric dysfunction. However, how the expanded CAG repeat in the huntingtin protein effects downstream biology and results in the disease phenotype is incompletely understood. One area of interest is the ways in which the transcriptome is altered in HD - the expression pattern of proteins within a cell. Transcriptional dysregulation is a central feature of HD pathogenesis (Hodges, 2006). Expression levels of specific genes, differential splicing and allele-specific expression of transcripts can be accurately determined by RNA sequencing (RNA-Seq) experiments, and it also can quantify low expressed transcripts ([Reviewed in Wang et al., 2009]). Studies using RNA-Seq have already altered our view of the extent and complexity of eukaryotic transcriptomes (Hensman Moss et al., 2017b).

HD research has largely focused on the brain due to the presence of characteristic mutant huntingtin protein aggregates in the brain (Bates et al., 2015), and because the prominent symptoms and signs can be linked to neurodegeneration in the basal ganglia and cerebral cortex (van der Burg et al., 2009). However, mutant HTT is ubiquitously expressed (Trottier et al., 1995) and mounting evidence suggests it has direct effects in peripheral tissues (van der Burg et al., 2009, Carroll et al., 2015), though whether these effects are distinct, or parallel those in the brain remains unclear. Clinically, HD patients demonstrate peripheral immune dysfunction pre-symptomatically (Tai et al., 2007, Bjorkqvist et al., 2008, Kwan et al., 2012c, Träger et al., 2015), as well as weight loss that leads to cachexia with advancing disease (Carroll et al., 2015). There is progressive muscle wasting (Busse et al., 2008), endocrine dysfunction (Saleh et al., 2009) liver impairment (Carroll et al., 2015), cardiac dysfunction (Lanska et al., 1988, Mihm et al., 2007, Pattison et al., 2008). Mutant HTT protein aggregates can be found in the peripheral tissues of HD mice (Orth et al., 2003), as well as advanced patients (Turner et al., 2007). These peripheral features may contribute to CNS pathology, disease progression and mortality (Carroll et al., 2015, van der Burg et al., 2009), and strongly suggest that HD is a systemic disorder.

Peripheral tissues have the research advantage that they can be sampled minimally invasively and inexpensively from living patients, enabling longitudinal study throughout disease course.
This is in contrast to brain tissue, the availability of which is limited and is mostly from post-mortem subjects with end-stage disease (Montanini et al., 2013, Tomita et al., 2004). While blood has been used for transcriptomic studies, studies of gene expression changes in HD blood have been inconsistent. Using microarray technology, Borovecki et al. (2005) identified 12 upregulated transcripts, seven of which were also upregulated in brain. However, subsequent studies did not replicate these results (Runne et al., 2007, Lovrecic et al., 2009, Mastrokolias et al., 2015). Using tag-based serial analysis of gene expression (SAGE) in blood, Mastrokolias et al. (2015) found 167 genes differentially expressed by motor score, 40 of which had previously been reported in at least one microarray study.

We therefore conducted a transcriptomic analysis of whole blood in human HD using RNA-Seq. We studied differential expression of individual gene transcripts and enrichment of differential expression in gene sets in two independent cohorts from Track-HD (Tabrizi et al., 2009b) and Leiden, looking for transcriptional signatures relating to disease status and disease stage. We then investigated whether transcriptional changes seen in blood parallel those from previous studies in HD brain.

One of the main aims of this thesis is to identify factors which modulate the progression of Huntington’s disease, and the identification of genetic variants modulating progression is discussed in Chapters 3-5. However, many genetic variants of small effect are likely to be regulatory rather than coding variants. In addition many non-genetic effects are likely to manifest in altered gene expression mediated by epigenetic changes (Lee et al., 2013, Horvath et al., 2016, Roubroeks et al., 2017, Majewski and Pastinen, 2011, Feil and Fraga, 2011). Therefore, in addition to investigating disease status related changes in transcript levels, in the second part of this chapter I discuss work to investigate whether there are transcriptional changes which correlate with the rate of disease progression in HD.

Regarding my involvement in the work in this chapter, much of the work was collaborative, and I was involved in this study from inception. I was responsible for the reporting and collaborative working within the Neuromics Consortium which funded the work. I selected the subjects to sequence, obtained permissions to use their samples, and arranged for it to be shipped to DeCODE in Iceland for sequencing. I was involved in setting up collaborations with Dr Vincent Plagnol and his postdoctoral assistant Dr Kitty Lo for bioinformatics support, and was involved in the discussions about the analysis plan with them. I helped set up the collaboration with Dr Willeke van Roon-Mom at Leiden University Medical Centre. Pathway and gene co-expression analysis was conducted by collaborators at Cardiff University led by
Professor Peter Holmans. I assisted with writing the manuscript reporting the data for the effect of disease status and stage, with the help of Dr Michael Flower and collaborators as I was on Maternity leave while we were writing the final manuscript, this paper was published in Scientific Reports (Hensman Moss et al., 2017a). Bioinformatic, pathway and gene co-expression analysis was conducted by collaborators. The material examining the role of disease progression on the transcriptome in HD has not been published elsewhere and much of the work is my own.

7.2 Materials and methods

All experiments we performed in accordance with the Declaration of Helsinki and approved by the University College London (UCL)/UCL Hospitals Joint Research Ethics Committee and the LUMC IRB. Peripheral blood samples were donated by genetically-diagnosed HD patients and controls, and all subjects provided informed written consent.

7.2.1 Cohorts

The Track-HD cohort is described in General Methods, Chapter 2. I pre-selected a representative sample from the Track-HD study (Table 7.1), to assure a wide range of disease risk, severity and progression rate. Control subjects were age and gender matched to individuals in the premanifest and manifest groups, and selected from spouses or partners to ensure consistency of environments. The sample from the Track-HD cohort consisted of 54 premanifest gene carriers, 63 manifest HD subjects and 23 controls. Manifest subjects demonstrated motor abnormalities that were unequivocal signs of HD, as rated by the assessor and supported by total motor scores (TMS) over 5 on the Unified Huntington’s Disease Rating Scale (UHDRS). Premanifest gene carriers had a burden of pathology score (age x [CAG – 36.5]) (Penney et al., 1997) greater than 250, and a TMS of 5 or lower and a diagnostic confidence score (DCS) less than 4 on the UHDRS (Group, 1996a), indicating no substantial motor signs (Tabrizi et al., 2009b). The unified Huntington’s disease progression score (Chapter 2) was used to select TRACK-HD subjects with fast, average and slow progression from both the Premanifest and manifest HD groups (Table 7.1) in order to get the maximum phenotypic range for analysis. Age and clinical scores at the time of blood collection were used in this analysis.

Through a collaboration I helped set up with Willeke van Roon-Mom (Leiden University Medical Centre, LUMC) as a part of the European Commission funded Neuromics project we had access to LUMC samples. The LUMC cohort (Mastrokolias et al., 2015) consisted of 18 premanifest gene carriers, 56 manifest HD subjects and 27 age and gender-matched controls.
Motor onset was determined by an experienced neurologist using the same UHDRS standard as in TRACK-HD. All premanifest carriers showed no substantial motor signs, with a TMS of 5 or less and a UHDRS diagnostic confidence level less than 4. All controls were free of known medical conditions. Blood sample collection protocols, storage, and analysis methods, described below, were identical for the two cohorts.

In order to conduct progression analysis on the LUMC samples, a novel cross sectional severity based progression measure was developed to assess progression within this cohort. Given this progression score was developed for this purpose, the full range of progression scores are reflected in this study.
<table>
<thead>
<tr>
<th>Cohort</th>
<th>Group</th>
<th>N</th>
<th>Mean age, y ± SD (range)</th>
<th>Gender (male/female)</th>
<th>Mean (CAG)n length ± SD (range)</th>
<th>Mean TMS ± SD (range)</th>
<th>Mean TFC ± (range)</th>
<th>Rate of progression (n)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TRACK-HD</td>
<td>HD</td>
<td>112</td>
<td>46 ± 10 (22-64)</td>
<td>50/62</td>
<td>44 ± 3 (39-59)</td>
<td>14 ± 13 (0-45)</td>
<td>12 ± 2 (7-13)</td>
<td>Fast 16  Average 14  Slow 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Premanifest</td>
<td>50</td>
<td>42 ± 9 (22-64)</td>
<td>24/26</td>
<td>43 ± 3 (39-52)</td>
<td>2 ± 2 (0-8)</td>
<td>13 ± 0 (12-13)</td>
<td>Fast 16  Average 14  Slow 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manifest</td>
<td>62</td>
<td>48 ± 10 (23-64)</td>
<td>26/36</td>
<td>44 ± 3 (39-59)</td>
<td>23 ± 11 (5-45)</td>
<td>11 ± 2 (7-13)</td>
<td>Fast 16  Average 14  Slow 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>22</td>
<td>45 ± 5 (34-53)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LUMC</td>
<td>HD</td>
<td>74</td>
<td>53 ± 11 (29-79)</td>
<td>34/40</td>
<td>44 ± 3 (39-53)</td>
<td>32 ± 31 (0-102)</td>
<td>8 ± 5 (0-13)</td>
<td>Fast 16  Average 14  Slow 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Premanifest</td>
<td>18</td>
<td>46 ± 10 (29-63)</td>
<td>5/13</td>
<td>42 ± 2 (39-47)</td>
<td>3 ± 2 (0-5)</td>
<td>12 ± 1 (10-13)</td>
<td>Fast 16  Average 14  Slow 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manifest</td>
<td>56</td>
<td>55 ± 11 (35-79)</td>
<td>29/27</td>
<td>44 ± 3 (39-53)</td>
<td>42 ± 30 (6-102)</td>
<td>7 ± 5 (0-13)</td>
<td>Fast 16  Average 14  Slow 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>27</td>
<td>43 ± 11 (26-65)</td>
<td>13/14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>HD</td>
<td>186</td>
<td>48 ± 11 (22-79)</td>
<td>84/102</td>
<td>44 ± 3 (39-59)</td>
<td>21 ± 24 (0-102)</td>
<td>10 ± 4 (0-13)</td>
<td>Fast 16  Average 14  Slow 20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>49</td>
<td>44 ± 9 (26-65)</td>
<td>22/27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.1: Track-HD and Leiden cohorts for RNA-Seq analysis.

Manifest subjects demonstrated motor abnormalities that were unequivocal signs of HD. Premanifest gene carriers had a total motor score of 5 or lower and a diagnostic confidence score (DCS) less than 4 on the UHDRS, indicating no substantial motor signs. The HD group consists of the combined premanifest and manifest subjects. Controls were matched for age and gender. Age and clinical scores considered for the analysis were at time of blood collection. SD – standard deviation; TFC – Total Functional Capacity; TMS – Total Motor Score.
7.2.2 Sample collection

Whole blood was collected in two PAXGene Blood RNA tubes (PreAnalytix, Qiagen/BD Company) per subject, and immediately placed upright at room temperature. They were checked at 5 hours for incomplete mixing or separation, and any showing separation were remixed with a further 10 inversions. Tubes were stored overnight at -20°C and transferred to -80°C the following morning. TRACK-HD samples were sent on dry ice to Biorep within 30 days. LUMC samples were stored on site until transfer to deCODE, Iceland.

7.2.3 RNA preparation

RNA preparation for the TRACK-HD cohort was done by Biorep, Italy; and for the LUMC samples it was done by deCODE, Iceland, however both followed the same standard protocol. Total RNA extraction was performed using the PAXGene Blood RNA kit (catalog N. 762174; PreAnalytix, Qiagen/BD Company), following the supplier’s instructions. Each solution in the kit was divided into aliquots to process batches of 12 samples. Replicate tubes for each subject were processed on different days. RNA was stored at -80°C before proceeding with the quality measurements and further use. RNA was collected by centrifugation, washing with 70% ethanol, and resuspended in buffer. Quality measurements of total RNA were made using spectrophotometric analysis (Nanodrop), 260/280 ratio denaturing agarose gel, and the RNA 6000 Nano kit for the Agilent Bioanalyzer (catalog N. 5067-1511, Agilent Technologies).

Erythrocytes contain high levels of haemoglobin, and globin transcripts are highly expressed so can constitute up to 76% of total mRNA: Van Roon-Mom and colleagues established that depletion of globin transcripts from whole can enrich data obtained from next generation sequencing-based expression profiling (Mastrokolias et al., 2012). Samples for this project were therefore globin reduced using the GLOBINclear™ method (catalog N. AM1980, ThermoFisher Scientific). Quality control measures were made on globin-reduced samples on the Bioanalyzer RNA 6000 Nano kit (Catalog N. 5067-1511, Agilent Technologies).

7.2.4 RNA Sequencing

RNA sequencing for all samples was done by DeCODE, Iceland. Indexed cDNA sequencing libraries were prepared using the TruSeq™ Poly-A mRNA method (Illumina)(Illumina, 2014). Using this method the poly-A containing mRNA molecules are purified using oligo-dT attached magnetic beads: the adenines form complementary base pairs to thymines. The purified RNA is fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments are copied into cDNA using reverse transcriptase and random
hexamer primers (a mixture of oligonucleotides representing all possible sequence for that size). DNA Polymerase and RNAase H catalyze the synthesis of the second cDNA strand. The cDNA fragments then go through an end repair process to convert the overhangs into blunt ends. An ‘A’ base is then added to the 3’ end of the blunt phosphorylated DNA fragments which prepares the next step in which DNA fragments are ligated to the adapters, which have a single ‘T’ base overhang at their 3’ end. The cDNA templates are purified on a gel, then enriched with PCR to create the final cDNA library.

Paired-end sequencing of indexed cDNA libraries on a HiSeq 2500 generated at least 50 M reads per sample. Sequencing was performed using sequencing by synthesis (SBS) and cluster kits from Illumina. With SDS sequencing, a fluorescently labelled reversible terminator is imaged as each deoxy-NTP (nucleotide triphosphate) is added, and then cleaved to allow incorporation of the next base, enabling each base to be detected as they are incorporated into the DNA template strands. Since all 4 reversible terminator-bound dNTPs are present during each sequencing cycle, natural competition occurs minimizing incorporation bias (Illumina, 2015). Indexed samples were de-multiplexed and FASTQ files were generated.

**7.2.5 Quality control**

Sequencing failed for six TRACK-HD samples, including four premanifest, one manifest and one control subject. Quality control analysis was performed using the RNA-SeQC package (DeLuca et al., 2012), ensuring measures including rRNA rate, mapping rate, concordance mapping rate and uniqueness rate were within acceptable ranges. Globin depletion was checked by inspecting read counts mapped to HBB, HBA1 and HBA2, confirming they made up less than 2% of reads for all samples. Four TRACK-HD and six Leiden samples failed quality control for duplication rate over 75%, GC bias or 5’ bias, and were removed, leaving 48 premanifest, 61 manifest and 21 control subjects in the TRACK-HD cohort and 15 premanifest, 54 manifest and 26 control subjects in the Leiden cohort.

**7.2.6 Gene expression analysis**

After planning meetings and discussions which I helped organize and took part in, the gene expression analysis was performed by Dr Kitty Lo under the supervision of Dr Vincent Plagnol, UCL Genetics Institute. The result of sequencing is multiple fragments of DNA. A critical step in the RNA-seq data analysis is the alignment of partial transcript reads to a reference genome sequence to establish where the sequence comes from. Reference-based alignment methods use the sequence of each read to find a potential mapping location either by an exact match for a reference or by scoring sequence similarity. Our RNA-Seq data were aligned to the
human reference genome hg19 using TopHat2 (Kim et al., 2013, Trapnell et al., 2009, Trapnell et al., 2012), which maps reads to the reference with Bowtie. TopHat2 can align reads of various lengths, and allow for various length indels and fusion breaks which can occur after genetic translocation. Read counts were summarized using HTSeq, keeping any duplicates and using the Ensembl transcript/gene database (http://www.ensembl.org/info/data/ftp/index.html, obtained in gtf format, genome build GRCh38.3, gene build (updated in June 2015). To remove residual batch effects the R package svaseq was used (Leek, 2014). Using the cleaned count data, differential expression analysis was conducted using the R package DESeq2 (Love et al., 2014) which uses shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. Outlier counts were removed using a Cooks distance cutoff of 5 in DESeq2 (Cooks distance is a statistical technique of evaluating the effect of outliers which may distort analysis). After filtering by the mean of normalized counts, 18,257 transcripts were detected. For the analysis of HD stage: firstly disease status: Premanifest / manifest / control, then all HD / control were used as categorical variables in DEseq2 and age and gender were used as covariates in the analysis. For the progression score analysis the unified Huntington’s disease progression score (see Chapters 2.5 and 3) was used as a numeric variable in DESeq2, and disease stage (HD / preHD) was used as a covariate.

### 7.2.7 Pathway analysis

Enrichment of differential expression among gene sets corresponding to biological hypotheses (pathways) was tested using the Gene Set Enrichment Analysis (GSEA) method (Subramanian et al., 2005) by Professor Peter Holmans, Cardiff University as a part of our collaborative project. Pathway analysis is introduced in General Methods, Chapter 2.7.8. Rather than defining a list of significant genes, GSEA ranks all genes in order of their differential expression statistic, and tests whether the genes in a particular gene set have a higher rank overall than would be expected by chance. The analysis is weighted by the differential expression statistic, thus giving more weight to more significant genes. Significance of enrichment was obtained by randomly permuting gene-wide association statistics among genes. One-sided p-values were calculated separately for differential upregulation and downregulation of expression in HD, and these were then converted into the corresponding chi-square statistic for use in the GSEA analysis. To avoid making a priori assumptions, a large pathway set from publicly available pathway databases was collected, including Gene Ontology (GO) (Consortium, 2016), Kyoto Encyclopedia of Genes and Genomes (KEGG) (KEGG, 2016), Mouse Genome Informatics (MGI) (MGI, 2016), PANTHER (PANTHER, 2016), BioCarta (BioCarta, 2016), REACTOME (REACTOME, 2016) and NCI (Institute, last updated: 18 September 2012). This resulted in a
total of 14,706 functional gene sets, many with overlapping members, containing between 3 and 500 genes. To correct for multiple testing of pathways we converted the GSEA p-values into q-values (Storey and Tibshirani, 2003), which can be interpreted as the minimum false discovery rate at which that q-value would be counted as significant.

I also interrogated the gene lists for evidence of enrichment within particular pathways using the online software GOrilla (Eden et al., 2009). This uses the order within the list of genes, which in my analysis were ranked according to differential expression with a given phenotype (see also Chapter 2.7.6).

Enrichment \((N, B, n, b)\) is defined as follows:

\[
N - \text{is the total number of genes} \\
B - \text{is the total number of genes associated with a specific GO term} \\
n - \text{is the number of genes in the top of the user's input list or in the target set when appropriate} \\
b - \text{is the number of genes in the intersection} \\
\text{Enrichment} = \frac{(b/n)}{(B/N)}
\]

For both the TRACK-HD and LUMC progression analysis GOrilla recognized 19115 genes out of 19184 gene terms entered, 19115 genes were recognized by gene, 219 duplicate genes were removed (keeping the highest ranking instance of each gene) leaving a total of 18896 genes. Only 17097 of these genes are associated with a GO term, so \(N= 17097\) in these analyses.

### 7.2.8 Gene co-expression networks

The use of public databases to provide pathways is limited by the pathway curation: due to poor annotation of many genes and limitations in our biological knowledge. To overcome this annotation gap, we also tested the sets of gene co-expression modules for enrichment of dysregulation, this part of the project was done by Professor Peter Holmans in Cardiff. Gene co-expression modules are constructed on the basis of similar expression patterns across samples (Stuart et al., 2003). Their co-expression suggests that they are controlled by the same transcriptional regulatory program, are functionally related, or are members of the same pathway or protein complex (Langfelder and Horvath, 2008) (Stuart et al., 2003). The following four sets of data were used for this work:

regions; the caudate nucleus (CN), BA4 region of the frontal cortex, which has motor function (FC-BA4), BA9 region of the frontal cortex, involved in association and cognitive functions (FC-BA9), and cerebellum (CB).

2. The set of 117 co-expression modules derived from the Gibbs et al. (2010) dataset, comprising microarray expression data from 150 human control individuals measured in four brain regions: cerebellum (CB), frontal cortex (FC), caudal pons (Pons) and temporal cortex (TCTX). Modules were generated using WGCNA as described in (International Genomics of Alzheimer’s Disease, 2015).

3. We generated a set of 213 co-expression modules from Braineac (2016), which consists of microarray expression data for 12 brain regions from 134 control brains; occipital cortex, frontal cortex, temporal cortex, hippocampus, intralobular white matter, cerebellar cortex, thalamus, putamen, substantia nigra, and medulla (inferior olivary nucleus). For each brain region, the array data was normalised in the R statistical-programming environment using the RMA algorithm (Carvalho and Irizarry, 2010). Principal Component Analysis (PCA) and hierarchical clustering were used to identify single outlier arrays for removal. In addition, small outlier clusters (<6 arrays) that were distinct from most of the other arrays were removed (i.e. small clusters appearing at the top of the dendrogram). Once outlier arrays were removed, the arrays were re-normalized and inspected again and re-processed if necessary until a homogenous dataset was produced. WGCNA was performed using the R package to derive modules (Langfelder and Horvath, 2008). Multiple probesets of the same gene were collapsed to a single value using the collapseRows() function, using default settings and based on gene annotation provided by Affymetrix (Affymetrix, 2016). Scale independence and mean connectivity were plotted to derive a soft threshold power of 6. Networks were unsigned.

4. The set of 111 co-expression modules from Zhang et al. (2013), generated using microarray expression data on 1,647 postmortem samples from three brain regions of late-onset Alzheimer’s disease (LOAD) and control subjects; prefrontal cortex (BA9), primary visual cortex (BA17), and cerebellum.

7.2.9 Concordance of fold change in gene expression between HD blood and cortex
Labadorf et al. (2015a) analyzed the transcriptome of human postmortem prefrontal cortex Brodmann area 9 (BA9) from 20 HD subjects and 49 controls using next-generation sequencing, identifying dysregulation of immune and developmental genes. Of the 15,834 genes common to both the combined Track-HD and Leiden blood dataset and the Labadorf et al. (2015a) prefrontal cortex dataset, 8447 had a fold change >1 (i.e. upregulated) in blood
and 7860 had a fold change >1 in cortex. Thus, if fold changes in the two datasets were assumed to be unrelated, the expected probability that a gene would show concordant fold change is equal to

\[
\left( \frac{8447}{15834} \right) \times \left( \frac{7860}{15834} \right) + \left( \frac{7387}{15834} \right) \times \left( \frac{7974}{15834} \right) = 0.4997
\]

The number of genes with concordant fold change in the absence of a relationship between the datasets is thus distributed as a binomial (15834, 0.4997) distribution. In the actual data, 8425 genes were observed to have concordant direction of fold change, significantly higher than the number expected by chance (7912).

7.3 Results: Effect of HD gene status and stage of disease on the transcriptome

7.3.1 No differential expression of individual transcripts in HD whole blood between disease stages or states

Attempting to identify both HD specific and stage-specific changes in gene expression (mRNA) level we compared premanifest, manifest and control subjects, whilst controlling for age and gender. Premanifest gene carriers had a mean total motor score (TMS) of 2 and total functional capacity (TFC) of 13 (Table 7.1), indicating no substantial motor signs. Manifest subjects demonstrated motor abnormalities that were unequivocal signs of HD. No transcripts were significantly differentially expressed (FDR < 0.05) between premanifest and manifest HD in either the Track-HD (Tabrizi et al., 2009b) or the independently collected Leiden cohort, or when they these cohorts combined (data for Track-HD shown in Table 7.2).

