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.

1. Chapter 3. **Hensman Moss DJ***, Pardiñas AF*, *et al*. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. ***The Lancet Neurology***. 2017. 16(9) 701-711. **Co-first author.* **(43)**
2. Chapter 4. Bettencourt C*, **Hensman Moss** D*, Flower M*, Wiethoff S* *et al*. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. ***Ann Neurol***. 2016 Jun;79(6):983-90. doi: 10.1002/ana.24656. **Co-first author.* **(51)**
3. Chapter 6. **Hensman Moss DJ** *et al*. C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. ***Neurology.*** 2014 Jan 28;82(4):292-9. **(140)**
4. Chapter 6. Beck J*, Poulter M*, **Hensman D** *et al*. Large C9orf72 Hexanucleotide Repeat Expansions Are Seen in Multiple Neurodegenerative Syndromes and Are More Frequent Than Expected in the UK Population. ***Am J Hum Genet***. 2013 Mar 7;92(3):345-53. **(221)**
5. Chapter 7. **Hensman Moss DJ** *et al*. Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. ***Nature Scientific Reports***. 2017. 7, 44849. **(5)**

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.
2. Huntington Study Group 'Insight of the Year' award for most influential paper in Huntington's disease in 2017. HSG Meeting, Denver, USA, November 2017.
3. Neuromics Consortium Prize for best poster presentation at the Neuromics Meeting, Berlin, May 2017.
4. European Huntington's Disease Network Prize for best presentation at the EHDN Meeting, The Hague, September 2016.

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 Scahill, 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

List of Figures

List of Tables

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.

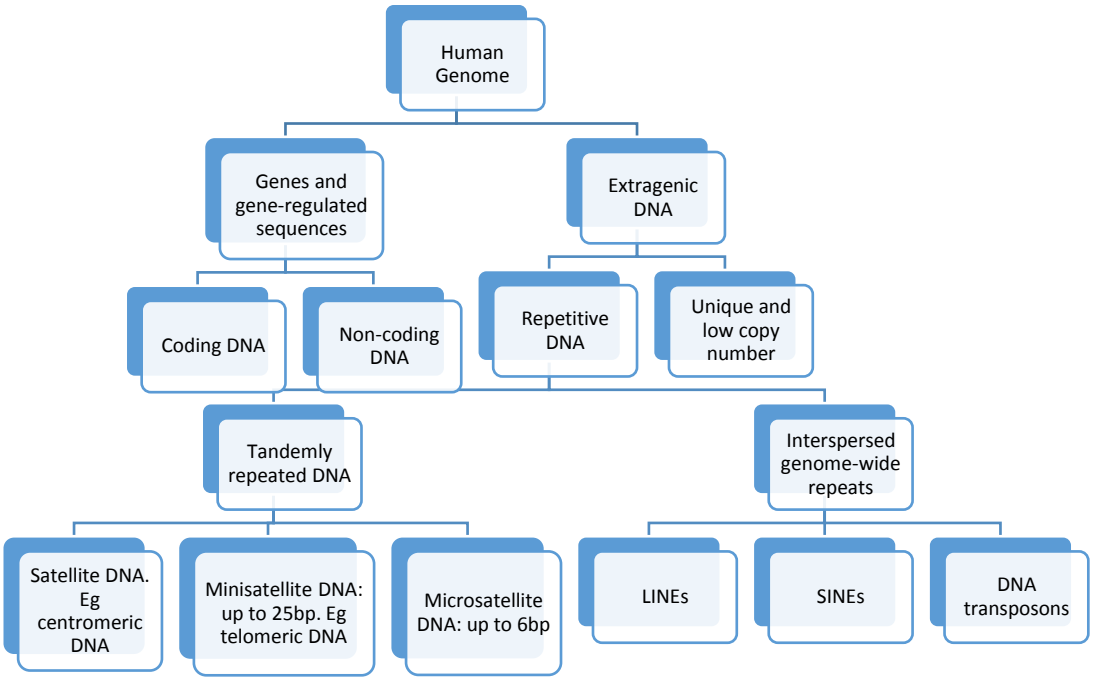


Figure 1.1*: Types of DNA within the human genome. Types of DNA within the human genome. Tandem repeats tend to be located in blocks at one or more locations on chromosomes. Interspersed repetitive sequences may be widespread over the genome, located over broad regions of one of more chromosomes. Bp: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