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Basic p-value</th>
<th>Adjusted p-value</th>
<th>z-score</th>
<th>Condition</th>
<th>Average count: Premanifest</th>
<th>Average count: Manifest</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHGDH</td>
<td>5.79E-06</td>
<td>0.1044</td>
<td>4.53381</td>
<td>Premanifest</td>
<td>4.6888</td>
<td>237.78</td>
</tr>
<tr>
<td>SPAG1</td>
<td>4.66E-04</td>
<td>1</td>
<td>-3.49935</td>
<td>Manifest</td>
<td>422.76</td>
<td>320.67</td>
</tr>
<tr>
<td>GATSL1</td>
<td>6.10E-04</td>
<td>1</td>
<td>-3.42713</td>
<td>Premanifest</td>
<td>1.54</td>
<td>0.38</td>
</tr>
<tr>
<td>IGF1</td>
<td>1.19E-03</td>
<td>1</td>
<td>3.24199</td>
<td>Manifest</td>
<td>4.69</td>
<td>11.56</td>
</tr>
<tr>
<td>IGSF23</td>
<td>1.46E-03</td>
<td>1</td>
<td>-3.18292</td>
<td>Premanifest</td>
<td>20.53</td>
<td>13.20</td>
</tr>
<tr>
<td>OTOGL</td>
<td>1.53E-03</td>
<td>1</td>
<td>-3.16988</td>
<td>Manifest</td>
<td>4.8844</td>
<td>0.49</td>
</tr>
<tr>
<td>MIR29B2</td>
<td>1.57E-03</td>
<td>1</td>
<td>-3.16142</td>
<td>Premanifest</td>
<td>125.69</td>
<td>100.47</td>
</tr>
<tr>
<td>ANKUB1</td>
<td>1.69E-03</td>
<td>1</td>
<td>-3.13974</td>
<td>Manifest</td>
<td>37.86</td>
<td>25.07</td>
</tr>
<tr>
<td>HIST1H3J</td>
<td>2.02E-03</td>
<td>1</td>
<td>3.08728</td>
<td>Premanifest</td>
<td>1.0005</td>
<td>2.08</td>
</tr>
</tbody>
</table>
As expression changes did not differ significantly between disease stages, all mutant HTT gene carriers were combined to increase the analytical power in a comparison of HD and controls. Once again there were no individually significant transcripts in independent or combined cohorts; the differential expression analysis in the combined cohort is given in Table 7.3.

### Table 7.2: Differential expression of transcripts for the TRACK-HD manifest HD vs premanifest HD samples showing that there are no individually significant differentially expressed transcripts. Only transcripts with unadjusted p-values <0.005 are shown.

<table>
<thead>
<tr>
<th>Entrez gene ID</th>
<th>Gene Symbol</th>
<th>p (diffexp)</th>
<th>q (diffexp)</th>
<th>log2(FC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>722</td>
<td>C4BPA</td>
<td>7.81E-06</td>
<td>1.41E-01</td>
<td>1.371</td>
</tr>
<tr>
<td>2297</td>
<td>FOXD1</td>
<td>9.09E-05</td>
<td>7.02E-01</td>
<td>-0.785</td>
</tr>
<tr>
<td>3805</td>
<td>KIR2DL4</td>
<td>1.93E-04</td>
<td>7.02E-01</td>
<td>0.651</td>
</tr>
<tr>
<td>196394</td>
<td>AMN1</td>
<td>2.11E-04</td>
<td>7.02E-01</td>
<td>0.208</td>
</tr>
<tr>
<td>94137</td>
<td>RP1L1</td>
<td>2.47E-04</td>
<td>7.02E-01</td>
<td>-1.350</td>
</tr>
<tr>
<td>158248</td>
<td>TTC16</td>
<td>2.67E-04</td>
<td>7.02E-01</td>
<td>-0.347</td>
</tr>
<tr>
<td>100422824</td>
<td>MIR3128</td>
<td>2.86E-04</td>
<td>7.02E-01</td>
<td>0.930</td>
</tr>
<tr>
<td>5797</td>
<td>PTPRM</td>
<td>3.12E-04</td>
<td>7.02E-01</td>
<td>-0.359</td>
</tr>
<tr>
<td>84692</td>
<td>CCDC54</td>
<td>4.79E-04</td>
<td>9.58E-01</td>
<td>2.532</td>
</tr>
<tr>
<td>889</td>
<td>KRIT1</td>
<td>7.30E-04</td>
<td>9.58E-01</td>
<td>-0.081</td>
</tr>
<tr>
<td>54221</td>
<td>SNTG2</td>
<td>7.51E-04</td>
<td>9.58E-01</td>
<td>-0.689</td>
</tr>
<tr>
<td>22979</td>
<td>EFR3B</td>
<td>8.17E-04</td>
<td>9.58E-01</td>
<td>0.494</td>
</tr>
<tr>
<td>56934</td>
<td>CA10</td>
<td>8.42E-04</td>
<td>9.58E-01</td>
<td>2.036</td>
</tr>
<tr>
<td>8763</td>
<td>CD164</td>
<td>9.53E-04</td>
<td>9.58E-01</td>
<td>0.098</td>
</tr>
<tr>
<td>597</td>
<td>BCL2A1</td>
<td>1.06E-03</td>
<td>9.58E-01</td>
<td>0.423</td>
</tr>
<tr>
<td>4940</td>
<td>OAS3</td>
<td>1.12E-03</td>
<td>9.58E-01</td>
<td>0.688</td>
</tr>
<tr>
<td>49</td>
<td>ACR</td>
<td>1.13E-03</td>
<td>9.58E-01</td>
<td>1.237</td>
</tr>
<tr>
<td>9262</td>
<td>STK17B</td>
<td>1.19E-03</td>
<td>9.58E-01</td>
<td>0.132</td>
</tr>
</tbody>
</table>
Table 7.3: Differential expression analysis in HD (premanifest and manifest combined) versus controls for the combined Track-HD and Leiden cohorts.

\[ p \text{ (diffexp)} \] – \( p \) value for differential expression between HD and controls; \( q \text{ (diffexp)} \) – \( q \) value shows correction for multiple testing in the combined dataset; \( \log_{2}(FC) \) – \( \log_{2} \) of the ratio of the mean counts in HD and controls. Transcripts with \( P<0.002 \) are shown.

<table>
<thead>
<tr>
<th>ID</th>
<th>Gene</th>
<th>( \log_{2}(FC) )</th>
<th>( p \text{ (diffexp)} )</th>
<th>( q \text{ (diffexp)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>54407</td>
<td>SLC38A2</td>
<td>1.19E-03</td>
<td>9.58E-01</td>
<td>0.124</td>
</tr>
<tr>
<td>285590</td>
<td>SH3PD2B</td>
<td>1.26E-03</td>
<td>9.58E-01</td>
<td>0.414</td>
</tr>
<tr>
<td>60370</td>
<td>AVPI1</td>
<td>1.49E-03</td>
<td>9.58E-01</td>
<td>0.273</td>
</tr>
<tr>
<td>6425</td>
<td>SFRPS5</td>
<td>1.49E-03</td>
<td>9.58E-01</td>
<td>0.634</td>
</tr>
<tr>
<td>387849</td>
<td>REP15</td>
<td>1.49E-03</td>
<td>9.58E-01</td>
<td>1.526</td>
</tr>
<tr>
<td>283726</td>
<td>FAM154B</td>
<td>1.53E-03</td>
<td>9.58E-01</td>
<td>0.748</td>
</tr>
<tr>
<td>143502</td>
<td>OR52I2</td>
<td>1.54E-03</td>
<td>9.58E-01</td>
<td>2.469</td>
</tr>
<tr>
<td>1999</td>
<td>ELF3</td>
<td>1.58E-03</td>
<td>9.58E-01</td>
<td>-0.343</td>
</tr>
<tr>
<td>54957</td>
<td>TXN4B</td>
<td>1.65E-03</td>
<td>9.58E-01</td>
<td>0.088</td>
</tr>
<tr>
<td>23446</td>
<td>SLC44A1</td>
<td>1.65E-03</td>
<td>9.58E-01</td>
<td>0.116</td>
</tr>
<tr>
<td>693213</td>
<td>MIR628</td>
<td>1.65E-03</td>
<td>9.58E-01</td>
<td>-0.365</td>
</tr>
<tr>
<td>375757</td>
<td>SWI5</td>
<td>1.68E-03</td>
<td>9.58E-01</td>
<td>0.114</td>
</tr>
<tr>
<td>728340</td>
<td>GTF2H2C</td>
<td>1.69E-03</td>
<td>9.58E-01</td>
<td>0.206</td>
</tr>
<tr>
<td>146713</td>
<td>RBFOX3</td>
<td>1.86E-03</td>
<td>9.58E-01</td>
<td>-0.434</td>
</tr>
<tr>
<td>26834</td>
<td>RNU4-2</td>
<td>1.89E-03</td>
<td>9.58E-01</td>
<td>1.028</td>
</tr>
<tr>
<td>79725</td>
<td>THAP9</td>
<td>1.93E-03</td>
<td>9.58E-01</td>
<td>0.149</td>
</tr>
<tr>
<td>164668</td>
<td>APOBEC3H</td>
<td>1.98E-03</td>
<td>9.58E-01</td>
<td>-0.323</td>
</tr>
</tbody>
</table>

7.3.2 Pathways are dysregulated in HD blood compared with controls

We next asked whether networks of genes with similar functional annotation were dysregulated in HD relative to controls. Pathway annotations were collated from publicly available gene ontology databases to form a set of generic pathways using the same method as the recent HD genome-wide association study (GWAS) of modifiers of age at onset (Consortium, 2015a) (see General Methods, Chapter 2.7.8). The number of pathways significantly dysregulated in both Track-HD and Leiden blood datasets was significantly higher than would be expected by chance (Table 7.4). Our findings indicate shared biology between the two independent cohorts despite differences in demographic and disease stage; Leiden subjects were on average 7 years older and had correspondingly higher TMS (mean 32 versus 14 in Track-HD) and lower TFC (mean 8 versus 12 in Track-HD). The significance of the overlap was greatly increased in analyses specifying the direction of dysregulation (increased or
decreased expression) (Table 7.4). Therefore, directional analyses were used in the combined dataset as the primary analysis.

<table>
<thead>
<tr>
<th>Reference dataset</th>
<th>Comparison dataset</th>
<th>Direction of dysregulation in HD</th>
<th>Number of pathways significant in both datasets (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Generic pathways</td>
</tr>
<tr>
<td>LUMC</td>
<td>TRACK-HD</td>
<td>Nondirectional</td>
<td>69 (4.6E-02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Downregulated</td>
<td>139 (&lt;1.0E-03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upregulated</td>
<td>219 (&lt;1.0E-03)</td>
</tr>
<tr>
<td>LUMC</td>
<td>TRACK-HD</td>
<td>Nondirectional</td>
<td>69 (1.4E-01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Downregulated</td>
<td>130 (1.7E-02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upregulated</td>
<td>217 (&lt;1.0E-03)</td>
</tr>
</tbody>
</table>

Table 7.4: Overlap analysis of Track-HD and LUMC cohorts shows that a significant excess of pathways are associated with HD (p < 0.05) in both datasets.

Significance of overlap is greatest when directionality is taken into account. There is an excess of significantly enriched pathways and modules in the reference dataset conditional on the pathway being enriched (p < 0.05) in the comparison dataset. The generic pathways gene set is collated from publicly-available databases including GO and KEGG. HD brain modules are derived from Neuerd and Bates (2014). Control brain modules are from the Braineac (2016) and Gibbs et al. (2010) expression datasets.

Gene set enrichment analysis (GSEA), with a false discovery rate (q-value) threshold of q < 0.05 to correct for multiple testing, identified 53 upregulated (Figure 7.1 and Table 7.5) and 14 downregulated pathways (Figure 7.2 and Table 7.6) that are at least nominally significant in both cohorts. Multiple immune-related pathways were upregulated, and RNA processing, ATP metabolism and DNA repair were notably downregulated, and T cell related pathways approached significance for downregulation. The 10 most dysregulated genes (p < 0.01) from the significantly up or downregulated generic pathways (q < 0.05) are listed in Table 7.7.

Notably, the significantly upregulated pathways contain some of the most differentially
expressed transcripts (Table 7.3), with several more contained in pathways reaching nominal significance (p < 0.05) for dysregulation. Genes highlighted by MGI pathways appear distinct from other pathway databases, likely because they are based on knockout studies in mice.
<table>
<thead>
<tr>
<th>Pathway</th>
<th>Number of dysregulated genes</th>
<th>p-value (combined)</th>
<th>q-value (combined)</th>
<th>p-value (Track-HD)</th>
<th>p-value (Leiden)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGI: 2419</td>
<td>434</td>
<td>3.03E-10</td>
<td>4.32E-06</td>
<td>5.10E-05</td>
<td>3.01E-05</td>
<td>abnormal innate immunity</td>
</tr>
<tr>
<td>MGI: 3009</td>
<td>432</td>
<td>5.78E-09</td>
<td>4.13E-05</td>
<td>5.96E-06</td>
<td>1.65E-04</td>
<td>abnormal cytokine secretion</td>
</tr>
<tr>
<td>GO: 50792</td>
<td>117</td>
<td>2.59E-08</td>
<td>1.23E-04</td>
<td>1.12E-02</td>
<td>7.24E-05</td>
<td>regulation of viral process</td>
</tr>
<tr>
<td>GO: 9615</td>
<td>208</td>
<td>1.22E-07</td>
<td>4.36E-04</td>
<td>3.06E-02</td>
<td>5.34E-06</td>
<td>response to virus</td>
</tr>
<tr>
<td>MGI: 2451</td>
<td>278</td>
<td>1.68E-07</td>
<td>4.80E-04</td>
<td>1.26E-02</td>
<td>9.51E-06</td>
<td>abnormal macrophage physiology</td>
</tr>
<tr>
<td>GO: 19221</td>
<td>308</td>
<td>2.38E-07</td>
<td>5.45E-04</td>
<td>4.60E-05</td>
<td>1.71E-04</td>
<td>cytokine-mediated signalling pathway</td>
</tr>
<tr>
<td>GO: 2252</td>
<td>365</td>
<td>3.10E-07</td>
<td>5.45E-04</td>
<td>7.01E-03</td>
<td>1.14E-04</td>
<td>immune effector process</td>
</tr>
<tr>
<td>MGI: 5025</td>
<td>406</td>
<td>3.44E-07</td>
<td>5.45E-04</td>
<td>5.91E-05</td>
<td>2.02E-04</td>
<td>abnormal response to infection</td>
</tr>
<tr>
<td>MGI: 1793</td>
<td>372</td>
<td>4.33E-07</td>
<td>5.82E-04</td>
<td>5.93E-05</td>
<td>2.42E-04</td>
<td>altered susceptibility to infection</td>
</tr>
<tr>
<td>MGI: 8568</td>
<td>305</td>
<td>4.49E-07</td>
<td>5.82E-04</td>
<td>4.79E-05</td>
<td>6.25E-05</td>
<td>abnormal interleukin secretion</td>
</tr>
<tr>
<td>GO: 48525</td>
<td>48</td>
<td>6.05E-07</td>
<td>7.09E-04</td>
<td>1.83E-02</td>
<td>6.41E-05</td>
<td>negative regulation of viral process</td>
</tr>
<tr>
<td>MGI: 8250</td>
<td>462</td>
<td>6.46E-07</td>
<td>7.09E-04</td>
<td>4.99E-04</td>
<td>4.84E-03</td>
<td>abnormal myeloid leukocyte morphology</td>
</tr>
<tr>
<td>REACTOME 287</td>
<td>264</td>
<td>8.59E-07</td>
<td>8.76E-04</td>
<td>1.24E-03</td>
<td>3.86E-05</td>
<td>REACT:CYTOKINE SIGNALLING IN IMMUNE SYSTEM</td>
</tr>
<tr>
<td>GO: 71345</td>
<td>403</td>
<td>1.15E-06</td>
<td>1.08E-03</td>
<td>4.41E-05</td>
<td>4.82E-04</td>
<td>cellular response to cytokine stimulus</td>
</tr>
<tr>
<td>GO: 45069</td>
<td>53</td>
<td>1.21E-06</td>
<td>1.08E-03</td>
<td>1.37E-02</td>
<td>2.99E-05</td>
<td>regulation of viral genome replication</td>
</tr>
<tr>
<td>GO: 45071</td>
<td>37</td>
<td>2.30E-06</td>
<td>1.93E-03</td>
<td>1.19E-02</td>
<td>1.36E-04</td>
<td>negative regulation of viral genome replication</td>
</tr>
<tr>
<td>GO: 1817</td>
<td>409</td>
<td>2.76E-06</td>
<td>2.19E-03</td>
<td>2.64E-03</td>
<td>1.81E-03</td>
<td>regulation of cytokine production</td>
</tr>
<tr>
<td>MGI: 8251</td>
<td>387</td>
<td>3.12E-06</td>
<td>2.34E-03</td>
<td>6.32E-04</td>
<td>2.83E-03</td>
<td>abnormal phagocyte morphology</td>
</tr>
<tr>
<td>GO: 31347</td>
<td>430</td>
<td>4.73E-06</td>
<td>3.38E-03</td>
<td>7.31E-04</td>
<td>6.90E-04</td>
<td>regulation of defence response</td>
</tr>
<tr>
<td>MGI: 2406</td>
<td>317</td>
<td>6.87E-06</td>
<td>4.67E-03</td>
<td>1.11E-03</td>
<td>1.33E-03</td>
<td>increased susceptibility to infection</td>
</tr>
<tr>
<td>GO: 50778</td>
<td>403</td>
<td>7.29E-06</td>
<td>4.73E-03</td>
<td>1.50E-02</td>
<td>3.41E-02</td>
<td>positive regulation of immune response</td>
</tr>
<tr>
<td>MGI: 8835</td>
<td>258</td>
<td>8.31E-06</td>
<td>5.16E-03</td>
<td>1.25E-02</td>
<td>1.67E-04</td>
<td>abnormal intercellular signalling peptide or protein level</td>
</tr>
<tr>
<td>MGI: 2444</td>
<td>438</td>
<td>1.14E-05</td>
<td>6.69E-03</td>
<td>8.89E-03</td>
<td>6.28E-03</td>
<td>abnormal T cell physiology</td>
</tr>
<tr>
<td>MGI: 5005</td>
<td>243</td>
<td>1.21E-05</td>
<td>6.69E-03</td>
<td>1.05E-03</td>
<td>1.25E-04</td>
<td>abnormal self-tolerance</td>
</tr>
<tr>
<td>MGI: 1844</td>
<td>242</td>
<td>1.28E-05</td>
<td>6.69E-03</td>
<td>8.83E-04</td>
<td>1.07E-04</td>
<td>autoimmune response</td>
</tr>
<tr>
<td>MGI: 8713</td>
<td>253</td>
<td>1.35E-05</td>
<td>6.69E-03</td>
<td>1.24E-02</td>
<td>3.66E-04</td>
<td>abnormal cytokine level</td>
</tr>
<tr>
<td>GO: 5773</td>
<td>379</td>
<td>1.36E-05</td>
<td>6.69E-03</td>
<td>7.03E-03</td>
<td>9.19E-03</td>
<td>vacuole</td>
</tr>
<tr>
<td>GO: 323</td>
<td>318</td>
<td>1.41E-05</td>
<td>6.69E-03</td>
<td>5.94E-03</td>
<td>8.00E-04</td>
<td>lytic vacuole</td>
</tr>
<tr>
<td>GO: 5764</td>
<td>318</td>
<td>1.41E-05</td>
<td>6.69E-03</td>
<td>5.94E-03</td>
<td>8.00E-04</td>
<td>lysosome</td>
</tr>
<tr>
<td>MGI: 5000</td>
<td>246</td>
<td>1.63E-05</td>
<td>7.52E-03</td>
<td>1.07E-03</td>
<td>1.71E-04</td>
<td>abnormal immune tolerance</td>
</tr>
<tr>
<td>MGI: 8195</td>
<td>412</td>
<td>1.94E-05</td>
<td>8.41E-03</td>
<td>3.20E-03</td>
<td>1.90E-04</td>
<td>abnormal antigen presenting cell morphology</td>
</tr>
<tr>
<td>MGI: 8469</td>
<td>437</td>
<td>2.68E-05</td>
<td>1.08E-02</td>
<td>2.01E-02</td>
<td>2.07E-04</td>
<td>abnormal protein level</td>
</tr>
<tr>
<td>GO: 2253</td>
<td>325</td>
<td>2.92E-05</td>
<td>1.13E-02</td>
<td>3.85E-02</td>
<td>4.33E-02</td>
<td>activation of immune response</td>
</tr>
<tr>
<td>REACTOME 589</td>
<td>161</td>
<td>3.02E-05</td>
<td>1.14E-02</td>
<td>2.65E-03</td>
<td>5.75E-04</td>
<td>REACT:INTERFERON SIGNALLING</td>
</tr>
<tr>
<td>MGI: 2441</td>
<td>257</td>
<td>3.42E-05</td>
<td>1.25E-02</td>
<td>1.02E-02</td>
<td>9.51E-03</td>
<td>abnormal granulocyte morphology</td>
</tr>
<tr>
<td>MGI: 2462</td>
<td>180</td>
<td>4.09E-05</td>
<td>1.40E-02</td>
<td>6.48E-04</td>
<td>2.50E-02</td>
<td>abnormal granulocyte physiology</td>
</tr>
<tr>
<td>ID</td>
<td>Count</td>
<td>p-value</td>
<td>q-value</td>
<td>Pathway Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
<td>-------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO: 71357</td>
<td>61</td>
<td>4.20E-05</td>
<td>1.40E-02</td>
<td>1.17E-02 8.60E-05 cellular response to type I interferon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO: 60337</td>
<td>61</td>
<td>4.20E-05</td>
<td>1.40E-02</td>
<td>1.17E-02 8.60E-05 type I interferon-mediated signalling pathway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO: 43903</td>
<td>134</td>
<td>4.28E-05</td>
<td>1.40E-02</td>
<td>1.11E-02 5.23E-04 regulation of symbiosis, encompassing mutualism through parasitism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 8248</td>
<td>307</td>
<td>4.39E-05</td>
<td>1.40E-02</td>
<td>4.16E-03 6.89E-03 abnormal mononuclear phagocyte morphology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO: 44437</td>
<td>243</td>
<td>4.50E-05</td>
<td>1.40E-02</td>
<td>6.10E-04 1.03E-02 vacuolar part</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO: 5765</td>
<td>135</td>
<td>4.97E-05</td>
<td>1.51E-02</td>
<td>1.88E-04 7.32E-03 lysosomal membrane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 2446</td>
<td>240</td>
<td>5.88E-05</td>
<td>1.75E-02</td>
<td>1.30E-03 1.97E-02 abnormal macrophage morphology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REACTOME 587</td>
<td>61</td>
<td>6.00E-05</td>
<td>1.75E-02</td>
<td>6.87E-03 1.32E-04 REACT:INTERFERON ALPHA BETA SIGNALLING</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 2425</td>
<td>192</td>
<td>7.94E-05</td>
<td>2.27E-02</td>
<td>2.34E-03 3.21E-04 altered susceptibility to autoimmune disorder</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 10210</td>
<td>188</td>
<td>8.70E-05</td>
<td>2.44E-02</td>
<td>2.09E-02 7.63E-03 abnormal circulating cytokine level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO: 34340</td>
<td>62</td>
<td>8.95E-05</td>
<td>2.46E-02</td>
<td>1.33E-02 6.41E-05 response to type I interferon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 2463</td>
<td>126</td>
<td>9.20E-05</td>
<td>2.48E-02</td>
<td>2.79E-03 1.49E-02 abnormal neutrophil physiology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 2498</td>
<td>226</td>
<td>1.01E-04</td>
<td>2.66E-02</td>
<td>1.18E-02 1.69E-02 abnormal acute inflammation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 2459</td>
<td>402</td>
<td>1.11E-04</td>
<td>2.77E-02</td>
<td>2.28E-02 1.08E-03 abnormal B cell physiology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG 5164</td>
<td>158</td>
<td>1.14E-04</td>
<td>2.80E-02</td>
<td>5.13E-02 2.53E-03 KEGG INFLUENZA A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 8704</td>
<td>106</td>
<td>1.54E-04</td>
<td>3.66E-02</td>
<td>4.76E-03 9.41E-03 abnormal interleukin-6 secretion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MGI: 8705</td>
<td>48</td>
<td>2.00E-04</td>
<td>4.68E-02</td>
<td>4.43E-02 4.94E-02 increased interleukin-6 secretion</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.5:*** 53 ‘generic’ pathways which are significantly upregulated in HD versus control blood GSEA. A total of 14,706 Generic pathways, each containing between 3 and 500 genes, were collated from publicly-available databases including GO and KEGG. Pathways are significantly dysregulated after multiple testing correction (q < 0.05). Enrichment p values in the current study for the Track-HD, Leiden and combined datasets are shown.
Figure 7.1: Upregulated pathways in HD versus control blood. Schematic representation of pathways collated from publicly available databases that are significantly upregulated in HD versus controls after correction for multiple testing (q < 0.05). Modules with similar gene content and functional annotation have been consolidated. Nodal shading is inversely proportional to false discovery rate threshold (q value); deep shades have low q values and pale shading is close to the 5% threshold. The weight of connecting lines is proportional to the number of genes shared between pathways. Figure prepared by T. Stone, from (Hensman Moss et al., 2017a).
<table>
<thead>
<tr>
<th>Pathway</th>
<th>Number of dysregulated genes</th>
<th>p-value (combined)</th>
<th>q-value (combined)</th>
<th>p-value (Track-HD)</th>
<th>p-value (Leiden)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO: 8380</td>
<td>282</td>
<td>5.22E-08</td>
<td>7.45E-04</td>
<td>4.25E-05</td>
<td>7.24E-05</td>
<td>RNA splicing</td>
</tr>
<tr>
<td>GO: 6397</td>
<td>359</td>
<td>2.38E-07</td>
<td>1.70E-03</td>
<td>1.48E-04</td>
<td>4.14E-04</td>
<td>mRNA processing</td>
</tr>
<tr>
<td>GO: 16887</td>
<td>329</td>
<td>1.37E-06</td>
<td>5.48E-03</td>
<td>1.96E-04</td>
<td>3.34E-02</td>
<td>ATPase activity</td>
</tr>
<tr>
<td>GO: 6200</td>
<td>333</td>
<td>1.54E-06</td>
<td>5.48E-03</td>
<td>2.42E-04</td>
<td>3.36E-02</td>
<td>ATP catabolic process</td>
</tr>
<tr>
<td>GO: 46034</td>
<td>361</td>
<td>5.36E-06</td>
<td>1.53E-02</td>
<td>1.74E-04</td>
<td>4.45E-02</td>
<td>ATP metabolic process</td>
</tr>
<tr>
<td>GO: 16607</td>
<td>144</td>
<td>9.06E-06</td>
<td>2.15E-02</td>
<td>4.68E-04</td>
<td>4.61E-03</td>
<td>nuclear speck</td>
</tr>
<tr>
<td>GO: 6281</td>
<td>356</td>
<td>1.66E-05</td>
<td>2.75E-02</td>
<td>2.00E-03</td>
<td>1.18E-04</td>
<td>DNA repair</td>
</tr>
<tr>
<td>GO: 16604</td>
<td>271</td>
<td>2.08E-05</td>
<td>2.75E-02</td>
<td>5.59E-03</td>
<td>2.46E-03</td>
<td>nuclear body</td>
</tr>
<tr>
<td>GO: 4386</td>
<td>135</td>
<td>2.12E-05</td>
<td>2.75E-02</td>
<td>2.83E-02</td>
<td>4.81E-02</td>
<td>helicase activity</td>
</tr>
<tr>
<td>GO: 375</td>
<td>184</td>
<td>2.40E-05</td>
<td>2.86E-02</td>
<td>1.14E-03</td>
<td>2.05E-03</td>
<td>RNA splicing, via transesterification reactions</td>
</tr>
<tr>
<td>MGI: 5094</td>
<td>219</td>
<td>4.60E-05</td>
<td>3.88E-02</td>
<td>1.87E-02</td>
<td>2.34E-02</td>
<td>Abnormal T cell proliferation</td>
</tr>
<tr>
<td>GO: 398</td>
<td>180</td>
<td>6.25E-05</td>
<td>3.88E-02</td>
<td>2.12E-03</td>
<td>2.01E-03</td>
<td>mRNA splicing, via spliceosome</td>
</tr>
<tr>
<td>GO: 377</td>
<td>180</td>
<td>6.25E-05</td>
<td>3.88E-02</td>
<td>2.12E-03</td>
<td>2.01E-03</td>
<td>RNA splicing, via transesterification reactions with bulged adenosine as nucleophile</td>
</tr>
<tr>
<td>GO: 5681</td>
<td>143</td>
<td>7.29E-05</td>
<td>4.34E-02</td>
<td>4.66E-03</td>
<td>3.52E-03</td>
<td>spliceosomal complex</td>
</tr>
</tbody>
</table>

**Table 7.6:** 14 'generic' pathways which are significantly downregulated in HD versus control blood GSEA. A total of 14,706 Generic pathways, each containing between 3 and 500 genes, were collated from publicly-available databases including GO and KEGG. Pathways are significantly dysregulated after multiple testing correction (q < 0.05). Enrichment p values in the current study for the Track-HD, Leiden and combined datasets are shown.
**Figure 7.2:** Downregulated pathways in HD versus control blood. Schematic representation of pathways collated from publicly available databases that are significantly downregulated in HD versus controls after correction for multiple testing ($q < 0.05$). Modules with similar gene content and functional annotation have been consolidated. Nodal shading is inversely proportional to false discovery rate threshold ($q$ value); deep shades have low $q$ values and pale shading is close to the 5% threshold. The weight of connecting lines is proportional to the number of genes shared between pathways. Figure prepared by T. Stone, from (Hensman Moss et al., 2017a).
| Direction     | Entrez
gene ID | Gene Symbol | p (Comb)  | log2FC (Comb) | p (Track-HD) | log2FC (Track-HD) | p (Leiden) | log2FC (Leiden) | Pathway membership (q < 0.05) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes in upregulated pathways</td>
<td>722</td>
<td>C4BPA</td>
<td>7.81E-06</td>
<td>1.371</td>
<td>1.29E-01</td>
<td>0.437</td>
<td>7.36E-01</td>
<td>0.187</td>
<td>GO:2252, GO:2253, GO:5773, GO:31347, GO:44437, GO:50778</td>
</tr>
<tr>
<td></td>
<td>8763</td>
<td>CD164</td>
<td>9.53E-04</td>
<td>0.098</td>
<td>2.97E-01</td>
<td>0.083</td>
<td>5.57E-03</td>
<td>0.101</td>
<td>GO:323, GO:5764, GO:5765, GO:5773, GO:44437</td>
</tr>
<tr>
<td></td>
<td>597</td>
<td>BCL2A1</td>
<td>1.06E-03</td>
<td>0.423</td>
<td>8.85E-02</td>
<td>0.319</td>
<td>1.20E-02</td>
<td>0.393</td>
<td>MGI:1793, MGI:2419, MGI:2462, MGI:2463, MGI:5025</td>
</tr>
<tr>
<td></td>
<td>4940</td>
<td>OAS3</td>
<td>1.12E-03</td>
<td>0.688</td>
<td>5.14E-02</td>
<td>0.602</td>
<td>6.45E-02</td>
<td>0.455</td>
<td>GO:2252, GO:9615, GO:19221, GO:34340, GO:43903, GO:45069, GO:45071, GO:48525, GO:50792, GO:60337, GO:71345, GO:71357, KEGG:5164, REACTOME:287, REACTOME:587, REACTOME:589</td>
</tr>
<tr>
<td></td>
<td>49</td>
<td>ACR</td>
<td>1.13E-03</td>
<td>1.237</td>
<td>7.54E-03</td>
<td>1.417</td>
<td>1.79E-01</td>
<td>0.768</td>
<td>GO:5773, GO:44437</td>
</tr>
<tr>
<td></td>
<td>9262</td>
<td>STK17B</td>
<td>1.19E-03</td>
<td>0.132</td>
<td>4.42E-02</td>
<td>0.134</td>
<td>3.56E-02</td>
<td>0.136</td>
<td>MGI:1844, MGI:2425, MGI:2444, MGI:3009, MGI:5000, MGI:5005, MGI:8568</td>
</tr>
<tr>
<td></td>
<td>164668</td>
<td>APOBEC3H</td>
<td>1.98E-03</td>
<td>-0.323</td>
<td>1.41E-01</td>
<td>-0.208</td>
<td>4.33E-03</td>
<td>-0.476</td>
<td>GO:2252, GO:9615, GO:43903, GO:45069, GO:45071, GO:48525, GO:50792</td>
</tr>
<tr>
<td></td>
<td>79026</td>
<td>AHNAK</td>
<td>2.12E-03</td>
<td>-0.169</td>
<td>1.48E-02</td>
<td>-0.201</td>
<td>1.27E-01</td>
<td>-0.106</td>
<td>MGI:1793, MGI:2406, MGI:2444, MGI:3009, MGI:5025, MGI:8568</td>
</tr>
<tr>
<td></td>
<td>6614</td>
<td>SIGLEC1</td>
<td>4.39E-03</td>
<td>0.634</td>
<td>3.58E-01</td>
<td>0.291</td>
<td>9.79E-02</td>
<td>0.552</td>
<td>MGI:2459, MGI:8195</td>
</tr>
<tr>
<td></td>
<td>875</td>
<td>CBS</td>
<td>4.42E-03</td>
<td>0.592</td>
<td>1.15E-01</td>
<td>0.439</td>
<td>2.38E-02</td>
<td>0.681</td>
<td>MGI:8469, MGI:8713, MGI:8835</td>
</tr>
<tr>
<td>Genes in downregulated pathways</td>
<td>9262</td>
<td>STK17B</td>
<td>1.19E-03</td>
<td>0.132</td>
<td>4.42E-02</td>
<td>0.134</td>
<td>3.56E-02</td>
<td>0.136</td>
<td>MGI:5094</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------</td>
<td>---------</td>
<td>-----------</td>
<td>-------</td>
<td>-----------</td>
<td>-------</td>
<td>-----------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>54957</td>
<td>TXNL4B</td>
<td>1.65E-03</td>
<td>0.088</td>
<td>2.99E-02</td>
<td>0.088</td>
<td>2.67E-02</td>
<td>0.090</td>
<td>GO:5681, GO:6397, GO:8380</td>
<td></td>
</tr>
<tr>
<td>375757</td>
<td>SWI5</td>
<td>1.68E-03</td>
<td>0.114</td>
<td>3.22E-02</td>
<td>0.112</td>
<td>2.67E-02</td>
<td>0.130</td>
<td>GO:6281</td>
<td></td>
</tr>
<tr>
<td>146713</td>
<td>RBFOX3</td>
<td>1.86E-03</td>
<td>-0.434</td>
<td>3.81E-02</td>
<td>-0.396</td>
<td>7.65E-02</td>
<td>-0.357</td>
<td>GO:6397, GO:8380</td>
<td></td>
</tr>
<tr>
<td>79026</td>
<td>AHNAK</td>
<td>2.12E-03</td>
<td>-0.169</td>
<td>1.48E-02</td>
<td>-0.201</td>
<td>1.27E-01</td>
<td>-0.106</td>
<td>MGI:5094</td>
<td></td>
</tr>
<tr>
<td>29890</td>
<td>RBM15B</td>
<td>2.67E-03</td>
<td>-0.055</td>
<td>9.18E-02</td>
<td>-0.048</td>
<td>8.98E-02</td>
<td>-0.044</td>
<td>GO:6397, GO:8380</td>
<td></td>
</tr>
<tr>
<td>9987</td>
<td>HNRNPD1</td>
<td>3.38E-03</td>
<td>-0.078</td>
<td>2.98E-02</td>
<td>-0.088</td>
<td>9.41E-03</td>
<td>-0.098</td>
<td>GO:6381</td>
<td></td>
</tr>
<tr>
<td>23499</td>
<td>MACF1</td>
<td>3.72E-03</td>
<td>-0.120</td>
<td>4.52E-03</td>
<td>-0.172</td>
<td>2.15E-01</td>
<td>-0.068</td>
<td>GO:6200, GO:16887, GO:46034</td>
<td></td>
</tr>
<tr>
<td>146754</td>
<td>DNAH2</td>
<td>4.04E-03</td>
<td>-0.621</td>
<td>1.82E-01</td>
<td>-0.415</td>
<td>2.39E-02</td>
<td>-0.723</td>
<td>GO:6200, GO:16887, GO:46034</td>
<td></td>
</tr>
<tr>
<td>10236</td>
<td>HNRNPDR</td>
<td>5.92E-03</td>
<td>-0.069</td>
<td>1.15E-01</td>
<td>-0.053</td>
<td>5.62E-02</td>
<td>-0.074</td>
<td>GO:375, GO:377, GO:398, GO:5681, GO:6397, GO:8380</td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.7:** The 10 most significantly dysregulated genes (p<0.01) in up or downregulated generic pathways (q<0.05). p (Comb/Track-HD/Leiden) – p value for differential expression between HD and controls in the combined, Track-HD or Leiden datasets; Log2FC – log2 of the ratio of mean counts in HD and controls.
7.3.3 Pathway dysregulation in HD whole blood overlaps with HD myeloid cells

In a related RNA-Seq study led by James Miller, in which I assisted with design, sample collection and analysis, we investigated the effect of pro-inflammatory stimulation on HD and control monocytes. Primary monocytes of 30 manifest HD patients and 33 control subjects were cultured with and without pro-inflammatory stimulation, and then RNAseq was performed using the same sequencing technologies as described above. Transcriptional dysregulation was observed in unstimulated monocytes from HD cases relative to controls (Miller et al., 2016b): pathway analysis revealed widespread resting enrichment of proinflammatory functional gene sets in HD monocytes. The pathway enrichment analyses in the Miller et al study used the same set of pathways used in the Hensman Moss et al work (Hensman Moss et al., 2017a). We investigated whether the same pathways were dysregulated in the HD monocytes to the HD whole blood. We found a significant excess of pathways to be significantly (p<0.05) enriched for dysregulation in both the Miller et al. (2016a) data and the combined TRACK-HD and Leiden whole blood data (Table 7.8). This overlap was attributable to a significant excess of pathways enriched for upregulation in both datasets. Overlap in downregulated pathways was not significantly larger than expected by chance. The 15 pathways most significantly (p<0.05) enriched for up and downregulation in both myeloid and whole blood are listed in Table 7.9. Pathways that are significantly enriched for upregulation relate mainly to immunity.

<table>
<thead>
<tr>
<th>Direction of dysregulation in HD</th>
<th>Number of pathways significant in both datasets (p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nondirectional</td>
<td>132 (0.009)</td>
</tr>
<tr>
<td>Downregulated</td>
<td>36 (0.113)</td>
</tr>
<tr>
<td>Upregulated</td>
<td>339 (&lt;1.0E-03)</td>
</tr>
</tbody>
</table>

Table 7.8: Number of pathways nominally significantly enriched (uncorrected p<0.05) in both the combined Track-HD/Leiden blood dataset and the unstimulated myeloid data of Miller et al. (2016a). The p-value measures whether there is an excess of significantly enriched pathways in the blood dataset conditional on the pathway being enriched (p < 0.05) in the myeloid dataset. The set of pathways was collated from publicly-available databases including GO and KEGG.
<table>
<thead>
<tr>
<th>Direction of dysregulation in HD</th>
<th>Pathway</th>
<th>Number of genes</th>
<th>p (blood: London+ Leiden)</th>
<th>p (myeloid un-stimulated)</th>
<th>p (blood &amp; myeloid combined)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upregulated</td>
<td>MGI: 2419</td>
<td>434</td>
<td>3.03E-10</td>
<td>3.77E-08</td>
<td>4.44E-16</td>
<td>abnormal innate immunity</td>
</tr>
<tr>
<td></td>
<td>MGI: 3009</td>
<td>432</td>
<td>5.78E-09</td>
<td>4.26E-07</td>
<td>8.54E-14</td>
<td>abnormal cytokine secretion</td>
</tr>
<tr>
<td></td>
<td>GO: 31347</td>
<td>430</td>
<td>4.73E-06</td>
<td>8.96E-09</td>
<td>1.35E-12</td>
<td>regulation of defence response</td>
</tr>
<tr>
<td></td>
<td>GO: 9615</td>
<td>208</td>
<td>1.22E-07</td>
<td>9.68E-07</td>
<td>3.64E-12</td>
<td>response to virus</td>
</tr>
<tr>
<td></td>
<td>MGI: 2451</td>
<td>278</td>
<td>1.68E-07</td>
<td>1.83E-06</td>
<td>9.16E-12</td>
<td>abnormal macrophage physiology</td>
</tr>
<tr>
<td></td>
<td>GO: 2252</td>
<td>365</td>
<td>3.10E-07</td>
<td>1.95E-06</td>
<td>1.76E-11</td>
<td>immune effector process</td>
</tr>
<tr>
<td></td>
<td>MGI: 1793</td>
<td>372</td>
<td>4.33E-07</td>
<td>2.30E-06</td>
<td>2.85E-11</td>
<td>altered susceptibility to infection</td>
</tr>
<tr>
<td></td>
<td>MGI: 8568</td>
<td>305</td>
<td>4.49E-07</td>
<td>3.26E-06</td>
<td>4.13E-11</td>
<td>abnormal interleukin secretion</td>
</tr>
<tr>
<td></td>
<td>MGI: 5025</td>
<td>406</td>
<td>3.44E-07</td>
<td>4.42E-06</td>
<td>4.28E-11</td>
<td>abnormal response to infection</td>
</tr>
<tr>
<td></td>
<td>MGI: 8835</td>
<td>258</td>
<td>8.31E-06</td>
<td>1.92E-07</td>
<td>4.49E-11</td>
<td>abnormal intercellular signalling peptide or protein level</td>
</tr>
<tr>
<td></td>
<td>MGI: 8713</td>
<td>253</td>
<td>1.35E-05</td>
<td>2.72E-07</td>
<td>1.01E-10</td>
<td>abnormal cytokine level</td>
</tr>
<tr>
<td></td>
<td>GO: 51607</td>
<td>138</td>
<td>3.28E-07</td>
<td>1.34E-05</td>
<td>1.19E-10</td>
<td>defence response to virus</td>
</tr>
<tr>
<td></td>
<td>GO: 1817</td>
<td>409</td>
<td>2.76E-06</td>
<td>1.68E-06</td>
<td>1.26E-10</td>
<td>regulation of cytokine production</td>
</tr>
<tr>
<td></td>
<td>REACTOME 287</td>
<td>264</td>
<td>8.59E-07</td>
<td>6.56E-06</td>
<td>1.52E-10</td>
<td>REACT:CYTOKINE SIGNALING IN IMMUNE SYSTEM</td>
</tr>
<tr>
<td></td>
<td>GO: 6954</td>
<td>352</td>
<td>2.72E-05</td>
<td>2.38E-07</td>
<td>1.73E-10</td>
<td>inflammatory response</td>
</tr>
<tr>
<td>GO: 50792</td>
<td>117</td>
<td>2.59E-08</td>
<td>4.23E-04</td>
<td>2.88E-10</td>
<td>regulation of viral process</td>
<td></td>
</tr>
<tr>
<td>GO: 43202</td>
<td>66</td>
<td>4.25E-02</td>
<td>2.47E-04</td>
<td>1.31E-04</td>
<td>lysosomal lumen</td>
<td></td>
</tr>
<tr>
<td>GO: 10921</td>
<td>88</td>
<td>1.48E-03</td>
<td>2.41E-02</td>
<td>3.99E-04</td>
<td>regulation of phosphatase activity</td>
<td></td>
</tr>
<tr>
<td>GO: 38024</td>
<td>57</td>
<td>1.91E-03</td>
<td>3.51E-02</td>
<td>7.10E-04</td>
<td>cargo receptor activity</td>
<td></td>
</tr>
<tr>
<td>GO: 16874</td>
<td>450</td>
<td>1.70E-03</td>
<td>4.11E-02</td>
<td>7.38E-04</td>
<td>ligase activity</td>
<td></td>
</tr>
<tr>
<td>MGI: 358</td>
<td>276</td>
<td>2.71E-02</td>
<td>6.37E-03</td>
<td>1.67E-03</td>
<td>abnormal cell morphology</td>
<td></td>
</tr>
<tr>
<td>REACTOME 596</td>
<td>10</td>
<td>3.98E-03</td>
<td>4.49E-02</td>
<td>1.72E-03</td>
<td>REACT:INTERLEUKIN-7 SIGNALING</td>
<td></td>
</tr>
<tr>
<td>GO: 6399</td>
<td>126</td>
<td>4.03E-02</td>
<td>7.24E-03</td>
<td>2.67E-03</td>
<td>tRNA metabolic process</td>
<td></td>
</tr>
<tr>
<td>GO: 2285</td>
<td>55</td>
<td>1.71E-02</td>
<td>2.09E-02</td>
<td>3.20E-03</td>
<td>lymphocyte activation involved in immune response</td>
<td></td>
</tr>
<tr>
<td>GO: 6457</td>
<td>185</td>
<td>3.37E-02</td>
<td>1.10E-02</td>
<td>3.29E-03</td>
<td>protein folding</td>
<td></td>
</tr>
<tr>
<td>GO: 70286</td>
<td>8</td>
<td>2.03E-02</td>
<td>1.83E-02</td>
<td>3.30E-03</td>
<td>axonemal dynein complex assembly</td>
<td></td>
</tr>
<tr>
<td>GO: 6302</td>
<td>99</td>
<td>1.46E-02</td>
<td>3.04E-02</td>
<td>3.88E-03</td>
<td>double-strand break repair</td>
<td></td>
</tr>
<tr>
<td>GO: 30165</td>
<td>75</td>
<td>4.59E-02</td>
<td>1.06E-02</td>
<td>4.18E-03</td>
<td>PDZ domain binding</td>
<td></td>
</tr>
<tr>
<td>MGI: 2419</td>
<td>434</td>
<td>2.06E-02</td>
<td>2.47E-02</td>
<td>4.37E-03</td>
<td>abnormal innate immunity</td>
<td></td>
</tr>
<tr>
<td>MGI: 1701</td>
<td>56</td>
<td>1.05E-02</td>
<td>4.94E-02</td>
<td>4.45E-03</td>
<td>incomplete embryo turning</td>
<td></td>
</tr>
<tr>
<td>GO: 19320</td>
<td>66</td>
<td>2.34E-02</td>
<td>2.67E-02</td>
<td>5.22E-03</td>
<td>hexose catabolic process</td>
<td></td>
</tr>
<tr>
<td>KEGG 4146</td>
<td>77</td>
<td>1.92E-02</td>
<td>4.05E-02</td>
<td>6.35E-03</td>
<td>KEGG PEROXISOME</td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.9:** Pathways significantly (p<0.05) upregulated in both the combined Track-HD and Leiden whole blood data and the unstimulated myeloid cell dataset of Miller et al. (2016a). Pathways are ordered by their combined p-value, which was obtained by combining the blood and myeloid p-values by Fisher’s method.
7.3.4 Gene co-expression modules from HD striatum are significantly enriched for
dysregulation in HD blood

A limitation of using curated pathways from databases is the incomplete or incorrect
annotation. One way to overcome this is to use gene co-expression, because genes that are
co-expressed often have related functions. WGCNA identifies clusters (modules) of genes with
highly correlated expression, constructing original, unbiased gene co-expression networks
based on observed data (Gibbs et al., 2013). HD brain expression modules were generated by
Neueder and Bates (2014), who applied WGCNA to Hodges et al. (2006) data and annotated
each module that was associated with HD disease status. To further fill the annotation gap
and better define functional biological pathways, collaborators Timothy Stone and Amelia
Guinee generated co-expression modules for control brain from the Braineac (2016) and
Gibbs et al. (2010) datasets (Hensman Moss et al., 2017a).