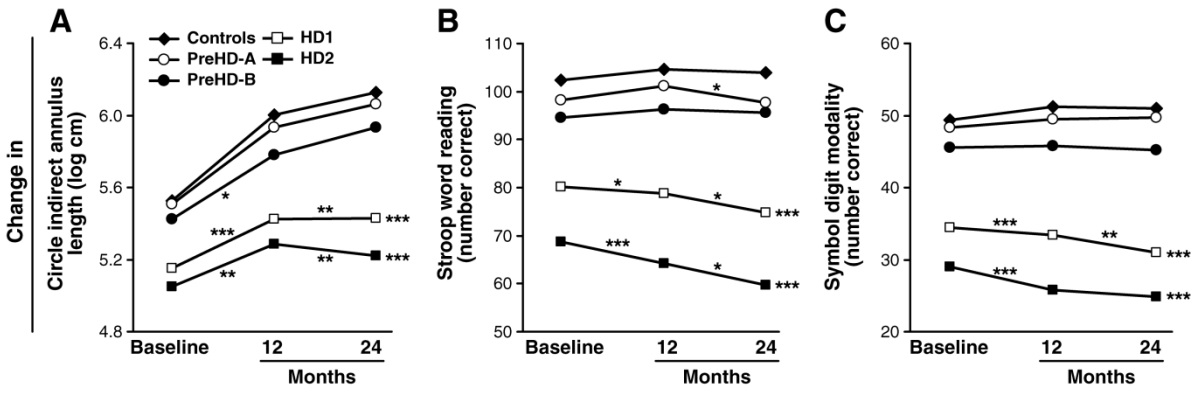


Figure 1.2*: Longitudinal changes in cognitive measures from the Track-HD study over 24 months. Significant change differences relative to controls over 0-12, 12-24, and 0-24 months are represented by *p<0.05, **p<0.01 and ***P<0.001. Groups determined at start of study; PreHD-A: more than 10.8 years 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).

CAG repeat length	<27	27 – 35	36 – 39	≥40	≥55 – 60
Clinical manifestation	Normal	Intermediate repeat allele Not generally pathogenic May expand into disease range in future generations in paternal line	Reduced penetrance but pathogenic	Fully penetrant	Usually have juvenile onset

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.

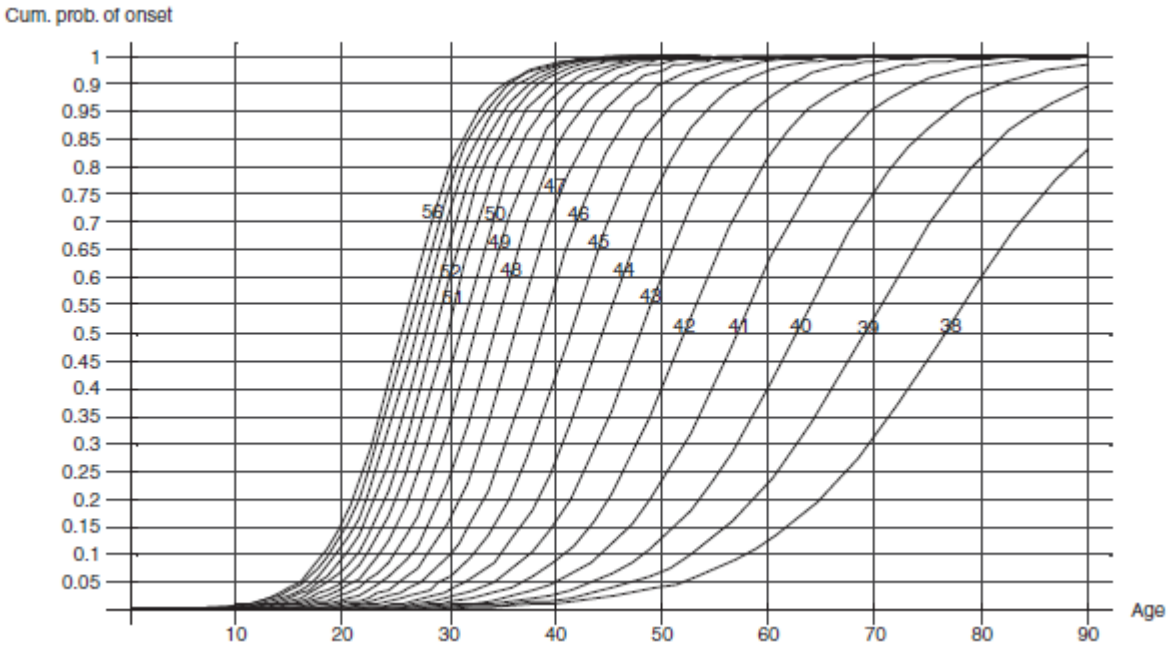


Figure 1.3: *Cumulative probability of Huntington's disease onset curves for various CAG lengths. Numbers indicate CAG repeat length. Cum. prob. onset = Cumulative probability of onset of Huntington's disease. From (Langbehn et al., 2004), Image reproduced with permission of the rights holder, Wiley-Blackwell.*

.

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).

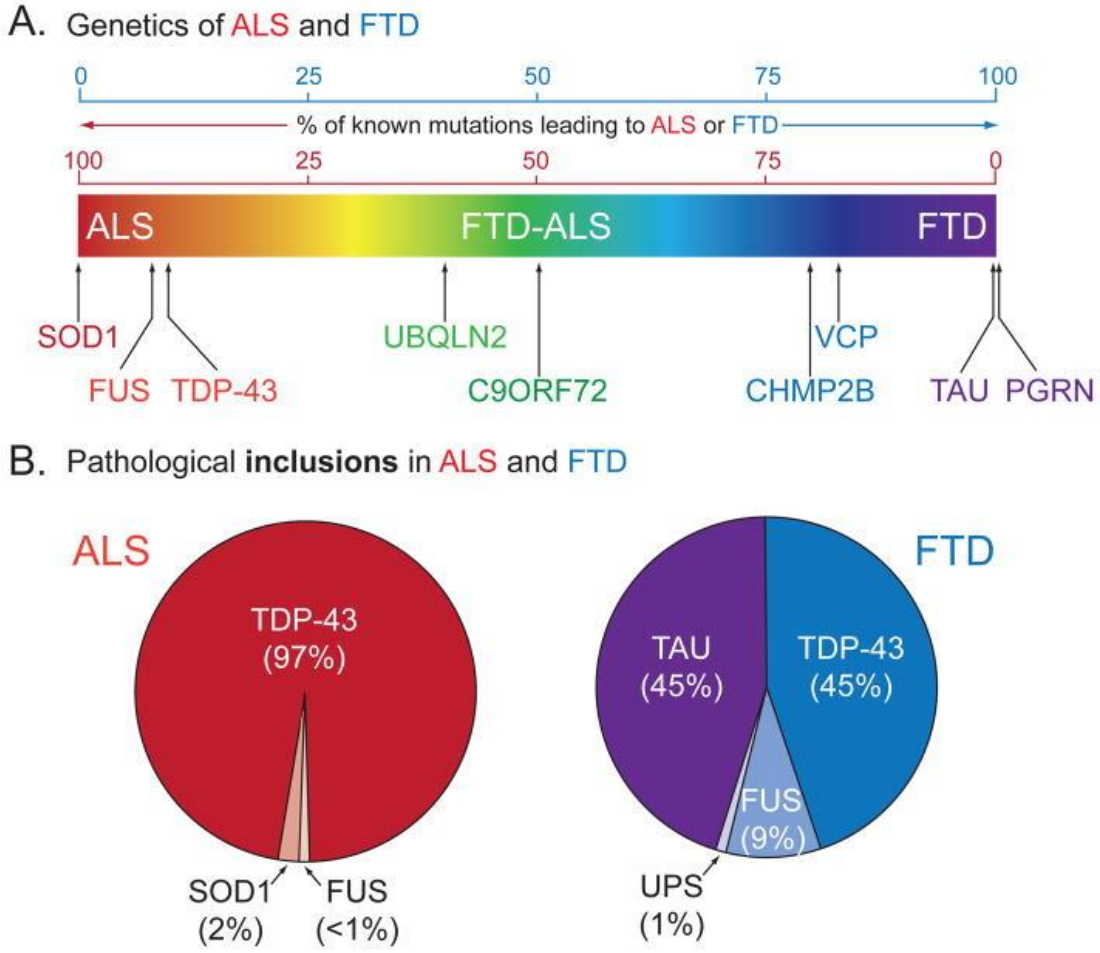


***Figure 1.4**: Clinical, genetic and pathological overlap of ALS and FTD.* (A) ALS and FTD represent a continuum of a broad neurodegenerative disorder with each presenting as extremes of a spectrum of overlapping clinical symptoms (ALS in red and FTD in purple). Major known genetic causes for ALS and 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), **i**mage 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 Baysean 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.