The full list of up and downregulated modules reaching nominal significant in both datasets is
given in Table 7.10. A list of genes from the modules in Table 7.10 that are themselves
nominally significantly dysregulated (p < 0.05) in the combined dataset is given in Table 7.11.
<table>
<thead>
<tr>
<th>Direction</th>
<th>Brain expression gene set</th>
<th>Module</th>
<th>Brain region</th>
<th>Annotation</th>
<th>Number of dysregulated genes</th>
<th>( p ) (Combined)</th>
<th>( p ) (Track-HD)</th>
<th>( p ) (Leiden)</th>
<th>Cor (HD)</th>
<th>BH (HD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upregulat ed</td>
<td>HD</td>
<td>111</td>
<td>FC_BA9</td>
<td>Immune response</td>
<td>514</td>
<td>7.81E-12</td>
<td>1.27E-04</td>
<td>7.53E-05</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>69 (FC4pos1)</td>
<td>FC_BA4</td>
<td>Inflammatory response</td>
<td>712</td>
<td>3.77E-08</td>
<td>3.05E-05</td>
<td>1.32E-03</td>
<td>0.61</td>
<td>3.77E-03</td>
</tr>
<tr>
<td>Control (B)</td>
<td></td>
<td>712</td>
<td>TCTX</td>
<td>Inflammatory response</td>
<td>213</td>
<td>1.41E-07</td>
<td>3.40E-05</td>
<td>8.14E-04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HD*</td>
<td>48 (CNpos2) *</td>
<td>CN</td>
<td>Lipid metabolism/regulation of transcription</td>
<td>1785</td>
<td>2.03E-07</td>
<td>3.85E-03</td>
<td>6.33E-03</td>
<td>0.72</td>
<td>2.21E-11</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td></td>
<td>110</td>
<td>FCTX</td>
<td>Inflammatory response</td>
<td>173</td>
<td>8.94E-07</td>
<td>1.04E-03</td>
<td>2.50E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (B)</td>
<td></td>
<td>909</td>
<td>White Matter</td>
<td>Activation of immune response</td>
<td>265</td>
<td>2.12E-06</td>
<td>1.24E-03</td>
<td>2.48E-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (B)</td>
<td></td>
<td>610</td>
<td>Substantia Nigra</td>
<td>Inflammatory response</td>
<td>178</td>
<td>1.21E-05</td>
<td>8.56E-04</td>
<td>5.57E-04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (B)</td>
<td></td>
<td>811</td>
<td>Thalamus</td>
<td>Inflammatory response</td>
<td>142</td>
<td>1.61E-05</td>
<td>3.94E-03</td>
<td>2.89E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (G)</td>
<td></td>
<td>56</td>
<td>Pons</td>
<td>Lipoprotein/ immune response /GTPase regulator activity</td>
<td>207</td>
<td>1.97E-05</td>
<td>2.44E-04</td>
<td>4.19E-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (B)</td>
<td></td>
<td>911</td>
<td>White Matter</td>
<td>Inflammatory response</td>
<td>159</td>
<td>3.00E-05</td>
<td>8.42E-04</td>
<td>1.39E-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>HD</td>
<td>28</td>
<td>CB</td>
<td>Immune response</td>
<td>209</td>
<td>3.11E-05</td>
<td>1.07E-02</td>
<td>1.19E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td></td>
<td>713</td>
<td>TCTX</td>
<td>Activation of immune response</td>
<td>171</td>
<td>4.02E-05</td>
<td>2.39E-02</td>
<td>4.67E-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-----</td>
<td>--------</td>
<td>-----------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>response</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>33</td>
<td>CB</td>
<td>Immune response</td>
<td>255</td>
<td>4.34E-05</td>
<td>1.08E-02</td>
<td>1.37E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>505</td>
<td>Putamen</td>
<td>Ether lipid metabolism</td>
<td>500</td>
<td>6.28E-05</td>
<td>3.16E-03</td>
<td>2.06E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>68</td>
<td>CN</td>
<td>Cilium</td>
<td>1268</td>
<td>1.09E-04</td>
<td>3.05E-02</td>
<td>5.00E-02</td>
<td>0.54</td>
<td>7.74E-06</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>516</td>
<td>Putamen</td>
<td>Cellular response to</td>
<td>133</td>
<td>3.07E-04</td>
<td>1.44E-02</td>
<td>1.71E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>64</td>
<td>CN</td>
<td>Inflammatory response</td>
<td>114</td>
<td>3.13E-04</td>
<td>1.18E-02</td>
<td>3.80E-02</td>
<td>0.46</td>
<td>2.28E-04</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>124</td>
<td>FC_BA9</td>
<td>NA</td>
<td>1176</td>
<td>2.91E-03</td>
<td>1.19E-02</td>
<td>2.37E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Downregulated</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>22</td>
<td>CB</td>
<td>Pro-rich region</td>
<td>831</td>
<td>1.83E-08</td>
<td>2.49E-03</td>
<td>2.06E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>28</td>
<td>FC</td>
<td>Intra-cellular transport/mitochondrion</td>
<td>3178</td>
<td>2.10E-08</td>
<td>6.30E-04</td>
<td>7.66E-05</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>304</td>
<td>Medulla</td>
<td>mRNA metabolic process</td>
<td>1811</td>
<td>2.91E-08</td>
<td>5.00E-15</td>
<td>4.01E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD*</td>
<td>66</td>
<td>CN</td>
<td>Synapse/ion channels</td>
<td>2645</td>
<td>2.71E-07</td>
<td>1.51E-04</td>
<td>2.13E-02</td>
<td>-0.80</td>
<td>6.03E-15</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>804</td>
<td>Thalamus</td>
<td>Regulation of cell</td>
<td>857</td>
<td>1.31E-06</td>
<td>4.03E-02</td>
<td>4.13E-04</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>522</td>
<td>Putamen</td>
<td>Regulation of RNA splicing</td>
<td>64</td>
<td>4.44E-06</td>
<td>6.26E-03</td>
<td>2.66E-04</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>74</td>
<td>Pons</td>
<td>Transcription/acetylation/protein transport</td>
<td>1183</td>
<td>9.22E-06</td>
<td>3.85E-08</td>
<td>7.44E-04</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>702</td>
<td>TCTX</td>
<td>Antigen processing:</td>
<td>4602</td>
<td>3.87E-04</td>
<td>1.22E-03</td>
<td>2.47E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>48</td>
<td>FC</td>
<td>Transcription corepressor/cell morphogenesis</td>
<td>648</td>
<td>4.65E-04</td>
<td>7.83E-03</td>
<td>2.05E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>202</td>
<td>Hippocampus</td>
<td>Mitochondrial membrane</td>
<td>2737</td>
<td>4.75E-04</td>
<td>1.16E-07</td>
<td>1.54E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>19</td>
<td>CB</td>
<td>Protein binding</td>
<td>155</td>
<td>7.44E-04</td>
<td>2.66E-02</td>
<td>2.26E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>906</td>
<td>White Matter</td>
<td>Uridyltransferase activity</td>
<td>416</td>
<td>1.12E-03</td>
<td>2.53E-02</td>
<td>1.12E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>93</td>
<td>Pons</td>
<td>Mitochondrion/nuclear lumen</td>
<td>317</td>
<td>1.30E-03</td>
<td>9.85E-03</td>
<td>8.74E-04</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>812</td>
<td>Thalamus</td>
<td>Transport of mature transcript to cytoplasm</td>
<td>114</td>
<td>1.42E-03</td>
<td>1.99E-02</td>
<td>4.70E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>102</td>
<td>FC_BA9</td>
<td>Cytoplasm</td>
<td>1908</td>
<td>1.47E-03</td>
<td>7.57E-03</td>
<td>1.31E-04</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>706</td>
<td>TCTX</td>
<td>Microtubule organising centre</td>
<td>481</td>
<td>1.93E-03</td>
<td>3.70E-05</td>
<td>3.80E-03</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>52</td>
<td>Pons</td>
<td>Acetylation/fatty acid metabolism</td>
<td>1590</td>
<td>3.28E-03</td>
<td>2.23E-02</td>
<td>1.31E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>3</td>
<td>CB (CBneg2)</td>
<td>mitochondrion</td>
<td>1164</td>
<td>3.19E-02</td>
<td>2.56E-02</td>
<td>1.29E-05</td>
<td>-0.45</td>
<td>1.66E-03</td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>25</td>
<td>CB</td>
<td>RNA binding</td>
<td>648</td>
<td>8.02E-01</td>
<td>1.72E-04</td>
<td>3.62E-02</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Table 7.10: All WGCNA brain expression modules significantly dysregulated (p < 0.05) in both Track-HD and Leiden datasets in HD versus control blood.

HD brain modules were defined by Neueder and Bates (2014), and Control brain modules were derived from Braineac (2016) or Gibbs et al. (2010) expression data. Neueder and Bates (2014) module identifiers are given in brackets where available. * denotes the caudate modules that were highly positively and negatively correlated with HD in their study. HTT is part of modules 66 (CNneg1) and 3 (CBneg2). HD co-expression modules defined by Neueder and Bates (2014); CTRL (B) – control brain co-expression modules from Braineac (2016); CTRL (G) – control brain co-expression modules from Gibbs et al. (2010). p (Combined/Track-HD/Leiden) – p value for differential expression between HD and controls in the combined, Track-HD or Leiden datasets; BH (HD) the Benjamini Hochberg significance value\(^1\) of correlation with HD in Neueder and Bates (2014) brain expression analysis, corrected for multiple comparisons; Cor (HD) the direction and size of correlation of a module with HD in Neueder and Bates (2014); CN – caudate nucleus; FC – frontal cortex; FC_BA4 - BA4 region of the frontal cortex; FC_BA9 – BA9 region of the frontal cortex; CB – cerebellum; TCTX – temporal cortex.

\(^1\) The Benjamini and Hochberg correction is a powerful method for dealing with multiple comparisons by controlling the false discovery rate which was used in the Neueder and Bates, 2014 paper NEUEDER, A. & BATES, G. P. 2014. A common gene expression signature in Huntington’s disease patient brain regions. BMC Med Genomics, 7, 60.. (This is an alternative method to the more widely recognized Bonferroni correction which instead controls the familywise error rate)
Table 7.11: Ten most significantly dysregulated genes \((p<0.05)\) from the WGCNA brain expression modules that were dysregulated (up or down) in HD blood.

\(p\) (Comb/Track-HD/Leiden) – \(p\) value for differential expression between HD and controls in the combined, Track-HD or Leiden datasets; Log2FC – \(\log_2\) of the ratio of the mean counts in HD and controls; HD co-expression modules defined by Neuder and Bates (2014); CTRL (B) – control brain co-expression modules from Braineac (2016); CTRL (G) – control brain co-expression modules from Gibbs et al. (2010).
In addition to reinforcing the biological conclusions, the significantly dysregulated modules from Table 7.10 also share genes with the top pathways, as illustrated in Figures 7.3 and 7.4.

We then investigated whether gene sets that are dysregulated in HD brain (Neueder and Bates, 2014) are also disrupted in peripheral blood. Table 7.12 lists the modules that were significantly dysregulated (after correcting for multiple testing of modules) in both HD brain (Neueder and Bates, 2014) and in our combined TRACK-HD and Leiden blood expression dataset. The direction of dysregulation in brain is shown by the correlation between the module eigengene and HD status (with a positive correlation corresponding to upregulation in the HD brain). Notably, two of the most significantly dysregulated modules in HD caudate (Neueder and Bates, 2014) were also significantly dysregulated in the same direction in blood (Table 7.5), not only in the combined dataset, but in each of the Track-HD and Leiden datasets independently; these being module 48 (CNpos2), which is upregulated in HD, and module 66 (CNneg1), which is downregulated.

<table>
<thead>
<tr>
<th>Module</th>
<th>Brain Region</th>
<th>Module name</th>
<th>Number of genes</th>
<th>p (combined)</th>
<th>p (TRACK-HD)</th>
<th>p (LUMC)</th>
<th>cor (HD brain)</th>
<th>p (HD brain)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>FC_BA4</td>
<td>FC4pos1</td>
<td>712</td>
<td>3.77E-08</td>
<td>3.05E-05</td>
<td>1.32E-03</td>
<td>0.61</td>
<td>3.77E-03</td>
<td>Inflammatory response</td>
</tr>
<tr>
<td>48</td>
<td>CN</td>
<td>CNpos2</td>
<td>1785</td>
<td>2.03E-07</td>
<td>3.85E-03</td>
<td>6.33E-03</td>
<td>0.72</td>
<td>2.21E-11</td>
<td>Lipid metabolism/regulation of transcription</td>
</tr>
<tr>
<td>64</td>
<td>CN</td>
<td>CNpos6</td>
<td>114</td>
<td>3.13E-04</td>
<td>1.17E-02</td>
<td>3.80E-02</td>
<td>0.46</td>
<td>2.28E-04</td>
<td>Inflammatory response</td>
</tr>
<tr>
<td>66</td>
<td>CN</td>
<td>CNneg1</td>
<td>2644</td>
<td>2.71E-07</td>
<td>1.51E-04</td>
<td>2.13E-02</td>
<td>-0.80</td>
<td>6.03E-15</td>
<td>Synapse</td>
</tr>
</tbody>
</table>

Table 7.12: Brain expression modules significantly dysregulated both in HD brain and HD blood. All modules in this table are significantly dysregulated after correction for multiple testing (q < 0.05) in the combined blood sample, and are nominally significantly dysregulated (p<0.05) in both Track-HD and Leiden datasets separately. Cor(HD brain) – the correlation between module eigengene and HD status observed by Neueder and Bates (2014) in brain.
expression data, with a positive correlation corresponding to upregulation in HD. \( p(\text{HD brain}) \) is the p-value for that correlation (corrected for multiple testing of modules).
Figure 7.3: Network diagram of the relationship between significantly (q<0.05) upregulated gene modules (Table 7.10) and generic biological pathways (Table 7.5) based on shared gene membership. The thickness of the edges corresponds to the proportion of overlap from the smaller term to the larger (overlap coefficient). Intensity of shading indicates p-value (darker colours have lower p-values), node size indicates size of gene content, node shape indicates origin of data (modules or pathways). For clarity, biological pathways with similar gene content are grouped together, and the shading reflects the most significant pathway in the group. Nodes are arranged such that the distance between them reflects similarity in gene content. Diagram rendered in Cytoscape, from (Hensman Moss et al., 2017a), prepared by T. Stone.
Figure 7.4: Network diagram of the relationship between significantly (q<0.05) downregulated gene modules (Table 7.10) and generic biological pathways (Table 6) based on shared gene membership. The thickness of the edges corresponds to the proportion of overlap from the smaller term to the larger (overlap coefficient). Intensity of shading indicates p-value (darker colours have lower p-values), node size indicates size of gene content, node shape indicates origin of data (modules or pathways). For clarity, biological pathways with similar gene content are grouped together, and the shading reflects the most significant pathway in the group. Nodes are arranged such that the distance between them reflects similarity in gene content. Diagram rendered in Cytoscape, from (Hensman Moss et al., 2017a), prepared by T. Stone.

The module membership (kME) of a gene is measured by the correlation of its expression with the eigengene, which is representative of all gene expression profiles in the module (Langfelder and Horvath, 2008); highly connected ‘hub’ genes have high kME values. Interestingly, among genes in module 48 (CNpos2), the Neueder and Bates (2014) HD caudate module that was also significantly upregulated in blood, there was a significant (p = 7.6 x 10^{-4}) correlation between dysregulation p-value in the direction of interest (positive) in HD blood and degree of module membership (kME) (Neueder and Bates, 2014). This suggests that
highly connected “hub” genes in this module may play a role in transcriptional dysregulation in HD. Genes in module 48 (CNpos2) that are dysregulated (p < 0.05) in both blood and caudate are shown in https://www.nature.com/articles/srep44849#supplementary-information. A similar, although much stronger, effect was noted in caudate (Neueder and Bates, 2014). There was no significant correlation in module 66 (CNneg1).

7.3.5 Expression changes in HD blood replicate those in HD prefrontal cortex

Labadorf et al. (2015a) identified dysregulated expression of immune and developmental genes in human HD postmortem prefrontal cortex (BA9). Fold changes in expression of individual genes in the combined Track-HD and Leiden data were compared by Timothy Stone and Amelia Guinee to those observed in Labadorf et al. (2015a), and were found to be in the same direction for 8,425 out of the 15,834 genes present in both datasets. This is a highly significant (p < 2.2x10^{-16}) excess (see Materials and Methods, 7.2.9), suggesting some concordance in signal at the individual gene level. Furthermore, a significant excess of generic pathways was found to be significantly (p < 0.05) dysregulated in both datasets, most markedly in the positive (p < 0.001) direction, but also in the negative (p = 0.028), thus showing an overlap in biological signal. Pathways significantly upregulated in both datasets are mainly related to immune response (Table 7.13 and Hensman Moss et al Table S12 https://www.nature.com/articles/srep44849#supplementary-information (Hensman Moss et al., 2017a)), a pattern also observed in the upregulated brain co-expression modules (Hensman Moss et al Table S13 https://www.nature.com/articles/srep44849#supplementary-information (Hensman Moss et al., 2017a)). Pathways downregulated in both datasets are shown in Table 7.13 and Hensman Moss et al Table S14 https://www.nature.com/articles/srep44849#supplementary-information (Hensman Moss et al., 2017a), with downregulated modules in Hensman Moss et al Table S15 https://www.nature.com/articles/srep44849#supplementary-information (Hensman Moss et al., 2017a). Notably, several modules related to the synapse and neuron projection are downregulated in both datasets. The two HD-related caudate modules from Neueder and Bates (2014) that were significantly dysregulated in blood were also significantly dysregulated in the same direction in Labadorf et al. (2015a). Module 48 (CNpos2) was significantly upregulated (p < 1x10^{-16}, Table S13) and module 66 (CNneg1) significantly downregulated (p < 1x10^{-16}), as are several other significant modules from Neueder and Bates (2014).
<table>
<thead>
<tr>
<th>Direction of effect</th>
<th>Pathway</th>
<th>Number of dysregulated genes</th>
<th>Blood p (Combined)</th>
<th>Brain p (Labadorf)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Upregulated</strong></td>
<td>MGI: 2459</td>
<td>402</td>
<td>1.11E-04</td>
<td>1.39E-13</td>
<td>abnormal B cell physiology</td>
</tr>
<tr>
<td></td>
<td>MGI: 2419</td>
<td>434</td>
<td>3.03E-10</td>
<td>2.05E-12</td>
<td>abnormal innate immunity</td>
</tr>
<tr>
<td></td>
<td>MGI: 1800</td>
<td>361</td>
<td>5.45E-04</td>
<td>2.58E-12</td>
<td>abnormal humoral immune response</td>
</tr>
<tr>
<td></td>
<td>MGI: 8195</td>
<td>412</td>
<td>1.94E-05</td>
<td>8.78E-12</td>
<td>abnormal antigen presenting cell morphology</td>
</tr>
<tr>
<td></td>
<td>MGI: 2490</td>
<td>333</td>
<td>8.00E-04</td>
<td>3.52E-11</td>
<td>abnormal immunoglobulin level</td>
</tr>
<tr>
<td></td>
<td>MGI: 8250</td>
<td>462</td>
<td>6.46E-07</td>
<td>4.04E-11</td>
<td>abnormal myeloid leukocyte morphology</td>
</tr>
<tr>
<td></td>
<td>MGI: 4939</td>
<td>381</td>
<td>3.31E-03</td>
<td>1.68E-10</td>
<td>abnormal B cell morphology</td>
</tr>
<tr>
<td></td>
<td>GO: 50778</td>
<td>403</td>
<td>7.29E-06</td>
<td>2.11E-10</td>
<td>positive regulation of immune response</td>
</tr>
<tr>
<td></td>
<td>MGI: 8251</td>
<td>387</td>
<td>3.12E-06</td>
<td>3.29E-10</td>
<td>abnormal phagocyte morphology</td>
</tr>
<tr>
<td></td>
<td>MGI: 3009</td>
<td>432</td>
<td>5.78E-09</td>
<td>5.24E-10</td>
<td>abnormal cytokine secretion</td>
</tr>
<tr>
<td><strong>Downregulated</strong></td>
<td>GO: 5874</td>
<td>327</td>
<td>4.97E-05</td>
<td>8.10E-05</td>
<td>microtubule</td>
</tr>
<tr>
<td></td>
<td>GO: 86</td>
<td>120</td>
<td>8.11E-03</td>
<td>1.70E-04</td>
<td>G2/M transition of mitotic cell cycle</td>
</tr>
<tr>
<td></td>
<td>GO: 48812</td>
<td>455</td>
<td>3.45E-02</td>
<td>2.20E-04</td>
<td>neuron projection morphogenesis</td>
</tr>
<tr>
<td></td>
<td>PAN-PW 29</td>
<td>120</td>
<td>4.80E-02</td>
<td>2.67E-04</td>
<td>Huntington disease</td>
</tr>
<tr>
<td></td>
<td>GO: 15631</td>
<td>187</td>
<td>4.14E-04</td>
<td>2.84E-04</td>
<td>tubulin binding</td>
</tr>
<tr>
<td></td>
<td>GO: 7017</td>
<td>372</td>
<td>7.94E-04</td>
<td>4.13E-04</td>
<td>microtubule-based process</td>
</tr>
<tr>
<td></td>
<td>MGI: 1828</td>
<td>233</td>
<td>1.66E-04</td>
<td>7.14E-04</td>
<td>abnormal T cell activation</td>
</tr>
</tbody>
</table>
Table 7.13: Ten most significantly upregulated and downregulated generic pathways in both HD blood and prefrontal cortex. Comparing gene expression changes in the combined Track-HD and Leiden HD blood dataset with HD prefrontal cortex from Labadorf, et al. (Labadorf et al., 2015b), a significant ($p < 0.001$) excess of generic pathways are significantly upregulated ($p < 0.05$) in both datasets; there is also a significant ($p = 0.028$) excess of generic pathways are significantly downregulated ($p < 0.05$) in both datasets. Blood/brain $p$ the $p$ value for pathway enrichment in HD relative to controls in the combined Track-HD and Leiden blood dataset (Combined) or the prefrontal cortex dataset (Labadorf).
7.3.6 Pathways dysregulated in the blood of HD subjects are associated with motor score

We investigated the effect of disease severity by testing for correlation between gene expression and UHDRS total motor score (TMS) in the 112 gene positive Track-HD subjects (Table 7.14). After correcting for multiple testing, expression of phosphatidylincholine transfer protein (PCTP) was significantly positively correlated with TMS. However, this was not found to be significantly correlated with TMS by Mastrokolias et al (Mastrokolias et al., 2015).