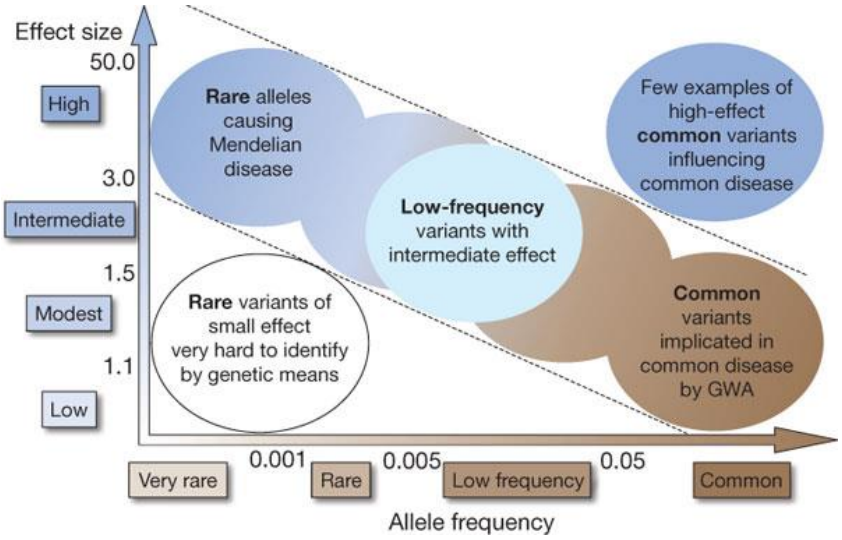


***Figure 1.5**: Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).* From (Manolio et al., 2009)**,** *image reproduced with permission of the rights holder, Nature Publishing Group.*

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).

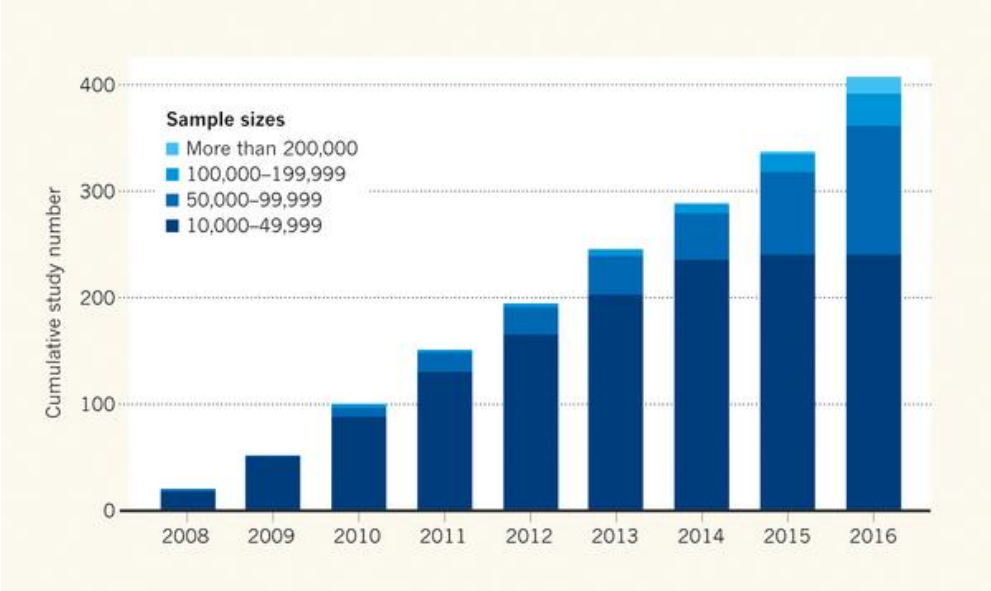


Figure 1.6*: Ever-increasing sample sizes for genome-wide association studies (GWAS).* This graph shows the cumulative number of GWAS involving 10,000 samples or more published per year, with those involving different sample sizes indicated in different colours. Graph from

(Manolio, 2017); Data taken from www.ebi.ac.uk/gwas., image reproduced with permission of the rights holder, Nature Publishing Group.

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* promotor 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 PPARGC1A (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). For

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 NHNN/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.

	Stage	TFC
Early HD	1	11 - 13
Early HD	2	7 – 10
Moderate HD	3	3 – 6
Advanced HD	4	1 – 2
Advanced HD	5	0

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.

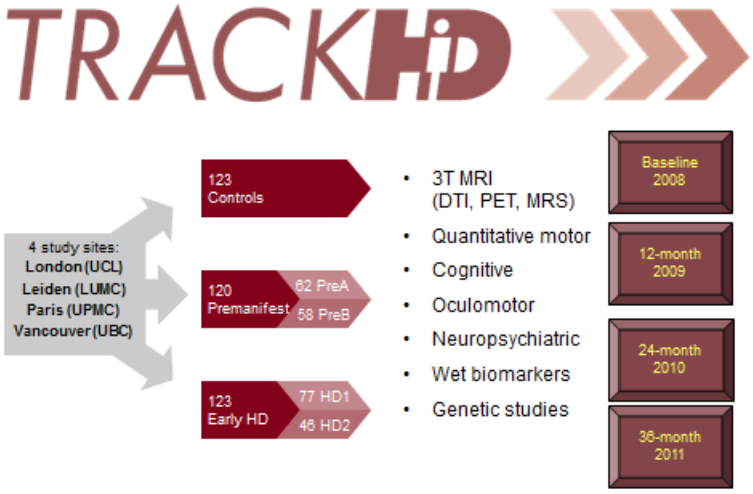


***Figure 2.1**: Study outline of TRACK-HD. Study sites, numbers of subjects in each disease group at baseline, principle assessment modalities and years assessed are shown.*

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**.

Time	Assessment
09:00 (45 min)	Consent, Clinical Rating, Neuropsychiatric assessment
9:45 (120-150 min)	Imaging
12:30	Lunch
13:30 (60 min)	Cognitive

14:30 (70 min)	Quantitative Motor
15:40 (15 min)	Oculomotor
15:55 (10 min)	Neuropsychiatric & Functional assessments
16.10 (75 min)	Transcranial Magnetic Stimulation

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**.

TRACK-HD variable	Domain
Symbol digit modality test (number correct)	Cognitive

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

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.

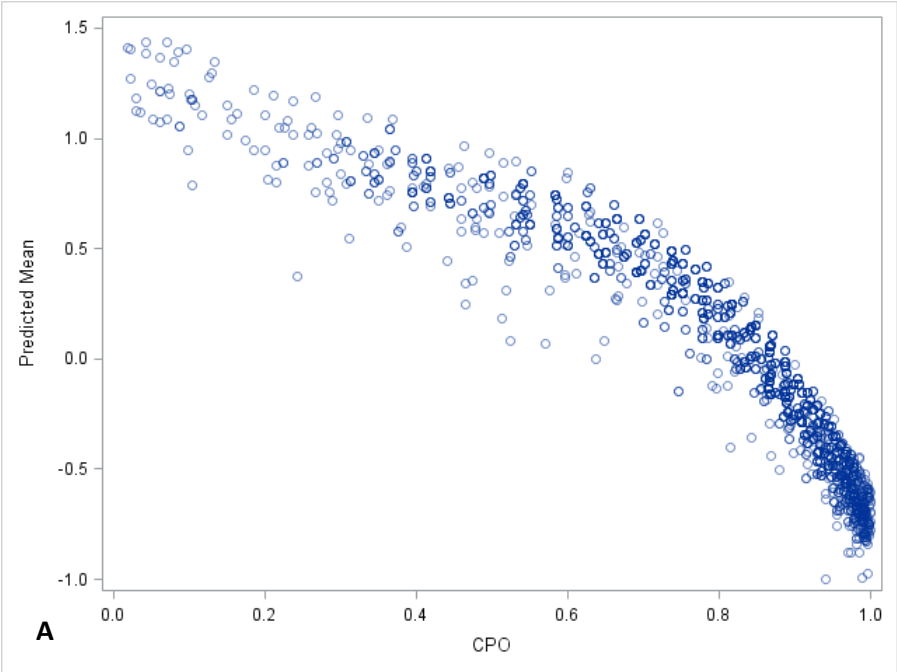
Atypical severity was defined as the residual between each subject's observed and marginal predicted value. The final averaged multiple imputation model used a 2 degree of freedom restricted cubic spline (Harrell, 2001) of cumulative probability of onset (CPO), plus main effects of gender and CAG length 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 10^{th}, 50^{th}, and 90^{th} 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**.