<table>
<thead>
<tr>
<th>Entrez gene ID</th>
<th>Gene Symbol</th>
<th>p (corr-TMS)</th>
<th>q (corr-TMS)</th>
<th>log2(FC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>58488</td>
<td>PCTP</td>
<td>1.82E-06</td>
<td>3.25E-02</td>
<td>8.00E-03</td>
</tr>
<tr>
<td>51060</td>
<td>TXNDC12</td>
<td>4.42E-05</td>
<td>1.79E-01</td>
<td>5.30E-03</td>
</tr>
<tr>
<td>57096</td>
<td>RPGRI1</td>
<td>4.64E-05</td>
<td>1.79E-01</td>
<td>1.25E-02</td>
</tr>
<tr>
<td>9258</td>
<td>MFHAS1</td>
<td>5.12E-05</td>
<td>1.79E-01</td>
<td>-8.40E-03</td>
</tr>
<tr>
<td>3667</td>
<td>IRS1</td>
<td>6.73E-05</td>
<td>1.79E-01</td>
<td>-1.37E-02</td>
</tr>
<tr>
<td>158293</td>
<td>FAM120AOS</td>
<td>6.88E-05</td>
<td>1.79E-01</td>
<td>3.80E-03</td>
</tr>
<tr>
<td>84263</td>
<td>HSDL2</td>
<td>7.01E-05</td>
<td>1.79E-01</td>
<td>6.30E-03</td>
</tr>
<tr>
<td>56925</td>
<td>LIXN</td>
<td>1.01E-04</td>
<td>2.22E-01</td>
<td>1.05E-02</td>
</tr>
<tr>
<td>118881</td>
<td>COMTD1</td>
<td>1.12E-04</td>
<td>2.22E-01</td>
<td>-8.10E-03</td>
</tr>
<tr>
<td>597</td>
<td>BCL2A1</td>
<td>1.44E-04</td>
<td>2.35E-01</td>
<td>1.44E-02</td>
</tr>
<tr>
<td>23002</td>
<td>DAAM1</td>
<td>1.58E-04</td>
<td>2.35E-01</td>
<td>-7.80E-03</td>
</tr>
<tr>
<td>3655</td>
<td>ITGA6</td>
<td>1.58E-04</td>
<td>2.35E-01</td>
<td>-8.00E-03</td>
</tr>
<tr>
<td>137835</td>
<td>TMEM71</td>
<td>1.96E-04</td>
<td>2.53E-01</td>
<td>7.00E-03</td>
</tr>
</tbody>
</table>

Table 7.14: Correlation between gene expression and TMS in gene positive Track-HD subjects. Genes with p<0.0002 are shown (full table: S16 in Hensman Moss et al (Hensman Moss et al., 2017a)). p (corr-TMS) – p value for correlation between expression and TMS; q (corr-TMS) – q value shows correction for multiple testing of genes; Log2(FC) – the change in log2 (expression) per unit increase of TMS.

We then tested whether generic pathways, that were significantly enriched for upregulated (Table 7.5) or downregulated (Table 7.6) genes, for enrichment of genes correlated with TMS in the expected direction using a similar method to that previously used to test for enrichment of differentially expressed genes (Table 7.15). Several immune related pathways were positively correlated with TMS, including MGI:2419, the most significantly dysregulated
pathway in HD blood. Downregulated pathways that correlated with TMS were related to T-cells, ATP metabolism and DNA repair.

Similarly, we tested whether modules dysregulated in HD blood relative to controls (Table 7.10) also correlated with TMS in the expected direction (Table 7.16). Many modules significantly correlated with TMS, including 68 (CNpos5) and 66 (CNneg1), which were also dysregulated in the HD caudate (Neueder and Bates, 2014).
<table>
<thead>
<tr>
<th>Direction of effect</th>
<th>Pathway</th>
<th>p (combined-diffexp)</th>
<th>p (TRACK-diffexp)</th>
<th>p (TRACK-TMS)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positively correlated with TMS</td>
<td>MGI: 2419</td>
<td>3.03E-10</td>
<td>5.10E-05</td>
<td>2.18E-03</td>
<td>Abnormal innate immunity</td>
</tr>
<tr>
<td></td>
<td>GO: 10942</td>
<td>8.79E-02</td>
<td>4.70E-02</td>
<td>3.21E-03</td>
<td>positive regulation of cell death</td>
</tr>
<tr>
<td></td>
<td>MGI: 2462</td>
<td>4.09E-05</td>
<td>6.48E-04</td>
<td>6.39E-03</td>
<td>Abnormal granulocyte physiology</td>
</tr>
<tr>
<td></td>
<td>MGI: 8556</td>
<td>6.85E-04</td>
<td>8.68E-03</td>
<td>7.91E-03</td>
<td>Abnormal tumour necrosis factor secretion</td>
</tr>
<tr>
<td></td>
<td>MGI: 2463</td>
<td>9.20E-05</td>
<td>2.79E-03</td>
<td>8.99E-03</td>
<td>Abnormal neutrophil physiology</td>
</tr>
<tr>
<td></td>
<td>MGI: 8704</td>
<td>1.54E-04</td>
<td>4.76E-03</td>
<td>9.56E-03</td>
<td>abnormal_interleukin-6_secretion</td>
</tr>
<tr>
<td></td>
<td>GO: 5773</td>
<td>1.36E-05</td>
<td>7.03E-03</td>
<td>1.62E-02</td>
<td>vacuole</td>
</tr>
<tr>
<td></td>
<td>GO: 50792</td>
<td>2.59E-08</td>
<td>1.12E-02</td>
<td>1.64E-02</td>
<td>regulation of viral process</td>
</tr>
<tr>
<td></td>
<td>MGI: 5351</td>
<td>6.95E-03</td>
<td>2.20E-02</td>
<td>2.76E-02</td>
<td>Decreased susceptibility to autoimmune disorder</td>
</tr>
<tr>
<td></td>
<td>GO: 44437</td>
<td>4.50E-05</td>
<td>6.10E-04</td>
<td>3.48E-02</td>
<td>vacuolar part</td>
</tr>
<tr>
<td></td>
<td>MGI: 3627</td>
<td>5.45E-02</td>
<td>3.61E-02</td>
<td>3.62E-02</td>
<td>Abnormal leukocyte tethering or rolling</td>
</tr>
<tr>
<td></td>
<td>GO: 50427</td>
<td>1.64E-01</td>
<td>4.26E-02</td>
<td>4.30E-02</td>
<td>3'-phosphoadenosine 5'-phosphosulfate metabolic process</td>
</tr>
<tr>
<td></td>
<td>GO: 34035</td>
<td>1.64E-01</td>
<td>4.26E-02</td>
<td>4.30E-02</td>
<td>purine ribonucleoside bisphosphate metabolic process</td>
</tr>
<tr>
<td></td>
<td>MGI: 2451</td>
<td>1.68E-07</td>
<td>1.26E-02</td>
<td>4.33E-02</td>
<td>Abnormal macrophage physiology</td>
</tr>
<tr>
<td></td>
<td>GO: 6024</td>
<td>1.39E-02</td>
<td>2.77E-02</td>
<td>4.56E-02</td>
<td>glycosaminoglycan biosynthetic process</td>
</tr>
<tr>
<td>Negatively correlated with TMS</td>
<td>GO: 45786</td>
<td>8.92E-04</td>
<td>1.88E-02</td>
<td>3.23E-05</td>
<td>negative regulation of cell cycle</td>
</tr>
<tr>
<td></td>
<td>MGI: 706</td>
<td>9.70E-02</td>
<td>1.11E-02</td>
<td>9.09E-05</td>
<td>Small thymus</td>
</tr>
<tr>
<td>MGI</td>
<td>p(combined-diffexp)</td>
<td>p(TRACK-diffexp)</td>
<td>p(TRACK-TMS)</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---------------------</td>
<td>------------------</td>
<td>-----------------------</td>
<td>-------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>2364</td>
<td>6.81E-02</td>
<td>1.14E-02</td>
<td>2.57E-04</td>
<td>Abnormal thymus size</td>
<td></td>
</tr>
<tr>
<td>5018</td>
<td>6.50E-04</td>
<td>5.94E-03</td>
<td>2.70E-04</td>
<td>Decreased T cell number</td>
<td></td>
</tr>
<tr>
<td>2435</td>
<td>1.95E-04</td>
<td>6.87E-03</td>
<td>2.79E-04</td>
<td>Abnormal effector T cell morphology</td>
<td></td>
</tr>
<tr>
<td>8081</td>
<td>1.48E-03</td>
<td>5.61E-03</td>
<td>3.83E-04</td>
<td>Abnormal single-positive T cell number</td>
<td></td>
</tr>
<tr>
<td>2145</td>
<td>1.19E-03</td>
<td>5.67E-03</td>
<td>8.68E-04</td>
<td>Abnormal T cell differentiation</td>
<td></td>
</tr>
<tr>
<td>2444</td>
<td>3.61E-04</td>
<td>7.95E-03</td>
<td>8.74E-04</td>
<td>Abnormal T cell physiology</td>
<td></td>
</tr>
<tr>
<td>2432</td>
<td>6.45E-04</td>
<td>3.48E-02</td>
<td>1.01E-03</td>
<td>Abnormal CD4-positive T cell morphology</td>
<td></td>
</tr>
<tr>
<td>6387</td>
<td>8.81E-05</td>
<td>4.95E-03</td>
<td>1.31E-03</td>
<td>Abnormal T cell number</td>
<td></td>
</tr>
<tr>
<td>8083</td>
<td>7.34E-03</td>
<td>3.36E-02</td>
<td>1.77E-03</td>
<td>Decreased single-positive T cell number</td>
<td></td>
</tr>
<tr>
<td>8077</td>
<td>7.89E-03</td>
<td>3.31E-02</td>
<td>2.14E-03</td>
<td>abnormal_CD8-positive_T_cell_number</td>
<td></td>
</tr>
<tr>
<td>1823</td>
<td>5.18E-02</td>
<td>4.61E-02</td>
<td>2.15E-03</td>
<td>Thymus hypoplasia</td>
<td></td>
</tr>
<tr>
<td>6200</td>
<td>1.54E-06</td>
<td>2.42E-04</td>
<td>2.34E-03</td>
<td>ATP catabolic process</td>
<td></td>
</tr>
<tr>
<td>46034</td>
<td>5.36E-06</td>
<td>1.74E-04</td>
<td>2.56E-03</td>
<td>ATP metabolic process</td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.15**: Enrichment of up or downregulated pathways from HD vs. control blood with TMS in the combined Track-HD and Leiden cohort. p(combined-diffexp) – enrichment p-value for upregulated genes in the combined Track-HD and Leiden sample. p/TRACK-diffexp) - enrichment p-value for upregulated genes in the Track-HD sample alone. p/TRACK-TMS - enrichment p-value for genes positively correlated with TMS in the TRACK-HD sample.
<table>
<thead>
<tr>
<th>Direction</th>
<th>Brain expression gene set</th>
<th>Module</th>
<th>Brain region</th>
<th>Annotation</th>
<th>Number of dysregulated genes</th>
<th>p (Combined diffexp)</th>
<th>p (TRACK-diffexp)</th>
<th>p (TRACK-TMS)</th>
<th>Cor (HD)</th>
<th>BH (HD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upregulated</td>
<td>HD</td>
<td>68</td>
<td>CN</td>
<td>Cilium</td>
<td>1268</td>
<td>1.09E-04</td>
<td>3.05E-02</td>
<td>5.52E-07</td>
<td>0.54</td>
<td>7.74E-06</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>909</td>
<td>White Matter</td>
<td>Activation of immune response</td>
<td>265</td>
<td>2.12E-06</td>
<td>1.24E-03</td>
<td>8.22E-04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>713</td>
<td>TCTX</td>
<td>Activation of immune response</td>
<td>171</td>
<td>4.02E-05</td>
<td>2.39E-02</td>
<td>1.69E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>111</td>
<td>FC_BA9</td>
<td>Immune response</td>
<td>514</td>
<td>7.81E-12</td>
<td>1.27E-04</td>
<td>3.75E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (G)</td>
<td>56</td>
<td>Pons</td>
<td>Lipoprotein/ immune response</td>
<td>207</td>
<td>1.97E-05</td>
<td>2.44E-04</td>
<td>7.72E-03</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>28</td>
<td>CB</td>
<td>Immune response</td>
<td>209</td>
<td>3.11E-05</td>
<td>1.07E-02</td>
<td>8.70E-03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>505</td>
<td>Putamen</td>
<td>Ether lipid metabolism</td>
<td>500</td>
<td>6.28E-05</td>
<td>3.16E-03</td>
<td>6.43E-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>911</td>
<td>White Matter</td>
<td>Inflammatory response</td>
<td>159</td>
<td>3.00E-05</td>
<td>8.42E-04</td>
<td>7.75E-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>124</td>
<td>FC_BA9</td>
<td>NA</td>
<td>1176</td>
<td>2.91E-03</td>
<td>1.19E-02</td>
<td>9.14E-02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (B)</td>
<td>110</td>
<td>FCTX</td>
<td>Inflammatory response</td>
<td>173</td>
<td>8.94E-07</td>
<td>1.04E-03</td>
<td>1.34E-01</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>33</td>
<td>CB</td>
<td>Immune response</td>
<td>255</td>
<td>4.34E-05</td>
<td>1.08E-02</td>
<td>1.52E-01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (B)</td>
<td>610</td>
<td>Substantia Nigra</td>
<td>Inflammatory response</td>
<td>178</td>
<td>1.21E-05</td>
<td>8.56E-04</td>
<td>2.00E-01</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>64 (CNpos6)</td>
<td>CN</td>
<td>Inflammatory response</td>
<td>114</td>
<td>3.13E-04</td>
<td>1.18E-02</td>
<td>2.22E-01</td>
<td>0.46</td>
<td>2.28E-04</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
<td>-------</td>
<td>-----------------------</td>
<td>-----</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>811</td>
<td>Thalamus</td>
<td>Inflammatory response</td>
<td>142</td>
<td>1.61E-05</td>
<td>3.94E-03</td>
<td>2.28E-01</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>712</td>
<td>TCTX</td>
<td>Inflammatory response</td>
<td>213</td>
<td>1.41E-07</td>
<td>3.40E-05</td>
<td>2.35E-01</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>516</td>
<td>Putamen</td>
<td>Cellular response to cytokine stimulus</td>
<td>133</td>
<td>3.07E-04</td>
<td>1.44E-02</td>
<td>4.16E-01</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD</td>
<td>69 (FC4pos1)</td>
<td>FC_BA4</td>
<td>Inflammatory response</td>
<td>712</td>
<td>3.77E-08</td>
<td>3.05E-05</td>
<td>5.22E-01</td>
<td>0.61</td>
<td>3.77E-03</td>
<td></td>
</tr>
<tr>
<td>HD*</td>
<td>48 (CNpos2)</td>
<td>CN</td>
<td>Lipid metabolism/regulation of transcription</td>
<td>1785</td>
<td>2.03E-07</td>
<td>3.85E-03</td>
<td>6.14E-01</td>
<td>0.72</td>
<td>2.21E-11</td>
<td></td>
</tr>
<tr>
<td>Downregulated</td>
<td>Control (B)</td>
<td>304</td>
<td>Medulla</td>
<td>mRNA metabolic process</td>
<td>1811</td>
<td>2.91E-08</td>
<td>5.00E-15</td>
<td>6.11E-16</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Control (B)</td>
<td>702</td>
<td>TCTX</td>
<td>Antigen processing: ubiquitination and proteasome degradation</td>
<td>4602</td>
<td>3.87E-04</td>
<td>1.22E-03</td>
<td>2.04E-13</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (B)</td>
<td>202</td>
<td>Hippocampus</td>
<td>Mitochondrial membrane</td>
<td>2737</td>
<td>4.75E-04</td>
<td>1.16E-07</td>
<td>1.44E-09</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Control (G)</td>
<td>28</td>
<td>FC</td>
<td>Intra-cellular transport/mitochondrion</td>
<td>3178</td>
<td>2.10E-08</td>
<td>6.30E-04</td>
<td>4.16E-09</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>HD*</td>
<td>66 (CNneg1)</td>
<td>CN</td>
<td>Synapse/ion channels</td>
<td>2645</td>
<td>2.71E-07</td>
<td>1.51E-04</td>
<td>1.05E-07</td>
<td>-0.8</td>
<td>6.03E-15</td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Control (G)</td>
<td>Control (B)</td>
<td>HD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pons</td>
<td>74</td>
<td>522</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control (G)</td>
<td>Control (B)</td>
<td>Control (G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>906</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>804</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>22</td>
<td>3 (CBneg2)</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>522</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>102</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control (G)</td>
<td>Control (G)</td>
<td>HD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC</td>
<td>Control (G)</td>
<td>Control (B)</td>
<td>HD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1183</td>
<td>706</td>
<td>448</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1590</td>
<td>522</td>
<td>648</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pons</td>
<td>1183</td>
<td>804</td>
<td>648</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1590</td>
<td>804</td>
<td>648</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>906</td>
<td>448</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>22</td>
<td>3 (CBneg2)</td>
<td>448</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>906</td>
<td>448</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB</td>
<td>22</td>
<td>3 (CBneg2)</td>
<td>448</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>522</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC</td>
<td>48</td>
<td>48</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td>48</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table:**

<table>
<thead>
<tr>
<th>Region</th>
<th>Condition</th>
<th>Area</th>
<th>GO Term</th>
<th>Adj. FDR</th>
<th>Benjamini FDR</th>
<th>Log2 Fold Change</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pons</td>
<td>Control (G)</td>
<td>52</td>
<td>Acetylation/fatty acid metabolism</td>
<td>1590</td>
<td>3.28E-03</td>
<td>2.23E-02</td>
<td>1.30E-07</td>
</tr>
<tr>
<td></td>
<td>Control (G)</td>
<td>74</td>
<td>Transcription/acetylation/protein transport</td>
<td>1183</td>
<td>9.22E-06</td>
<td>3.85E-08</td>
<td>1.19E-05</td>
</tr>
<tr>
<td></td>
<td>Control (G)</td>
<td>22</td>
<td>Pro-rich region</td>
<td>831</td>
<td>1.83E-08</td>
<td>2.49E-03</td>
<td>7.72E-05</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>804</td>
<td>Regulation of cell morphogenesis</td>
<td>857</td>
<td>1.31E-06</td>
<td>4.03E-02</td>
<td>8.29E-05</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>706</td>
<td>Microtubule organising center</td>
<td>481</td>
<td>1.93E-03</td>
<td>3.70E-05</td>
<td>3.00E-04</td>
</tr>
<tr>
<td></td>
<td>Control (G)</td>
<td>48</td>
<td>Transcription corepressor/cell morphogenesis</td>
<td>648</td>
<td>4.65E-04</td>
<td>7.83E-03</td>
<td>7.14E-04</td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>102</td>
<td>Cytoplasm</td>
<td>1908</td>
<td>1.47E-03</td>
<td>7.57E-03</td>
<td>9.26E-03</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>906</td>
<td>Uridyltransferase activity</td>
<td>416</td>
<td>1.12E-03</td>
<td>2.53E-02</td>
<td>1.34E-02</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>812</td>
<td>Transport of mature transcript to cytoplasm</td>
<td>114</td>
<td>1.42E-03</td>
<td>1.99E-02</td>
<td>1.36E-02</td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>19</td>
<td>Protein binding</td>
<td>155</td>
<td>7.44E-04</td>
<td>2.66E-02</td>
<td>2.18E-02</td>
</tr>
<tr>
<td></td>
<td>HD</td>
<td>3 (CBneg2)</td>
<td>mitochondrion</td>
<td>1164</td>
<td>3.19E-02</td>
<td>2.56E-02</td>
<td>6.17E-02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.45</td>
<td>1.66E-03</td>
</tr>
<tr>
<td></td>
<td>Control (G)</td>
<td>93</td>
<td>Mitochondrion/nuclear lumen</td>
<td>317</td>
<td>1.30E-03</td>
<td>9.85E-03</td>
<td>1.24E-01</td>
</tr>
<tr>
<td></td>
<td>Control (B)</td>
<td>522</td>
<td>Regulation of RNA splicing</td>
<td>64</td>
<td>4.44E-06</td>
<td>6.26E-03</td>
<td>2.52E-01</td>
</tr>
<tr>
<td></td>
<td>Control (G)</td>
<td>25</td>
<td>RNA binding</td>
<td>648</td>
<td>8.02E-01</td>
<td>1.72E-04</td>
<td>9.99E-01</td>
</tr>
</tbody>
</table>

**Table Notes:**

- Adj. FDR: Adjusted False Discovery Rate
- Benjamini FDR: Benjamini False Discovery Rate
- Log2 Fold Change: Logarithm (base 2) of the fold change
- q-value: Corrected p-value
**Table 7.16:** Enrichment of modules from HD vs control blood (Table S9) with TMS in the combined Track- HD and Leiden cohort. Table is sorted by \( p \) (TRACK-TMS).

- \( p(\text{combined-diffexp}) \) – enrichment \( p \)-value for downregulated genes in the combined Track-HD and Leiden sample.
- \( p(\text{TRACK-diffexp}) \) - enrichment \( p \)-value for downregulated genes in the Track-HD sample alone.
- \( p(\text{TRACK-TMS}) \) - enrichment \( p \)-value for genes negatively correlated with TMS in the TRACK-HD sample.

\( BH \) (HD) the Benjamini Hochberg significance value of correlation with HD in Neueder and Bates (Neueder and Bates, 2014) brain expression analysis, corrected for multiple comparisons; \( \text{Cor} \) (HD) the direction and size of correlation of the module with HD in Neueder and Bates.
Mastrokolias and colleagues (Mastrokolias et al., 2015) listed 170 genes significantly associated with TMS, of which 142 passed quality control in our RNA-Seq data. We tested for correlation between these genes and TMS in gene positive subjects from the Track-HD cohort (Table 7.17, and extended version published in Supplementary table S20). 14 genes were nominally significant (p<0.05), which is significantly higher than expected by chance (p=7.89x10\(^{-3}\)). Using the same method as for concordance with Labadoorf et al. (2015a) (see 7.2.9), we compared fold changes in expression of individual genes between Track-HD and Mastrokolias (Mastrokolias et al., 2015). Strikingly, 101 genes showed consistent direction of effect, as measured by log(FC), significantly greater than expected by chance (p=4.78x10\(^{-7}\)).