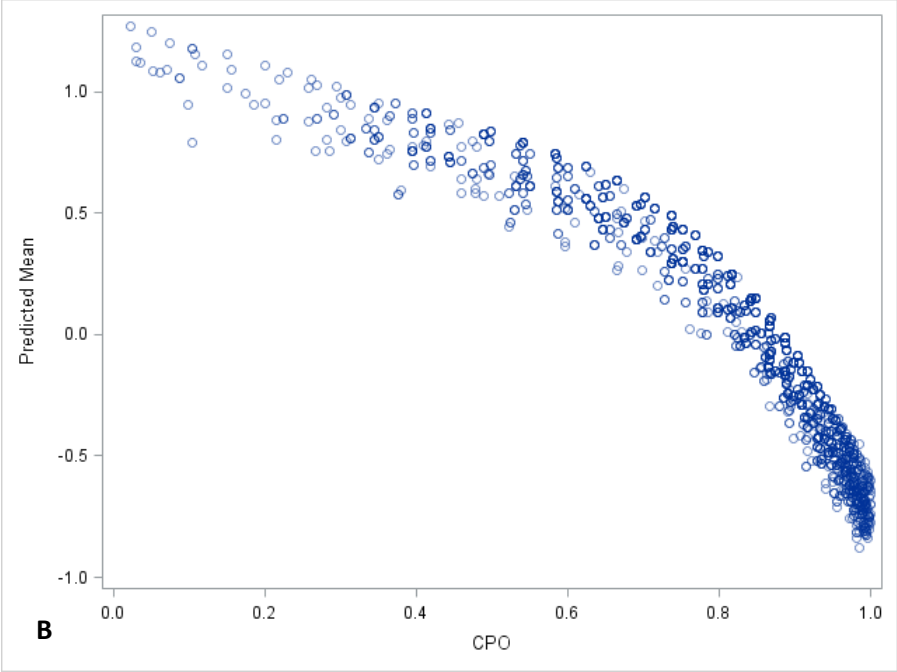
Parameter	gender	Estimate	Std Error	95% Confidence Limits		DF	t for H0:	P Val
Intercept		2.075589	0.267283	1.55102	2.60016	897.01	7.77	<.0001
cpo_1		-0.9142	0.21009	-1.32638	-0.50201	1191.6	-4.35	<.0001
cpo_2		-7.00283	0.911001	-8.79025	-5.2154	1141.5	-7.69	<.0001
cag		-0.01919	0.005133	-0.02927	-0.00912	862.96	-3.74	0.0002
gender	F	-0.13631	0.042605	-0.21992	-0.05271	1030.1	-3.2	0.0014
gender	M	0	0	.	.	.	.	.

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)**.


Predicted Mean
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
0.6
0.8
1.0
CPO
A


Predicted Mean
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
0.6
0.8
1.0
CPO
B

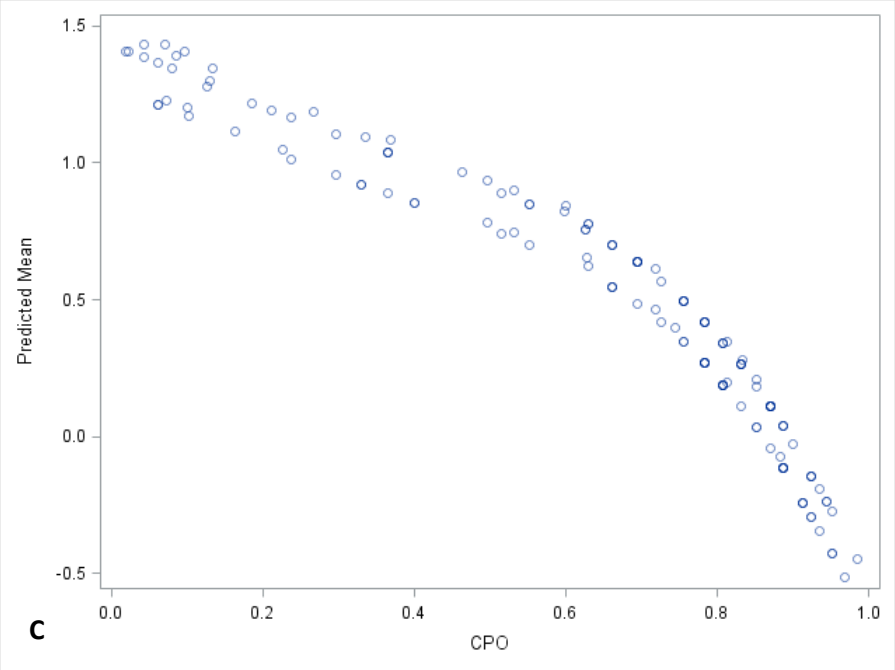


***Figure 2.2**: Age-CAG severity function against clinical probability of onset (CPO) in REGISTRY. A: plot showing predicted values for all subjects. B: plot of predicted values using only subjects in the CAG 41– 55 range. C: Plot based on extrapolating the severity model to subjects with CAG in the 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.

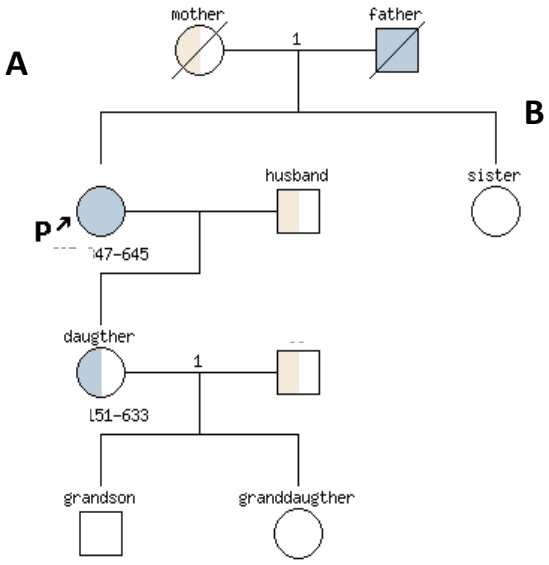
2.6 Assessment of Relatedness

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**).

Column number	Column entry
1	Individual's family ID ('FID')
2	Individual's within-family ID ('IID'; cannot be '0')
3	Within-family ID of father ('0' if father isn't in dataset)
4	Within-family ID of mother ('0' if mother isn't in dataset)
5	Sex code ('1' = male, '2' = female, '0' = unknown)

***Table 2.5**: Format for family history encoding*

A simple example is shown in **Figure 2.3** below.



1	2	3	4	5
1	1	0	0	1
1	2	0	0	2
1	3	1	2	2
1	4	0	0	1
1	5	4	3	2

***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 *Eco*R1. 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 oligonucelotides: 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 5×10^{-8}. Other thresholds have been suggested however: by subsampling genotypes at increasing density and extrapolating to infinite density, a genome-wide significance threshold of 7.2×10^{-8} was obtained; sequence simulation under various demographic and evolutionary models found a genome-wide significance threshold of 3.1×10^{-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.