<table>
<thead>
<tr>
<th>Entrez gene ID</th>
<th>Gene name</th>
<th>log(FC)-Mastrokolias</th>
<th>p (Mastrokolias)</th>
<th>log(FC)-TRACK</th>
<th>p (TRACK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>84263</td>
<td>HSDL2</td>
<td>7.00E-03</td>
<td>4.86E-02</td>
<td>6.00E-03</td>
<td>7.01E-05</td>
</tr>
<tr>
<td>10114</td>
<td>HIPK3</td>
<td>7.00E-03</td>
<td>3.77E-02</td>
<td>-6.00E-03</td>
<td>9.52E-03</td>
</tr>
<tr>
<td>79751</td>
<td>SLC25A22</td>
<td>-7.00E-03</td>
<td>3.58E-02</td>
<td>-3.00E-03</td>
<td>1.14E-02</td>
</tr>
<tr>
<td>366</td>
<td>AQP9</td>
<td>1.20E-02</td>
<td>4.68E-02</td>
<td>7.00E-03</td>
<td>1.41E-02</td>
</tr>
<tr>
<td>79581</td>
<td>SLC52A2</td>
<td>-8.00E-03</td>
<td>4.60E-02</td>
<td>-3.00E-03</td>
<td>1.50E-02</td>
</tr>
<tr>
<td>388228</td>
<td>SBK1</td>
<td>-7.00E-03</td>
<td>4.54E-02</td>
<td>-5.00E-03</td>
<td>1.56E-02</td>
</tr>
<tr>
<td>4773</td>
<td>NFATC2</td>
<td>-1.10E-02</td>
<td>1.69E-02</td>
<td>-8.00E-03</td>
<td>1.83E-02</td>
</tr>
<tr>
<td>2357</td>
<td>FPR1</td>
<td>9.00E-03</td>
<td>2.77E-02</td>
<td>6.00E-03</td>
<td>2.39E-02</td>
</tr>
<tr>
<td>23195</td>
<td>MDN1</td>
<td>-6.00E-03</td>
<td>9.10E-03</td>
<td>-5.00E-03</td>
<td>2.45E-02</td>
</tr>
<tr>
<td>54497</td>
<td>HEATR5B</td>
<td>-7.00E-03</td>
<td>4.68E-02</td>
<td>-3.00E-03</td>
<td>2.69E-02</td>
</tr>
<tr>
<td>84181</td>
<td>CHD6</td>
<td>-5.00E-03</td>
<td>4.93E-02</td>
<td>-4.00E-03</td>
<td>3.30E-02</td>
</tr>
<tr>
<td>729230</td>
<td>CCR2</td>
<td>7.00E-03</td>
<td>3.24E-02</td>
<td>-6.00E-03</td>
<td>4.82E-02</td>
</tr>
<tr>
<td>440503</td>
<td>PLIN5</td>
<td>1.20E-02</td>
<td>1.98E-02</td>
<td>9.00E-03</td>
<td>4.88E-02</td>
</tr>
<tr>
<td>4552</td>
<td>MTRR</td>
<td>2.20E-02</td>
<td>4.54E-02</td>
<td>-3.00E-03</td>
<td>6.22E-02</td>
</tr>
</tbody>
</table>

Table 7.17: Correlation between genes differentially expressed in HD from Mastrokolias et al. (Mastrokolias et al., 2015) and TMS in the Track-HD gene positive subjects. p(Mastrokolias) – p-value for correlation between expression and TMS in Mastrokolias et al. p(TRACK) – p-value for correlation between expression and TMS in TRACK. log2(FC) – the change in log2 (expression) per unit increase of TMS.
The Alzheimer's disease brain transcriptional signature is significantly dysregulated in HD blood

In Alzheimer's disease, an early inflammatory response involving microglia contributes to pathogenesis (Gomez-Nicola et al., 2013, Olmos-Alonso et al., 2016, Hong et al., 2016a). Given the upregulation of immune-related gene sets in HD, we next asked whether co-expression modules dysregulated in Alzheimer's disease (AD) brain were also disrupted in HD blood. Recently the International Genomics of Alzheimer's Disease Consortium (IGAP) identified four modules from the Gibbs et al. (2010) brain co-expression network that showed enrichment of signal in the GWAS of >70,000 late-onset Alzheimer's disease (LOAD) and control subjects (International Genomics of Alzheimer's Disease, 2015). These four modules, each derived from a different brain region, are all involved in the immune response and also significantly enrich for upregulation in our combined HD blood dataset (Table 7.18). The module derived from pontine data was also significantly enriched in both Track-HD and Leiden datasets independently. IGAP identified 151 genes that were present in two or more of these modules and showed the most significant enrichment with LOAD GWAS signal (International Genomics of Alzheimer's Disease, 2015). These 151 genes were also significantly enriched for upregulation in the combined HD blood dataset (p = 2.50 x 10⁻⁴).

<table>
<thead>
<tr>
<th>Module</th>
<th>Brain Region</th>
<th>Number of genes</th>
<th>p (IGAP)</th>
<th>p (Comb)</th>
<th>p (Track-HD)</th>
<th>p (Leiden)</th>
<th>Module Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>Frontal Cortex</td>
<td>109</td>
<td>1.00E-05</td>
<td>1.45E-03</td>
<td>7.06E-03</td>
<td>9.48E-02</td>
<td>GO:0006955 immune response</td>
</tr>
<tr>
<td>99</td>
<td>Temporal Cortex</td>
<td>145</td>
<td>4.00E-05</td>
<td>2.22E-04</td>
<td>5.25E-03</td>
<td>9.13E-02</td>
<td>GO:0006955 immune response</td>
</tr>
<tr>
<td>56</td>
<td>Pons</td>
<td>207</td>
<td>6.00E-05</td>
<td>1.97E-05</td>
<td>2.44E-04</td>
<td>4.19E-02</td>
<td>GO:0006955 immune response</td>
</tr>
<tr>
<td>5</td>
<td>Cerebellum</td>
<td>135</td>
<td>6.80E-04</td>
<td>1.09E-03</td>
<td>4.24E-02</td>
<td>8.15E-02</td>
<td>GO:0006955 immune response</td>
</tr>
</tbody>
</table>

Table 7.18: WGCNA co-expression modules from the Gibbs et al. (2010) control brain expression dataset significantly associated with late-onset Alzheimer's disease (LOAD) in the IGAP GWAS are upregulated in HD blood. The four immune-related modules that were the most significantly enriched modules in LOAD are also significantly enriched for upregulation in the combined Track-HD and Leiden HD blood dataset. p (IGAP) – p value for enrichment of the gene set between LOAD and controls in the IGAP GWAS; p (Combined/Track-HD/Leiden) – p
value for enrichment of the gene set between HD and controls in our HD blood expression dataset.

Zhang et al. (2013) identified co-expression modules that were differentially connected between LOAD and controls. Ten of these were also significantly enriched for upregulation in our HD blood expression dataset (https://www.nature.com/articles/srep44849#supplementary-information) after correction for multiple testing (q < 0.05), with their most significant module, yellow, being particularly highly enriched (combined Track-HD and Leiden p < 1x10^-16). Notably, this module has immune and microglia-specific functions (Zhang et al., 2013). This enrichment for modules from the IGAP GWAS (International Genomics of Alzheimer’s Disease, 2015) and Zhang et al. (2013) in the HD blood transcriptome suggests a shared immune-related mechanism between different neurodegenerative diseases, at least including HD and Alzheimer’s disease.

7.4 Results: Relationship between rate of HD progression and the transcriptome

7.4.1 No differential expression of individual transcripts in HD whole blood with changing rate of disease progression

In order to investigate whether the rate at which an individual with HD progresses is associated with any differences in gene expression in whole blood we first looked at the relationship between our unified Huntington’s disease progression score (described in General Methods, Chapter 2) and transcript levels in the TRACK-HD samples; progression score was used as continuous variables in this analysis. There was no association between individual transcripts and rate of progression that remained significant once corrected for multiple comparisons, however it is apparent when looking at the function of the most significant proteins that many are involved in the cell cycle (Table 7.19).
<table>
<thead>
<tr>
<th>Gene ID</th>
<th>log2 Fold Change</th>
<th>Lfc SE</th>
<th>stat</th>
<th>p-value</th>
<th>Adjusted p-value</th>
<th>Selected information from Genecards about protein function (accessed 02/06/2015) (GeneCards)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSH4</td>
<td>-0.40704</td>
<td>0.102704</td>
<td>19.64606</td>
<td>9.32E-06</td>
<td>0.16819</td>
<td>This gene encodes a member of the DNA mismatch repair mutS family. This member is a meiosis-specific protein that is not involved in DNA mismatch correction, but is required for reciprocal recombination and proper segregation of homologous chromosomes at meiosis I. GO annotations related to this gene include mismatched DNA binding and DNA-dependent ATPase activity.</td>
</tr>
<tr>
<td>RRM2</td>
<td>-0.22329</td>
<td>0.060098</td>
<td>17.14964</td>
<td>3.45E-05</td>
<td>0.311739</td>
<td>Subunit for a ribonucleotide reductase, which catalyses the formation of deoxyribonucleotides from ribonucleotides. Paralog of RRM2B which has been implicated in HD elsewhere (Consortium, 2015a)</td>
</tr>
<tr>
<td>PBK</td>
<td>-0.37372</td>
<td>0.097875</td>
<td>16.35145</td>
<td>5.26E-05</td>
<td>0.31652</td>
<td>Phosphorylates MAP kinase p38. Seems to be active only in mitosis. May also play a role in the activation of lymphoid cells. When phosphorylated, forms a complex with TP53, leading to TP53 destabilization and attenuation of G2/M checkpoint during doxorubicin-induced DNA damage</td>
</tr>
<tr>
<td>ANLN</td>
<td>-0.21363</td>
<td>0.056665</td>
<td>15.57077</td>
<td>7.95E-05</td>
<td>0.358565</td>
<td>GO annotations related to this gene include actin binding and phospholipid binding.</td>
</tr>
<tr>
<td>REN</td>
<td>1.556558</td>
<td>0.421303</td>
<td>14.67092</td>
<td>0.000128</td>
<td>0.436492</td>
<td>Renin catalyzes the first step in the activation pathway of angiotensinogen--a cascade that can result in aldosterone release, vasoconstriction, and increase in blood pressure.</td>
</tr>
<tr>
<td>PKDCC</td>
<td>0.356707</td>
<td>0.106007</td>
<td>14.36259</td>
<td>0.000151</td>
<td>0.436492</td>
<td>Protein kinase which is required for longitudinal bone growth through regulation</td>
</tr>
</tbody>
</table>
of chondrocyte differentiation. Involved in protein transport from the Golgi apparatus to the plasma membrane (By similarity)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Log2(FC)</th>
<th>LFC</th>
<th>SE</th>
<th>Stat</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCAPG</td>
<td>-0.18097</td>
<td>0.052443</td>
<td>14.14435</td>
<td>0.000169</td>
<td>0.436492 Regulatory subunit of the condensin complex, a complex required for conversion of interphase chromatin into mitotic-like condense chromosomes. The condensin complex probably introduces positive supercoils into relaxed DNA in the presence of type I topoisomerases and converts nicked DNA into positive knotted forms in the presence of type II topoisomerases</td>
</tr>
<tr>
<td>SKA3</td>
<td>-0.23214</td>
<td>0.06639</td>
<td>13.72807</td>
<td>0.000211</td>
<td>0.476607 A microtubule-binding subcomplex of the outer kinetochore that is essential for proper chromosome segregation</td>
</tr>
<tr>
<td>DTL</td>
<td>-0.19848</td>
<td>0.060308</td>
<td>12.88651</td>
<td>0.000331</td>
<td>0.653314 E3 ubiquitin-protein ligase complex required for cell cycle control, DNA damage response and translesion DNA synthesis</td>
</tr>
<tr>
<td>CCNB1</td>
<td>-0.10543</td>
<td>0.031051</td>
<td>12.71875</td>
<td>0.000362</td>
<td>0.653314 The protein encoded by this gene is a regulatory protein involved in mitosis.</td>
</tr>
<tr>
<td>BUB1</td>
<td>-0.15348</td>
<td>0.047313</td>
<td>12.00116</td>
<td>0.000532</td>
<td>0.77395 This gene encodes a serine/threonine-protein kinase that play a central role in mitosis</td>
</tr>
<tr>
<td>DLGAP5</td>
<td>-0.19198</td>
<td>0.061511</td>
<td>11.93955</td>
<td>0.000550</td>
<td>0.77395 Potential cell cycle regulator that may play a role in carcinogenesis of cancer cells. Mitotic phosphoprotein regulated by the ubiquitin-proteasome pathway.</td>
</tr>
<tr>
<td>CRYZ</td>
<td>-0.14291</td>
<td>0.041116</td>
<td>11.80478</td>
<td>0.000591</td>
<td>0.77395 Binds NADP and acts through a one-electron transfer process.</td>
</tr>
<tr>
<td>CCDC152</td>
<td>-0.20601</td>
<td>0.056894</td>
<td>11.77476</td>
<td>0.000673</td>
<td>0.77395 CCDC152 (coiled-coil domain containing 152) is a protein-coding gene.</td>
</tr>
<tr>
<td>ZNF684</td>
<td>-0.12772</td>
<td>0.037985</td>
<td>11.3534</td>
<td>0.000753</td>
<td>0.897588 May be involved in transcriptional regulation</td>
</tr>
</tbody>
</table>

**Table 7.19:** Differential expression analysis with rate of HD progression in gene positive members of the TRACK-HD cohort. Log2(FC) – log2 of the ratio of the mean counts in HD and controls. Lfc SE – standard error of the log2(FC). Stat – test statistic.
7.4.2 Pathways are dysregulated in HD subjects with faster vs slower rates of disease progression

We then tested whether transcripts sharing similar functional annotation were dysregulated in relation to rate of progression. Positive and negative directions of correlation were tested separately using GSEA. The same pathway annotations as used for the HD vs control analysis, described above, were used. With a false discovery rate (q-value) threshold of q < 0.05 to correct for multiple testing, there was a significant negative correlation between 119 pathways and rate of HD progression (top 20 pathways shown in Table 7.20). Many of these pathways relate to the cell cycle: faster progressors were found to have lower levels of cell cycle related gene expression in blood compared to slower progressors. Looking at the top genes in the cell cycle related pathways many of the most significant genes overall are seen including MSH4, RRM2, PBK (data not shown). Comparing with the gene set enrichment analysis of the TRACK-HD GWAS (Chapter 3), the pathways which are significantly downregulated in faster progressors are not pathways associated with differential rate of disease progression genetically (Table 7.20).

In contrast, there were no biological pathways with a positive correlation with disease progression, ie no pathways significantly more expressed in faster vs slower progressors (Table 7.20).
<table>
<thead>
<tr>
<th>Direction of effect</th>
<th>Pathway</th>
<th>#genes</th>
<th>Enrichment p-value (directional)</th>
<th>q-value (directional)</th>
<th>p (TRACK) GWAS</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative correlation with HD progression</td>
<td>GO: 280</td>
<td>137</td>
<td>1.11E-15</td>
<td>7.99E-12</td>
<td>6.19E-01</td>
<td>nuclear division</td>
</tr>
<tr>
<td></td>
<td>GO: 7067</td>
<td>7</td>
<td>1.11E-15</td>
<td>7.99E-12</td>
<td>6.19E-01</td>
<td>mitosis</td>
</tr>
<tr>
<td></td>
<td>GO: 48285</td>
<td>150</td>
<td>2.83E-15</td>
<td>1.36E-11</td>
<td>7.95E-01</td>
<td>organelle fission</td>
</tr>
<tr>
<td></td>
<td>GO: 793</td>
<td>481</td>
<td>4.55E-15</td>
<td>1.64E-11</td>
<td>1.51E-01</td>
<td>condensed chromosome</td>
</tr>
<tr>
<td></td>
<td>GO: 51301</td>
<td>211</td>
<td>4.03E-13</td>
<td>1.16E-09</td>
<td>8.07E-01</td>
<td>cell division</td>
</tr>
<tr>
<td></td>
<td>REACTOME 694</td>
<td>3</td>
<td>2.77E-12</td>
<td>6.64E-09</td>
<td>2.64E-01</td>
<td>REACTOME: MITOTIC_M-M_G1_PHASES</td>
</tr>
<tr>
<td></td>
<td>GO: 10564</td>
<td>62</td>
<td>8.48E-12</td>
<td>1.74E-08</td>
<td>9.89E-01</td>
<td>regulation of cell cycle process</td>
</tr>
<tr>
<td></td>
<td>GO: 7059</td>
<td>391</td>
<td>4.12E-11</td>
<td>7.42E-08</td>
<td>9.18E-01</td>
<td>chromosome segregation</td>
</tr>
<tr>
<td></td>
<td>REACTOME 642</td>
<td>25</td>
<td>5.74E-11</td>
<td>9.18E-08</td>
<td>3.20E-01</td>
<td>REACT: M_PHASE</td>
</tr>
<tr>
<td></td>
<td>GO: 775</td>
<td>32</td>
<td>1.11E-10</td>
<td>1.60E-07</td>
<td>3.17E-01</td>
<td>chromosome, centromeric region</td>
</tr>
<tr>
<td></td>
<td>GO: 7346</td>
<td>37</td>
<td>4.31E-10</td>
<td>5.44E-07</td>
<td>9.99E-01</td>
<td>regulation of mitotic cell cycle</td>
</tr>
<tr>
<td></td>
<td>GO: 779</td>
<td>110</td>
<td>4.53E-10</td>
<td>5.44E-07</td>
<td>4.02E-01</td>
<td>condensed chromosome, centromeric region</td>
</tr>
<tr>
<td></td>
<td>GO: 5813</td>
<td>123</td>
<td>8.10E-10</td>
<td>8.53E-07</td>
<td>6.62E-01</td>
<td>centrosome</td>
</tr>
<tr>
<td></td>
<td>GO:1901987</td>
<td>22</td>
<td>8.30E-10</td>
<td>8.53E-07</td>
<td>9.86E-01</td>
<td>regulation of cell cycle phase transition</td>
</tr>
<tr>
<td></td>
<td>GO: 777</td>
<td>130</td>
<td>2.44E-09</td>
<td>2.34E-06</td>
<td>3.35E-01</td>
<td>condensed chromosome kinetochore</td>
</tr>
<tr>
<td></td>
<td>REACTOME 695</td>
<td>169</td>
<td>3.63E-09</td>
<td>3.26E-06</td>
<td>2.34E-01</td>
<td>REACTOME: MITOTIC PROMETAPHASE</td>
</tr>
<tr>
<td></td>
<td>GO: 901988</td>
<td>53</td>
<td>4.33E-09</td>
<td>3.67E-06</td>
<td>9.37E-01</td>
<td>negative regulation of cell cycle phase transition</td>
</tr>
<tr>
<td></td>
<td>GO: 776</td>
<td>24</td>
<td>6.20E-09</td>
<td>4.96E-06</td>
<td>2.02E-01</td>
<td>kinetochore</td>
</tr>
</tbody>
</table>
Table 7.20: Relationship between generic pathways and rate of HD progression showing that while there are multiple pathways significantly downregulated with faster progression, but there are no pathways significantly upregulated with faster progression. The 20 most negatively correlated pathways, and 10 most positively pathways are shown. #genes: number of genes in the pathway; Enrichment p-value- p-value of the normalized enrichment score in the negative (top 20 rows) and positive (bottom 10 rows) direction; p(TRACK): p-value of association of this pathway in the TRACK-HD GWAS (Chapter 3).
A similar pattern of cell cycle pathway enrichment is observed when I used the online software GOrilla for the pathway analysis, this program identifies and visualizes enriched GO terms in ranked lists of genes rather than using the fold change values (Eden et al., 2009). 100 pathways were significantly associated with rate of progression (FDR q-value <0.05) (Figure 7.5 and Table 7.21).
Figure 7.5: Cell cycle pathways expression is associated with rate of HD progression. Diagrammatic illustration of the GO terms enriched in the top vs bottom of the list of transcripts in which transcripts are ranked according to differential expression with rate of HD progression. P-value cut-off for GO terms set at $10^{-7}$. GOrilla accessed 22/11/2017.
<table>
<thead>
<tr>
<th>GO term</th>
<th>Description</th>
<th>P-value</th>
<th>FDR q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:1903047</td>
<td>mitotic cell cycle process</td>
<td>3.72E-14</td>
<td>5.61E-10</td>
</tr>
<tr>
<td>GO:0022402</td>
<td>cell cycle process</td>
<td>1.25E-13</td>
<td>9.41E-10</td>
</tr>
<tr>
<td>GO:0007346</td>
<td>regulation of mitotic cell cycle</td>
<td>5.05E-11</td>
<td>2.54E-7</td>
</tr>
<tr>
<td>GO:0051726</td>
<td>regulation of cell cycle</td>
<td>7.23E-11</td>
<td>2.72E-7</td>
</tr>
<tr>
<td>GO:0007088</td>
<td>regulation of mitotic nuclear division</td>
<td>1.16E-10</td>
<td>3.5E-7</td>
</tr>
<tr>
<td>GO:0051301</td>
<td>cell division</td>
<td>1.46E-10</td>
<td>3.67E-7</td>
</tr>
<tr>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>3.99E-10</td>
<td>8.59E-7</td>
</tr>
<tr>
<td>GO:1901990</td>
<td>regulation of mitotic cell cycle phase transition</td>
<td>6.71E-10</td>
<td>1.26E-6</td>
</tr>
<tr>
<td>GO:0051783</td>
<td>regulation of nuclear division</td>
<td>7.43E-10</td>
<td>1.24E-6</td>
</tr>
<tr>
<td>GO:1901987</td>
<td>regulation of cell cycle phase transition</td>
<td>1.46E-9</td>
<td>2.2E-6</td>
</tr>
<tr>
<td>GO:0010564</td>
<td>regulation of cell cycle process</td>
<td>1.73E-9</td>
<td>2.37E-6</td>
</tr>
<tr>
<td>GO:0007049</td>
<td>cell cycle</td>
<td>2.55E-9</td>
<td>3.2E-6</td>
</tr>
<tr>
<td>GO:0007059</td>
<td>chromosome segregation</td>
<td>2.51E-8</td>
<td>2.91E-5</td>
</tr>
<tr>
<td>GO:0006260</td>
<td>DNA replication</td>
<td>2.86E-8</td>
<td>3.08E-5</td>
</tr>
<tr>
<td>GO:0007052</td>
<td>mitotic spindle organization</td>
<td>3.28E-8</td>
<td>3.3E-5</td>
</tr>
<tr>
<td>GO:0045786</td>
<td>negative regulation of cell cycle</td>
<td>7.45E-8</td>
<td>7.02E-5</td>
</tr>
<tr>
<td>GO:0051276</td>
<td>chromosome organization</td>
<td>1.03E-7</td>
<td>9.16E-5</td>
</tr>
<tr>
<td>GO:0090068</td>
<td>positive regulation of cell cycle process</td>
<td>1.5E-7</td>
<td>1.26E-4</td>
</tr>
<tr>
<td>GO:0045787</td>
<td>positive regulation of cell cycle</td>
<td>1.97E-7</td>
<td>1.56E-4</td>
</tr>
<tr>
<td>GO:0000075</td>
<td>cell cycle checkpoint</td>
<td>2.08E-7</td>
<td>1.57E-4</td>
</tr>
<tr>
<td>GO:1902850</td>
<td>microtubule cytoskeleton organization involved in mitosis</td>
<td>2.34E-7</td>
<td>1.68E-4</td>
</tr>
<tr>
<td>GO:0033044</td>
<td>regulation of chromosome organization</td>
<td>2.67E-7</td>
<td>1.83E-4</td>
</tr>
</tbody>
</table>

Table 7.21: Cell cycle pathways are enriched in GOrilla analysis of ranked transcripts from the TRACK-HD progression differential progression analysis. Transcripts with a FDR q-value < 2.0E-4 are shown. 'P-value' is the enrichment p-value. 'FDR q-value' is the correction of the p-value for multiple testing using the Benjamini and Hochberg (1995) method.
### 7.4.3 Gene co-expression modules and rate of HD Progression

GSEA for brain co-expression modules was applied to the HD progression differential expression dataset. Firstly we looked at HD modules from the Neueder and Bates paper (Neueder and Bates, 2014) which applied WGCNA to obtain 124 modules from the Hodges et al (Hodges et al., 2006) HD and control brain dataset. Slower progressors (negative correlation enrichment) have higher levels of transcripts from protein transport and folding modules. While in faster progressors (positive correlation enrichment) transcripts involved in ion channels, synapses and mitochondrial biology are enriched (Table 7.22).

<table>
<thead>
<tr>
<th>Direction of correlation enrichment</th>
<th>Module(s)</th>
<th>Brain Region</th>
<th>#genes</th>
<th>Enrichment p-value</th>
<th>cor(HD)</th>
<th>BH (HD)</th>
<th>Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>89 FC_BA 4</td>
<td>FC_BA 4</td>
<td>390</td>
<td>4.14E-06</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>119 FC_BA 9</td>
<td>FC_BA 9</td>
<td>696</td>
<td>1.90E-05</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>66 CN</td>
<td>CN</td>
<td>2624</td>
<td>5.40E-05</td>
<td>-</td>
<td>0.800</td>
<td>6.03E-15 Synapse / ion channels</td>
</tr>
<tr>
<td></td>
<td>25 CB</td>
<td>CB</td>
<td>362</td>
<td>6.86E-05</td>
<td>-</td>
<td>0.450</td>
<td>1.66E-03 mitochondrion</td>
</tr>
<tr>
<td></td>
<td>46 CN</td>
<td>CN</td>
<td>1016</td>
<td>5.71E-04</td>
<td>0.744</td>
<td>4.03E-12</td>
<td>Regulation of transcription/ mRNA/ chromatin modification</td>
</tr>
<tr>
<td></td>
<td>26 CB</td>
<td>CB</td>
<td>247</td>
<td>2.55E-03</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>100 FC_BA 4</td>
<td>FC_BA 4</td>
<td>78</td>
<td>2.66E-03</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>44 CB</td>
<td>CB</td>
<td>628</td>
<td>6.37E-03</td>
<td>0.336</td>
<td>2.90E-02</td>
<td>Zinc finger binding / chromatin modification</td>
</tr>
<tr>
<td></td>
<td>98 FC_BA 4</td>
<td>FC_BA 4</td>
<td>1350</td>
<td>1.82E-02</td>
<td>-</td>
<td>3.35E-02</td>
<td>Glycolysis / protein transport</td>
</tr>
<tr>
<td></td>
<td>75 FC_BA</td>
<td>FC_BA</td>
<td>44</td>
<td>3.75E-02</td>
<td>-</td>
<td>1.21E-01</td>
<td>Fibronectin</td>
</tr>
</tbody>
</table>
Table 7.22: Correlation enrichment between HD modules from Neueder & Bates (Neueder and Bates, 2014) and differential transcription according to progression.

Next we looked for concordance with the Gibbs control brain modules. Slower progressors have higher levels of proteasome, mRNA, transcription, protein modification and transport modules compared to faster progressors (Table 7.23). Faster progressors have higher levels of transcripts from Proline-rich regions, and those involved with the golgi apparatus (Table 7.23).

<table>
<thead>
<tr>
<th>Direction of correlation enrichment</th>
<th>Module</th>
<th>Brain region</th>
<th>#genes</th>
<th>Enrichment p</th>
<th>Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>113</td>
<td>TCTX</td>
<td>619</td>
<td>&lt;10-16</td>
<td>Proteasome/acetylation/mRNA metabolic process</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>CB</td>
<td>788</td>
<td>3.16E-15</td>
<td>Protein</td>
</tr>
<tr>
<td>Rank</td>
<td>Tissue</td>
<td>Gene</td>
<td>FDR</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>FCTX</td>
<td>1001</td>
<td>1.07E-13</td>
<td>Membrane enclosed lumen</td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>Pons</td>
<td>1182</td>
<td>5.44E-13</td>
<td>Transcription/acetylation/protein folding/histone deacetylase</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>CB</td>
<td>591</td>
<td>7.42E-13</td>
<td>Protein transport/Golgi apparatus/proteolysis</td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>TCTX</td>
<td>422</td>
<td>2.69E-09</td>
<td>Nucleus</td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>Pons</td>
<td>476</td>
<td>1.88E-08</td>
<td>KRAB/phosphoprotein/GTP binding</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>FCTX</td>
<td>868</td>
<td>5.57E-08</td>
<td>Small conjugating protein ligase/golgi apparatus</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>Pons</td>
<td>432</td>
<td>1.30E-05</td>
<td>Phosphoprotein/lytic vacuole</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>TCTX</td>
<td>1135</td>
<td>7.84E-05</td>
<td>Vesicle-mediated transport/cytoplasm</td>
<td></td>
</tr>
</tbody>
</table>

**Negative**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tissue</th>
<th>Gene</th>
<th>FDR</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>CB</td>
<td>831</td>
<td>2.71E-12</td>
<td>Compositionally biased region: Pro-rich</td>
</tr>
<tr>
<td>95</td>
<td>TCTX</td>
<td>361</td>
<td>1.57E-08</td>
<td>Golgi membrane/Compositionally biased region: Pro-rich</td>
</tr>
<tr>
<td>48</td>
<td>FCTX</td>
<td>647</td>
<td>2.24E-08</td>
<td>Golgi cistern/transcription corepressor activity/Pro-rich</td>
</tr>
<tr>
<td>85</td>
<td>Pons</td>
<td>556</td>
<td>1.18E-05</td>
<td>Lipid binding/nucleoside-triphasatase regulator activity</td>
</tr>
<tr>
<td>66</td>
<td>Pons</td>
<td>136</td>
<td>1.23E-05</td>
<td>Phosphoprotein</td>
</tr>
<tr>
<td>63</td>
<td>Pons</td>
<td>1542</td>
<td>1.40E-05</td>
<td>DNA binding/diencephalon development/protease</td>
</tr>
<tr>
<td>111</td>
<td>TCTX</td>
<td>926</td>
<td>1.34E-04</td>
<td>MAPKinase signalling pathway/chromatin remodelling</td>
</tr>
<tr>
<td>39</td>
<td>FCTX</td>
<td>1325</td>
<td>2.25E-04</td>
<td>DNA binding/hormone/disulphide bond</td>
</tr>
<tr>
<td>103</td>
<td>TCTX</td>
<td>72</td>
<td>3.87E-04</td>
<td>nucleoplasm part</td>
</tr>
<tr>
<td>112</td>
<td>TCTX</td>
<td>731</td>
<td>5.23E-04</td>
<td>Calcium binding/actin binding/tight junction/endocytosis</td>
</tr>
</tbody>
</table>
### 7.4.4 Comparison of HD progression results to the HD vs control WGCNA results

Comparing the Neueder & Bates HD modules and progression results to the HD vs control results, Module 76 and 110 significantly enriched for control-manifest downregulated genes and Module 111 significantly enriched for control-manifest upregulated genes, however overall there was no consistent direction of association seen in either the positive or negative direction (Table 7.22).

For the Gibbs modules, modules 23, 26, 45 and 115, which are among those with the strongest negative correlation enrichment, are also significantly enriched for control-manifest downregulated genes; by contrast there is no overlap in the positive correlation enrichment modules (Table 7.23).

### 7.4.5 Attempted replication of TRACK-HD progression RNAseq results in the LUMC dataset

Given that a subset of the individuals from the LUMC dataset had serial data we investigated whether we could replicate the result showing an enrichment of cell cycling related pathways in TRACK-HD slower progressors in the LUMC samples. We therefore developed a progression score using the available data from LUMC as described in General Methods (Chapter 2.5.3). Results for TRACK-HD and LUMC were not meta-analysed for the progression RNAseq analysis due to concerns that the progression scoring methods were not sufficiently similar to justify it.

### 7.4.6 No individual transcripts are differentially expressed according to rate of HD progression in the LUMC cohort

We investigated whether any transcripts were differentially expressed with rate of progression according to the LUMC atypical severity score: no transcripts were significantly differentially expressed (Table 7.24).

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>log2FoldChange</th>
<th>Standard error of LFC</th>
<th>stat</th>
<th>p-value</th>
<th>Brief function from GeneCards</th>
</tr>
</thead>
<tbody>
<tr>
<td>TME</td>
<td>0.2414876</td>
<td>0.058</td>
<td>17.72</td>
<td>2.55E-2</td>
<td>Enhances production of pro-inflammatory</td>
</tr>
<tr>
<td>Gene</td>
<td>Log2FoldChange</td>
<td>p-value</td>
<td>FoldChange</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>---------</td>
<td>------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>M9</td>
<td>152</td>
<td>0.05</td>
<td>727</td>
<td>cytokines induced by TNF, IL1B, and TLR ligands.</td>
<td></td>
</tr>
<tr>
<td>CHST2</td>
<td>-0.2747902</td>
<td>0.067</td>
<td>503</td>
<td>Among its related pathways are Disease and Metabolism. GO annotations related to this gene include <em>sulfotransferase activity</em> and <em>N-acetylglucosamine 6-O-sulfotransferase activity</em>.</td>
<td></td>
</tr>
<tr>
<td>OCM</td>
<td>0.4134942</td>
<td>0.102</td>
<td>0.05</td>
<td>Oncomodulin is a high-affinity calcium ion-binding protein. It belongs to the superfamily of calmodulin proteins, also known as the EF-hand proteins.</td>
<td></td>
</tr>
<tr>
<td>PTPRS</td>
<td>-0.3103974</td>
<td>0.074</td>
<td>553</td>
<td>PTPs are known to be signalling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation.</td>
<td></td>
</tr>
<tr>
<td>TRA2A</td>
<td>-0.1048107</td>
<td>0.025</td>
<td>705</td>
<td>This gene is a member of the transformer 2 homolog family and encodes a protein with several RRM (RNA recognition motif) domains. This phosphorylated nuclear protein binds to specific RNA sequences and plays a role in the regulation of pre-mRNA splicing.</td>
<td></td>
</tr>
<tr>
<td>NUAK1</td>
<td>-0.6623118</td>
<td>0.168</td>
<td>239</td>
<td>Serine/threonine-protein kinase involved in various processes such as cell adhesion, regulation of cell ploidy and senescence, cell proliferation and tumour progression.</td>
<td></td>
</tr>
<tr>
<td>LPCAT1</td>
<td>-0.1464862</td>
<td>0.037</td>
<td>763</td>
<td>Lysophosphatidylcholine (LPC) acyltransferase (LPCAT; EC 2.3.1.23) catalyses the conversion of LPC to phosphatidylcholine (PC)</td>
<td></td>
</tr>
<tr>
<td>PRDM16</td>
<td>-0.9408971</td>
<td>0.283</td>
<td>15</td>
<td>Binds DNA and functions as a transcriptional regulator.</td>
<td></td>
</tr>
<tr>
<td>ZSCA N32</td>
<td>-0.0716239</td>
<td>0.018</td>
<td>526</td>
<td>GO annotations related to this gene include sequence-specific DNA binding RNA.</td>
<td></td>
</tr>
</tbody>
</table>
This gene encodes a protein which contains a death domain. Death domain-containing proteins function in signaling pathways and formation of signaling complexes, as well as the apoptosis pathway.

<table>
<thead>
<tr>
<th>GO term</th>
<th>Description</th>
<th>P-value</th>
<th>FDR q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0043434</td>
<td>response to peptide hormone</td>
<td>4.75E-7</td>
<td>7.15E-3</td>
</tr>
<tr>
<td>GO:0070887</td>
<td>cellular response to chemical stimulus</td>
<td>6.18E-7</td>
<td>4.66E-3</td>
</tr>
<tr>
<td>GO:0034164</td>
<td>negative regulation of toll-like receptor 9 signalling pathway</td>
<td>2.22E-6</td>
<td>1.12E-2</td>
</tr>
<tr>
<td>GO:0032870</td>
<td>cellular response to hormone stimulus</td>
<td>5.3E-6</td>
<td>2E-2</td>
</tr>
<tr>
<td>GO:1901698</td>
<td>response to nitrogen compound</td>
<td>7.53E-6</td>
<td>2.27E-2</td>
</tr>
<tr>
<td>GO:0032687</td>
<td>negative regulation of interferon-alpha production</td>
<td>8.06E-6</td>
<td>2.02E-2</td>
</tr>
<tr>
<td>GO:1901652</td>
<td>response to peptide</td>
<td>8.18E-6</td>
<td>1.76E-2</td>
</tr>
<tr>
<td>GO:0071310</td>
<td>cellular response to organic substance</td>
<td>8.33E-6</td>
<td>1.57E-2</td>
</tr>
<tr>
<td>GO:0006793</td>
<td>phosphorus metabolic process</td>
<td>8.86E-6</td>
<td>1.48E-2</td>
</tr>
<tr>
<td>GO:0010243</td>
<td>response to organonitrogen compound</td>
<td>9.1E-6</td>
<td>1.37E-2</td>
</tr>
</tbody>
</table>

Table 7.25: Ten pathways most enriched in a GOrilla pathway analysis of the differential transcription in the LUMC samples according to cross-sectional severity score. 'P-value' is the enrichment p-value. 'FDR q-value' is the correction of the p-value for multiple testing using the Benjamini and Hochberg (1995) method.
Figure 7.6: Pathways related to progression in the LUMC cohort. Diagrammatic illustration using GOrilla (Eden et al., 2009) of the GO terms enriched in the top vs bottom of the list of transcripts in which transcripts are ranked according to differential expression with rate of HD progression in the LUMC cohort. P-value cutoff for GO terms set at $10^{-6}$. GOrilla accessed 22/11/2017.

7.5 Discussion

HD research has focused on the brain as the most conspicuous clinical features can be clearly linked to progressive degeneration of specific brain regions (van der Burg et al., 2009, Bates et al., 2015). However, HD is a systemic condition with peripheral expression of mutant huntingtin directly driving abnormalities such as immune dysfunction, metabolic derangement and transcriptional dysregulation that contribute to onset, progression, quality of life and mortality (van der Burg et al., 2009, Carroll et al., 2015, Aziz et al., 2018).
In this chapter I have discussed work in which we conducted RNA-Seq of whole blood in two independent cohorts of HD patients. Using gene set enrichment analysis (GSEA) with publicly-available pathway databases and WGCNA modules from HD and control brain datasets, dysregulated gene sets were identified in HD blood that were replicated in both independent cohorts. These correspond to the most significantly dysregulated modules in caudate nucleus, the most prominently affected region in HD brain. This suggests mutant huntingtin drives a common pathogenic signature in both blood and brain.

RNA-Seq more comprehensively and accurately quantifies mRNA than hybridisation-based microarrays or tag-based methods (Costa et al., 2010). However, it is perhaps unsurprising that there was no significant differential expression of individual transcripts by disease stage or state in either the independent or combined cohorts (Table 7.3) given that the major cell types known to contribute to symptoms are not present and the haematogenous cells known to be dysfunctional in HD, such as monocytes and macrophages (Bjorkqvist et al., 2008, Wild et al., 2011), constitute only a small proportion of circulating cells (Whitney et al., 2003). The variation of gene expression in blood with age, gender, cell type and time of day is also likely to add to the sample heterogeneity (Whitney et al., 2003, Horvath et al., 2012). Our results are consistent with previous studies that have shown weak correlation at the transcript level between blood and brain (Cai et al., 2010).

Despite these limitations, in the disease status experiments, gene set enrichment analysis identified significantly overlapping dysregulated pathways in the Track-HD and Leiden HD blood datasets, even though they differed in age and disease severity. The observed upregulation of immune-related pathways in HD is consistent with that previously identified in transcriptional and functional studies (Mastrokoliás et al., 2015, Carroll et al., 2015, van der Burg et al., 2009). HD patients are known to have immune dysfunction, both in the central nervous system (CNS) with microglial activation (Tai et al., 2007), and peripherally with elevated pro-inflammatory cytokines in premanifest carriers up to 16 years before predicted onset (Bjorkqvist et al., 2008, Wild et al., 2011). The migration of phagocytic cells is impaired in HD (Kwan et al., 2012b, Träger et al., 2015) and patient-derived monocytes are hyperactive on stimulation, an effect reduced by HTT lowering (Bjorkqvist et al., 2008). Modulation of the peripheral immune system with a type 2 cannabinoid receptor (CB2) agonist (Bouchard et al., 2012) or bone marrow transplantation (Kwan et al., 2012a) can increase lifespan and reduce motor deficits and synaptic loss in HD mouse models.
RNA processing pathways were downregulated in HD, which is congruent with known decreases in miRNAs and altered expression of key miRNA processing enzymes in HD (Seredenina and Luthi-Carter, 2012, Langfelder et al., 2018). Consistent with effects we observe on pathways involved in energy metabolism, mitochondrial ATP is reduced in HD brain (Mochel et al., 2012) and blood (Seong, 2005), and PGC-1α, a member of the dysregulated ATP metabolic process pathway, is a key protective regulator of mitochondrial genes that is repressed in HD mouse models (Cui et al., 2006, Chaturvedi et al., 2010).

Downregulation of genes involved in DNA repair is likely to be relevant to somatic expansion that may influence disease onset and progression (Jonson et al., 2013, Hensman Moss et al., 2017b, Consortium, 2015a). The signature of pathway dysregulation we identified in HD whole blood significantly overlaps with that recently found in unstimulated HD monocytes (Miller et al., 2016a). This enrichment was driven primarily by upregulation of immune pathways, as might be expected given that Miller et al. (2016a) isolated myeloid cells.

To overcome the annotation gap commonly observed with publicly-derived pathway databases and to investigate whether gene expression changes from HD brain are also present in blood, we performed GSEA using brain co-expression networks derived from HD (Neueder and Bates, 2014) and control (Gibbs et al., 2010, Braineac, 2016) subjects. Several HD brain modules were significantly dysregulated in HD blood, suggesting a common signature of transcriptional dysregulation between blood and brain. Brain modules upregulated in blood were enriched for immune-related genes, confirming the results of our pathway analysis. Strikingly, two of the modules most significantly dysregulated in HD caudate, 48 (CNpos2) and 66 (CNneg1), were also significantly dysregulated in the same direction in both independent blood datasets. Compared with other brain regions, the caudate has the largest number of expression changes and the highest correlation with HD (Neueder and Bates, 2014). Module 48 (CNpos2), the second most significantly upregulated module in caudate, is enriched for transcriptional regulators, chromatin modifiers and genes involved in mRNA processing (Neueder and Bates, 2014). We also find this module to be significantly enriched for immune response genes, giving further support to the pathway results. Module 66 (CNneg1), the most significantly downregulated module in caudate, contains genes involved in neuronal function, particularly synaptic function and plasticity, and ion channels. Around half of its hub genes are implicated in synaptic function and all were significantly downregulated in Hodges et al. (2006). Though synapses are not present in blood, synaptic genes may be dysregulated in circulating cells without significant pathogenic impact, or alternatively they may serve distinct functions in blood cells. Indeed, Cai et al. (2010) found that the synaptic module was well
preserved between brain and blood. We also found that gene expression and pathway
dysregulation from HD prefrontal cortex (Labadorf et al., 2015a) was replicated in HD blood.
The high degree of replication increases confidence in the shared signal between blood and
brain. This overlap is important for future studies: blood is a readily available tissue, our
findings support the use of blood from people who have Huntington’s disease to give insights
on HD brain.

Mina et al. (2016) performed WGCNA on the Leiden blood sample, finding modules related to
immune response that were associated with TFC and motor score. Furthermore, by comparing
biological annotations of their HD blood modules with those they derived from the Hodges et
al. (2006) brain expression data, they showed a common signature between blood and
caudate related to immune response. These analyses, using different methodology to ours,
give further support to our conclusions.

Analysis of the impact of rate of disease progression on the transcriptome of HD gene positive
individuals was investigated in 117 gene positive TRACK-HD, and interestingly suggested that
cell cycle transcripts are markedly and significantly less expressed in faster progressors
compared to slower progressors. This is particularly intriguing given that circulating blood is
largely a post-mitotic tissue, however, as we have established transcriptional dysregulation in
blood reflects that occurring in brain. These results were not replicated in our progression
analysis in the LUMC cohort, however there are several limitations of the study which could
be responsible for this. The progression statistic that we developed for the LUMC cohort was
based on very limited phenotypic data, just TMS and TFC from two time points, in contrast to
the TRACK-HD progression measure which was based on 24 variables over four time points.
This means that the LUMC measure is likely to be much less robust and reproducible, such
that much larger sample sizes would be needed to use with reasonable study power.
However, instead we had fewer samples in the LUMC cohort. Thus our power to investigate
progression related changes in transcription was very limited in the LUMC analysis. HD studies
with both clinical and biosample data available were previously limited. However Enroll-HD, a
global study collecting annual phenotypic data, and including biosamples may be worth
considering to further address this question in the future.

Another limitation is that the faster progressors also tended to have more advanced stage
disease at time of sampling that the slow progressors, thus disease stage could confound our
result, though were this to be the case one would expect to find a similar result to the analysis
looking at the effect of disease stage, which was not the case.
While any change in cell cycle transcripts could be a direct effect of transcriptional dysregulation, it could also be related to the downstream effects of huntingtin. Expression of misfolded proteins such as HTT often leads to the formation of intracellular aggregates (Ross and Tabrizi, 2011). When the capacity of the autophagy and ubiquitin-proteosome systems are exceeded a large juxtanuclear aggregate known as the aggresome forms (Lu et al., 2015). Lu and colleagues (Lu et al., 2015) have established that perinuclear aggresome accumulation is associated with abnormal nuclear morphology and DNA double-strand breaks resulting in cell-cycle arrest via the phosphorylated p53 dependent pathway. Aggresomes can also have a detrimental effect on mitosis by steric interference with chromosome alignment, chromosome positioning and spindle formation. It would be interesting to investigate whether cells from faster and slower progressing subjects have a difference in their cell turnover, and look at the levels of aggresomes as it would be plausible for faster progressors to have higher levels of aggresomes via an accelerated disease process, and that this could result in lower cell cycle gene expression and lower rates of mitosis. Huntingtin has a highly conserved role in modulating mitotic spindle orientation through the dynein/dynactin complex (Godin et al., 2010). However given that there was no significant difference in HTT expression with change in progression rate, the change in level of cell-cycle related transcripts does not seem likely to be driven by this.

A particularly intriguing result presented in this chapter is the evidence of the shared immune transcriptomic signature between Alzheimer Disease (AD) and HD. Alzheimer Disease (AD) is the most common cause of dementia, typically presenting with a progressive loss of cognitive function and memory (Guerreiro et al., 2013). Like HD, AD is associated with protein misfolding and aggregation: it is characterized by amyloid plaques and accumulations of tau called neurofibrillary tangles (Fitzpatrick et al., 2017). In AD, amyloid plaques are surrounded by chronically activated microglia (Gomez-Nicola et al., 2013, Olmos-Alonso et al., 2016) and GWA studies have identified immune-related genes as risk factors for late onset Alzheimer Disease (LOAD) (Wyss-Coray and Rogers, 2012). Recently Hong et al. (2016a) showed that early in the disease process, before plaque formation, microglia and complement activation drive synaptic loss, a process that resembles and may reflect reactivation of developmental synaptic pruning (Hong et al., 2016b). In HD blood we found significant upregulation of immune modules associated with AD in the IGAP GWAS (International Genomics of Alzheimer's Disease, 2015), a subset of genes with shared membership of several of these modules, and the most significant immune and microglia-related modules from Zhang et al. (2013). In a co-expression network generated from prefrontal cortex of 194 HD patients,
Zhang et al. (2013) found that their most significant immune and microglia module was well conserved, though was not significantly dysregulated in HD and did not correlate with CAG repeat length. This may be because cortex shows less severe pathology and transcriptional dysregulation than caudate (Hodges, 2006). Overlapping immune upregulation in HD and AD suggests these two distinct neurodegenerative diseases share some common pathogenic mechanisms, with parallel signalling cascades initiated in macrophages upon pathogen phagocytosis and in microglia involved in synaptic pruning in the AD brain (Hong et al., 2016a). St-Amour and colleagues (St-Amour et al., 2017) recently showed that mixed proteinopathies occur in late-stage human HD brain: tau is abnormally phosphorylated and is aberrantly spliced, and there is increased aggregation of TDP-43, α-synuclein and phosphorylated-Tau as HD progresses, possibly pointing to common mechanisms leading to the abnormal accumulation of aggregation-prone proteins in neurodegenerative diseases.

In this chapter I have shown that transcriptional analysis of deeply phenotyped subjects from the TRACK-HD cohort has yielded important insights about the effect of disease status and disease progression on the expression pattern in patients with HD.
Chapter 8: Conclusion and future directions

It has been the aim of this thesis to better understand the genetic factors underpinning phenotypic diversity in neurodegenerative diseases, particularly those caused by repeat expansion mutations. I used a variety of genetic strategies to look for variants, both genetic and transcriptomic, which are associated with rate of progression and age of disease onset in Huntington’s disease and the polyglutamine spinocerebellar ataxias. In Chapter 3 I have presented work in which I successfully identified a locus in MSH3 as being associated with HD progression ($1.58 \times 10^{-8}$) (Hensman Moss et al., 2017b), and went on, in Chapter 5, to identify the likely causal variant underlying this genetic signal. Further study of this variant I identified has already been done by members of the Tabrizi lab and collaborators and this work is the topic of a manuscript recently accepted for publication in Brain (Flower et al., 2019). Work on the functional impact of this variant, and confirmation that it is the presence of the MSH3 repeat, and not close-by variants in high linkage disequilibrium would be of interest.

MSH3 is a DNA repair gene which is involved in the repair of DNA mismatches: loops of around 10 mispaired bases in the DNA (Figure 3.15). The results from Chapters 3 and 4 show that DNA repair pathways more broadly are associated with the rate of progression and age at onset, in not only Huntington’s disease, but across a range of disorders caused by CAG repeat expansion mutations. This suggests a common mechanism, acting at the level of the CAG repeat tract rather than being a protein/disease specific mechanism. Because of their repetitive nature, CAG repeats are susceptible to forming unusual structures such as imperfect hairpins and slipped strand structures to which DNA repair proteins are recruited (Mirkin, 2007). Through a process of aberrant repair at these repetitive regions additional bases are either added or removed, resulting in somatic instability of the CAG repeats. Repeat expansion occurs in dividing and non-dividing cells, and is tissue specific, cell specific, and disease specific (Jones et al., 2017). Somatic expansion of the CAG repeat tract has been discussed in this thesis as the likely mechanism through which DNA repair variants modulate the course of disease: in fast progressing individuals aberrant DNA repair mechanisms result in expansion of the CAG repeat tract in susceptible tissues (Figure 3.16).

Proteins with larger CAG repeat tracts are associated with higher disease-related toxicity, resulting in faster disease progression and slower onset (Bates et al., 2015). While it is proteins of the DNA mismatch repair pathway which have been primarily implicated by the pathway analysis, DNA repair proteins other than those involved in mismatch repair are also involved. The bidirectional associations at FAN1 (Chapter 3, Chapter 5, (Bettencourt et al.,
2016, Hensman Moss et al., 2017b, GeM-HD-Consortium, 2015), which is involved in
interstrand cross-link repair as well as interacting with mismatch repair proteins,
demonstrates that members of various DNA repair pathways are implicated in somatic
instability of CAG repeats. Indeed, it seems that we should view the trinucleotide repeat
expansion pathway as separate pathway, involving proteins associated with various DNA
repair pathways, and likely having its own unique mechanism specific to repetitive DNA
(Figure 8.1).
Figure 8.1: The main DNA damage response (DDR) pathways with the proteins suspected to be involved in each. The postulated role of the DDR in trinucleotide repeat instability is shown on the right, shaded in yellow, the proteins listed have been implicated by genetic data. Pathways of DSB repair are in blue-shaded area; pathways of SSB repair are in red-shaded area. Main targets of drug development are in red. Figure prepared by me, after Brown et al, 2017.
The GeM-HD Consortium have been collecting many more samples for further genetic analysis of the determinants of AAO in HD. Many of these new samples come from the ENROLL-HD study, a worldwide study with annual study visits collecting biosamples, cognitive, motor and psychiatric data (Landwehrmeyer et al., 2016). Thus not only are larger studies looking into modifiers of AAO now possible and underway, but the greater clinical data in ENROLL may enable further analysis of modifiers of HD progression. There were several peaks just under the genome-wide significance level in the published version of the GeM-HD study (GeM-HD-Consortium, 2015), and further data in which more loci attain significance has been presented at conferences and meetings. By conducting larger studies looking for genetic modifiers of onset and progression in HD it is hoped that greater knowledge of the genetics will improve our understanding of the key molecular events accelerating/ decelerating pathology.

There is, quite correctly, caution about applying knowledge of an individual patients’ genotype in terms of genetic modifiers to the clinic as part of genetic counselling: given the residual variability it would not be possible to accurately predict how an individual patient’s HD will progress. However, with a raft of therapeutic studies underway in HD, it may be of value for the genetic data to be used in these trials. Polygenic scores, similar to those described in Chapter 4, could be used to predict which subjects are likely to progress faster/slower based on their genetics. This may be particularly applicable to early phase studies in which sample sizes are low and thus results could be subject to bias if fast/slow progressing subjects were overrepresented on one arm of the study.

There is considerable scope for further valuable work looking for genetic modifiers of the polyglutamine SCAs. An unbiased genetic screen looking for modifiers of age of onset across the polyglutamine SCAs would be of great interest. This would enable not only further analysis of the role of DNA repair pathway variants, but also to examine whether any other novel areas of biology are implicated. Efforts to collect a much larger sample of cases is underway by collaborators. The MSH3 repeat identified as a modifier of HD progression in this thesis was not tagged in the genetic analysis looking at whether DNA repair gene variants modify onset in the polyglutamine SCAs. Sequencing of SCA subjects to look for this variant and determine its effect on AAO would further expand our understanding of the role of this variant in repeat disorders.

The C9orf72 hexanucleotide repeat expansion is observed with much longer pathological repeats than the repeats associated with HD or the SCAs (repeat numbers of thousands rather than tens or hundreds), meaning that accurate sizing of these large repeats is challenging. In
Chapter 6 I show that intergenerational instability is observed in families with high-normal C9orf72 repeat lengths. Somatic instability of the C9orf72 repeat has also been observed (Suh et al., 2015, McGoldrick et al., 2018), and length of repeat was inversely correlated with disease duration in those with an FTD phenotype in a small study (Suh et al., 2015). In HD, intermediate alleles with 27–35 repeats are not associated with disease symptoms but can expand into the affected range upon (predominantly paternal) germline transmission and thus cause HD in offspring. HTT CAG repeat expansions appear to occur before meiosis in dividing Spermatogonia, or after meiosis is complete in differentiating germ cells (McMurray, 2010). It is not fully established whether there are mechanistic differences between expansions of long and short repeats, or whether the same pathways for expansion are used in different cell types. It would be interesting to investigate somatic instability in C9orf72 further, and look at whether the DNA repair variants which modulate progression/onset in polyglutamine repeat disorders have a similar phenotypic effect in C9orf72 associated ALS/FTD.

The preliminary work I present in Chapter 7 looking for transcriptomic signatures associated with rate of disease progression suggested that cell cycle transcripts are significantly less expressed in fast compared to slow progressing subjects, which may be related to perinuclear aggresome accumulation and resultant DNA damage and cell cycle arrest as discussed in Chapter 7. It would be interesting to adapt the REGISTRY progression score for use in ENROLL-HD, enabling the identification of fast progressing subjects using high quality phenotypic data who are still actively engaged in research studies. A replication cohort for my work looking at the association between rate of disease progression and the transcriptome could be identified, and potentially biosamples collected to investigate cell cycle rates further. Blood is largely a post-mitotic tissue, but using buccal swabs epithelial cells and leucocytes can be readily collected (Theda et al., 2018): cell turnover could be investigated in buccal endothelial cells in fast vs slow progressing HD subjects.

A common theme underpinning this thesis has been the value of high quality phenotypic data to assist in genetic analysis of neurodegenerative diseases: conditions which cause progressive loss of brain functions and overlapping clinical syndromes. The value of clinical phenotyping has been doubted by some, due to the availability of diagnostic tests, or because of overlapping presentations between conditions (Alexander et al., 2014). However, the work presented in this thesis demonstrates the value of careful phenotyping in order to probe the complex genotype phenotype relationships seen in the neurodegenerative diseases. Without high quality phenotypic data, I would not have had the power to detect association at MSH3 in this thesis since this increased the power of the genetic analysis. This is illustrated by the
fact that, due to better phenotypic data, the association in REGISTRY ($p = 1.39 \times 10^{-5}$) was much lower than in TRACK-HD ($p = 5.8 \times 10^{-8}$) despite a greater sample size (n=1773 vs 216 respectively, Figure 3.1). For the polyglutamine disease analysis, the age of onset data was essential to our study yet the phenotypic data available for the spinocerebellar ataxia cases was generally fairly limited. The age at onset data was often recorded retrospectively from the notes, and I suspect that if more accurate data had been available we may have had a stronger signal.

Although their clinical presentations vary, there are many features common to the neurodegenerative disorders including the accumulation of misfolded proteins or peptide fragments in the brain and spinal cord. Immune pathways have been implicated in both AD and Parkinson’s disease GWAS analysis, and as I showed in Chapter 7 there is an overlap in immune pathway expression upregulation in AD and HD. Indeed, a maladaptive innate immune response has emerged as a critical driving force in the pathogenesis of many neurodegenerative diseases (Figure 8.2). Other neuronal pathways that are altered in various neurodegenerative diseases include protein folding and quality control, autophagy and lysosomal dysfunction, mitochondrial damage and homeostasis, protein seeding and propagation, stress granules, synaptic toxicity, nucleocytoplasmic transport and unconventional translation (Gan et al., 2018).
**Figure 8.2:** Innate immune pathways in neurodegenerative diseases. A maladaptive innate immune response has emerged as a critical driving force in the pathogenesis of many neurodegenerative diseases. SNPs on many disease-associated genes induce maladaptive innate immune responses that are also associated with aging and epigenetic changes. Microglia, the resident immune cells in the brain, engage in cross-talk with astroglia and are modulated by peripheral immune system. Maladaptive microglia could damage neuronal circuits due to dysfunction in their detection or response to homeostasis imbalance, resulting in accumulation of protein aggregates, in concert with astroglia and possibly the peripheral immune system. Microglia could also cause neuronal and network dysfunction by altering cytokine signaling and synaptic pruning, independently of their effects on protein aggregates. Figure from (Gan et al., 2018) image reproduced with permission of the rights holder, Nature Publishing Group.

As considered in this thesis, particularly in my work on C9orf72 repeat expansion associated disease (Chapter 6), and the phenotypic analysis in Huntington’s disease (Chapters 2 and 3), neurodegenerative diseases can have diverse clinical manifestations. The clinical manifestation of a particular neurodegenerative disease reflects the region of the brain and the specific population of cells and synapses within it that are affected (Gan et al., 2018, Fu et al., 2018). However the variable penetrance and broad range of presentations from ALS to FTD to HD phenocopy to Parkinsonism seen in people with expanded C9orf72 repeats is a particular conundrum. The factors underlying selective neuronal vulnerability have been difficult to dissect, but expression levels of risk proteins, lysosomal and ubiquitin proteasome system function, calcium and energy homeostasis, neurotransmitters and neurotransmitter receptors, and aging have all been proposed as having a role (Fu et al., 2018).

In Huntington’s disease, high levels of somatic instability are observed in the striatum, the tissue particularly vulnerable in this condition. By contrast low levels of somatic instability are seen in the cerebellum, and the difference in levels of somatic instability have been linked to the expression of DNA repair proteins (Goula et al., 2009). It is hard to see how these data are compatible with a similar mechanism of DNA repair protein mediated somatic expansion of the CAG repeat operating in the polyglutamine SCAs, since they are primarily disorders of the cerebellum: further in vitro analysis would be illuminating.

As I have discussed, I have made important advances in the understanding of what genetic factors underpin phenotypic diversity in HD and other repeat disorders. Arguably the most
exciting of these is the identification of a variant in MSH3 associated with HD progression. Given that MSH3 is not constrained by selection pressures and variants within it are not closely associated with malignancy it is an attractive therapeutic target, not only in HD but in other repeat disorders. A small molecule inhibitor of MSH3 may have the potential to slow the progression of HD in patients, and work is now underway by several pharmaceutical companies studying MSH3 as a therapeutic target for HD and potentially other triplet repeat diseases. While there are many hurdles to be crossed to better understand the mechanisms through which this variant acts and assessing efficacy, further work on this potential therapeutic avenue has to be the most important future direction arising from this thesis.
References


BETTENCOURT, C., RAPOSO, M., KAZACHKOVA, N., CYMBRON, T., SANTOS, C., KAY, T., VASCONCELOS, J., MACIEL, P., DONIS, K. C., SARAIVA-PEREIRA, M. L., JARDIM,


LAHIRI, N. 2013. Identification of markers of disease onset and progression in Huntington's Disease. MD(RES), University College London.


LANGBEHN, D. 2012. RE: Principal Component Analysis, in Outlier Analysis Summary. Type to HENSMAN MOSS, D. J.


M., JONES, R., ASHIZAWA, T., FRANK, S., SAINT-HILAIRE, M. H., HERSCH, S. M.,
ROSAS, H. D., LUCENTE, D., HARRISON, M. B., ZANKO, A., ABRAMSON, R. K.,
Common SNP-based haplotype analysis of the 4p16.3 Huntington disease gene

LEEK, J. T. 2014. svaseq: removing batch effects and other unwanted noise from

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T.,
O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN,
T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F.,
ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N.,
DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J.,
GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L.,
NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-
RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS,
C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H. H., YU,
D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY,
S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN,
C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I.,
MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M.,
SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J.,
TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G.

LESAGE, S., LE BER, I., CONDROYER, C., BROUSSOLLE, E., GABELLE, A., THOBOIS, S.,
PASQUIER, F., Mondon, K., DION, P. A., ROCHEFORT, D., ROULEAU, G. A.,
DURR, A. & BRICE, A. 2013. C9orf72 repeat expansions are a rare genetic cause

product of C9orf72, a gene strongly implicated in neurodegeneration, is

LGC. 2018. *How does KASP work* [Online]. Available:

Genomics Hum Genet*, 10, 387-406.

PRITCHARD, J. K. 2016. RNA splicing is a primary link between genetic variation

LING, S. C., POLYMENIDOU, M. & CLEVELAND, D. W. 2013. Converging mechanisms in

LIU, T., GHOSAL, G., YUAN, J., CHEN, J. & HUANG, J. 2010. FAN1 acts with FANC-


LOCKE, D. P., HILLIER, L. W., WARREN, W. C., WORLEY, K. C., NAZARETH, L. V., MIZNY,
D. M., YANG, S. P., WANG, Z., CHINWALLA, A. T., MINX, P., MITREVA, M.,
COOK, L., DELEHAUNTY, K. D., FRONICK, C., SCHMIDT, H., FULTON, L. A.,
FULTON, R. S., NELSON, J. O., MAGRINI, V., POHL, C., GRAVES, T. A.,
MARKOVIC, C., CREE, A., DINH, H. H., HUME, J., KOVAR, C. L., FOWLER, G. R.,

282


MASTROKOLIAS, A., DEN DUNNEN, J. T., VAN OMMEN, G. B., T HOEN, P. A. & VAN ROON-MOM, W. M. 2012. Increased sensitivity of next generation sequencing-
based expression profiling after globin reduction in human blood RNA. BMC Genomics, 13, 28.


MICHEGAN, U. O. Minimac Tutorial.


Pharmacological targeting of CSF1R inhibits microglial proliferation and prevents the progression of Alzheimer's-like pathology. *Brain*, 139, 891-907.


a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 81, 559-75.


neuropathology of the polyglutamine spinocerebellar ataxias type 1, 2, 3, 6 and 7. *Prog Neurobiol*, 104, 38-66.


instability in maternally transmitted infantile spinocerebellar ataxia type 7. 
*JAMA Neurol*, 72, 219-23.


TRAPNEL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., 
transcript expression analysis of RNA-seq experiments with TopHat and 

TRINH, J., GUSTAVSSON, E. K., VILARIÑO-GÜELL, C., BORTNICK, S., LATOURELLE, J., 
MCKENZIE, M. B., TU, C. S., NOSOVA, E., KHINDA, J., MILNERWOOD, A., 
LESAGE, S., BRICE, A., TAZIR, M., AASLY, J. O., PARKKINEN, L., HAYTURAL, H., 
FOROUD, T., MYERS, R. H., SASSI, S. B., HENTATI, E., NABLI, F., FARHAT, E., 
AMOURI, R., HENTATI, F. & FARRER, M. J. 2016. DNM3 and genetic modifiers of 
age of onset in LRRK2 Gly2019Ser parkinsonism: a genome-wide linkage and 

TROTTIER, Y., DEVYS, D., IMBERT, G., SAUDOU, F., AN, I., LUTZ, Y., WEBER, C., AGID, Y., 
disease protein and discrimination of the normal and mutated form. *Nature 
Genetics*, 10, 104-110.

TSUJI, S. 1999. Dentatorubral-pallidoluysian atrophy (DRPLA): clinical features and 

TURNER, C., COOPER, J. M. & SCHAPIRA, A. H. V. 2007. Clinical correlates of 
mitochondrial function in Huntington's disease muscle. *Movement Disorders*, 
22, 1715-1721.

UDD, B., JUVONEN, V., HAKAMIES, L., NIEMINEN, A., WALLGREN-PETTERSSON, C., 
disease in Western Finland -- is the syndrome underdiagnosed? *Acta Neurol 
Scand*, 98, 128-33.


UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B., REMM, M. 
& ROZEN, S. Primer3--new capabilities and interfaces. *Nucleic Acids Res.*, 40, 
e115.

VAN DE WARRENBURG, B. P., HENDRIKS, H., DURR, A., VAN ZUIJLEN, M. C., STEVANIN, 
G., CAMUZAT, A., SINKE, R. J., BRICE, A. & KREMER, B. P. 2005. Age at onset 
variance analysis in spinocerebellar ataxias: a study in a Dutch-French cohort. 
*Ann Neurol*, 57, 505-12.

VAN DE WARRENBURG, B. P., SINKE, R. J., VERSCHUUREN-BEMELMANS, C. C., 
SCHEVERF, H., BRUNT, E. R., IPPEL, P. F., MAAT-KIEVIT, J. A., DOOIJES, D., 
Spinocerebellar ataxias in the Netherlands: prevalence and age at onset 

VAN DEN BROEK, W. J. A. A., NELEN, M. R., WANSINK, D. G., COERWINKE, M. M., 
expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in 
mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. 

VAN DER BURG, J. M., BJORKQVIST, M. & BRUNDIN, P. 2009. Beyond the brain: 


WILLIAMS, G. M. & SURTEES, J. A. 2015a. MSH3 Promotes Dynamic Behavior of Trinucleotide Repeat Tracts In Vivo. *Genetics*, 200, 737-+


ZHAO, X. N. & USDIN, K. 2018. FAN1 protects against repeat expansions in a Fragile X mouse model. *DNA Repair* 69, 1-5.
Appendix 1:

General PCR and Sequencing protocol

Stage 1 - PCR:
1) Prepare premix: 25µl MegaMix Blue per well; add stock primer so that final concentration of primer is 0.5µM
2) Pipette 24µl of premix per well of a 96-well plate
3) Add 1µl DNA to each well.
4) Cover the plate and spin down
5) Transfer plate to Tetrad2 thermal cycler and run PCR program with the following cycling conditions:
   a) 95ºC for 1 min
   b) 95ºC for 30 secs
   c) 58ºC for 30 secs
   d) 72ºC for 1 min
   e) Go to step b) for an additional 34 cycles
6) Assess PCR by electrophoresis of 5µl product on a 2% agarose gel stained with Red Safe (20µl Redsafe to 400ml gel). Load 5µl HyperLadder IV size standard. View gel using the Biorad transilluminator and Quantity One software.

Stage 2 - PCR Product Clean-up
1) Add an equal volume of Microclean to the PCR product, cover and mix well by vortexing.
2) Leave at room temperature for 5 minutes.
3) Spin the plate at 3000G for 40 minutes at RT.
4) Invert plate onto tissue paper and spin at 40G for 30 seconds.
5) Resuspend pellets by adding 200µl 18MΩ H2O to amplicons giving a strong signal on gel. (Resuspend in less H2O if PCRs are weaker.) Vortex plate. Leave for 5 mins.
6) Vortex again and spin down. PCR products are now ready to use.

Stage 3 - Sequencing Reactions
1) For each sequencing reaction prepare a premix of 1µl BigDye, 5µl BetterBuffer and 7.25µl 18MΩ ddH2O.
2) Mix and pipette 13.25µl of premix into each well of a 96 well plate.
3) Pipette 0.75µl sequencing primer (at 5µM concentration) into their respective wells.
4) Pipette 1µl of PCR product into well.
5) Cover plate, spin down and run program ‘BD2’ (or BD23 or BD22; the last digit refers to the extension time) in the Diagnostics folder on a Tetrad2 thermal cycler in room 4.09b. Cycling is as follows:
   a) 96ºC for 1 min
   b) 96ºC for 10 secs
   c) 50ºC for 5 secs
   d) 60ºC for 3 mins
   e) Go to step b) for an additional 24 cycles

Stage 4 - Sequencing Product Clean-up
1) To each sequencing reaction add 3.75µl 125mM EDTA ensuring the solution is pipetted into the bottom of the well.
2) Add 45µl 100% EtOH to each reaction and mix by pipetting up and down.
3) Leave plate at room temperature for 15 minutes
4) Spin plate at 3000G for 30 minutes at 4°C
5) Remove cover, invert plate onto tissue paper and spin up to 185G.
6) Add 60µl 70% EtOH.
7) Spin plate at 1650G for 15 minutes at 4°C.
8) Remove cover, invert plate onto tissue paper and spin at 185g for 1 minute.
9) Place plate on PCR block, uncovered, held at 37°C for 5 minutes to remove final traces of EtOH.

Stage 5 - Electrophoresis on the 3730XL DNA Analyzer

1) Add 10µl Hi-Di formamide loading solution to each well of the plate containing dry sequencing product pellets. (Caution!! Wear appropriate protective eyewear, clothing, and gloves).
2) Cover plate and spin down. Denature samples by placing on PCR block held at 95°C for 2 minutes, and transfer to ice/4°C tetrad block immediately for 2 mins.
3) Set-up a new sample sheet by opening the Plate Manager and clicking on ‘Find All’. Use the next run number in the sequence, ‘3730Runx’, where x is the run number. Click on ‘New’. Fill in the plate name with the run number in the format as described and choose Sequencing Analysis for the application. Write the run number on the ‘sequencing reaction set-up and sample sheet’. Fill in the Sample Name column using the following format: [DGnumber]_[[primer number/name]. i.e. 12345_122. This is important for analysis using ‘Seqscape’. For the Results Group column, choose ‘Sequencing_HumanGenetics’.
4) For the Instrument Protocol choose ‘BD1_LongSeq50_POP7’. (This is a 15 second injection run for BigDye Version 1.1 using 50cm capillaries and POP-7 polymer). For the analysis protocol choose ‘BD1’
5) Remove the plate from ice/Tetrad and place in a black plate base and remove cover; make sure the plate is orientated correctly. Seal the plate by placing a clean septa on top of the plate, aligning with the wells and pressing firmly down. Snap a white plate retainer on top of this.
6) Place the assembled cassette into the input stack of the sequencer.
7) Link the plate by selecting the plate in ‘Run Scheduler’ which will add it to the input stack. Close the instrument doors and wait for the light on the front of the machine to turn green. The run can now be started by clicking on the green arrow at the top left of the screen. The run takes 2 hours per plate.
8) Data is automatically saved to the storage drive: ‘Z:\3730XL_Runs\Human Genetics Group’
9) Once the run is complete, view the raw data by opening up the run in ‘Run History’: signal strength and general run quality can be quickly assessed.
10) Both raw and analysed data can also be viewed using the Sequence Analysis Software V5.2 which is installed on the instrument computer. This is done as follows: double click on program icon to open, fill in password, ‘sequence’, click on ‘add samples’. Select the run folder, click on ‘add selected samples’ and press ‘OK’. Data for each sample can now be viewed as raw data, base-called electropherogram, sequence data and annotation data (run parameters, signal strengths etc.).
Appendix 2:

Published papers and book chapters

Published papers to which I have contributed during the course of my PhD, including both those which have been discussed in this thesis and those beyond the scope of this thesis.