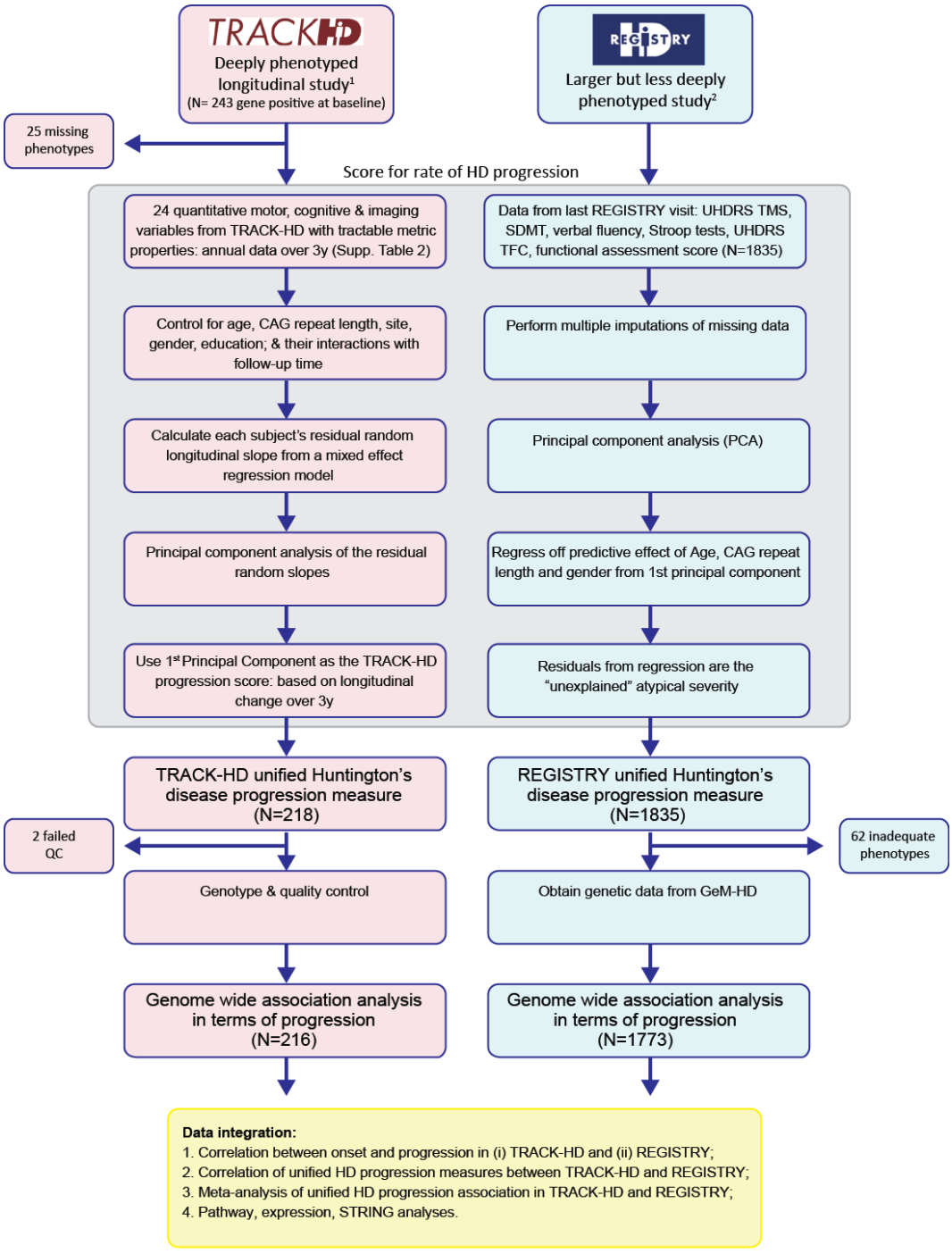


Figure 3.1*: Study Design. After establishing that brain imaging, quantitative motor and cognitive variables are correlated and follow a similar trajectory, we scored the TRACK-HD subjects using principal component 1 as a unified progression measure, and used this measure to look for genome-wide associations 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 principalcomponent 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**).

Factor Pattern	
	Factor1
Square root of raw UHDRS total motor score	-0.91567
Symbol digit modality test (number correct)	0.90797
Stroop word reading test (number correct)	0.87904
UHDRS Total functional capacity	0.86045

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 ($\hat{\pi}$) 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**).

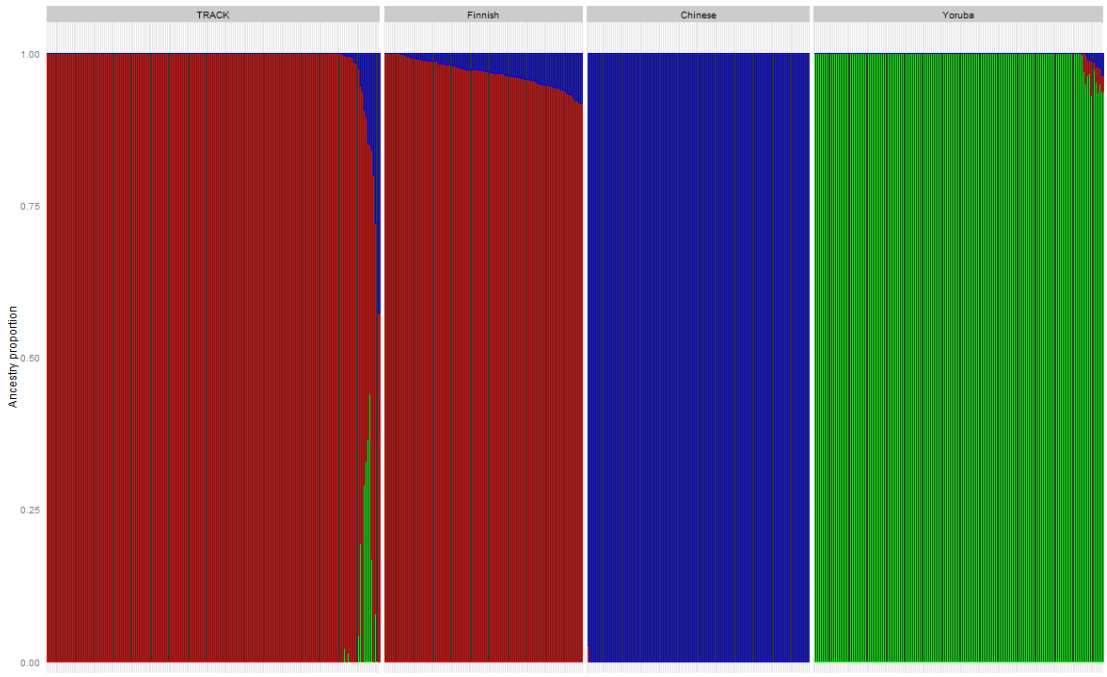


***Figure 3.2**: Ancestry analysis of the TRACK-HD cohort (left hand box) in comparison to Finnish, Chinese, and Yoruban populations from left to right respectively. 6 subjects had >25% non-European ancestry. Figure prepared by Dr Pardiñas.*

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.

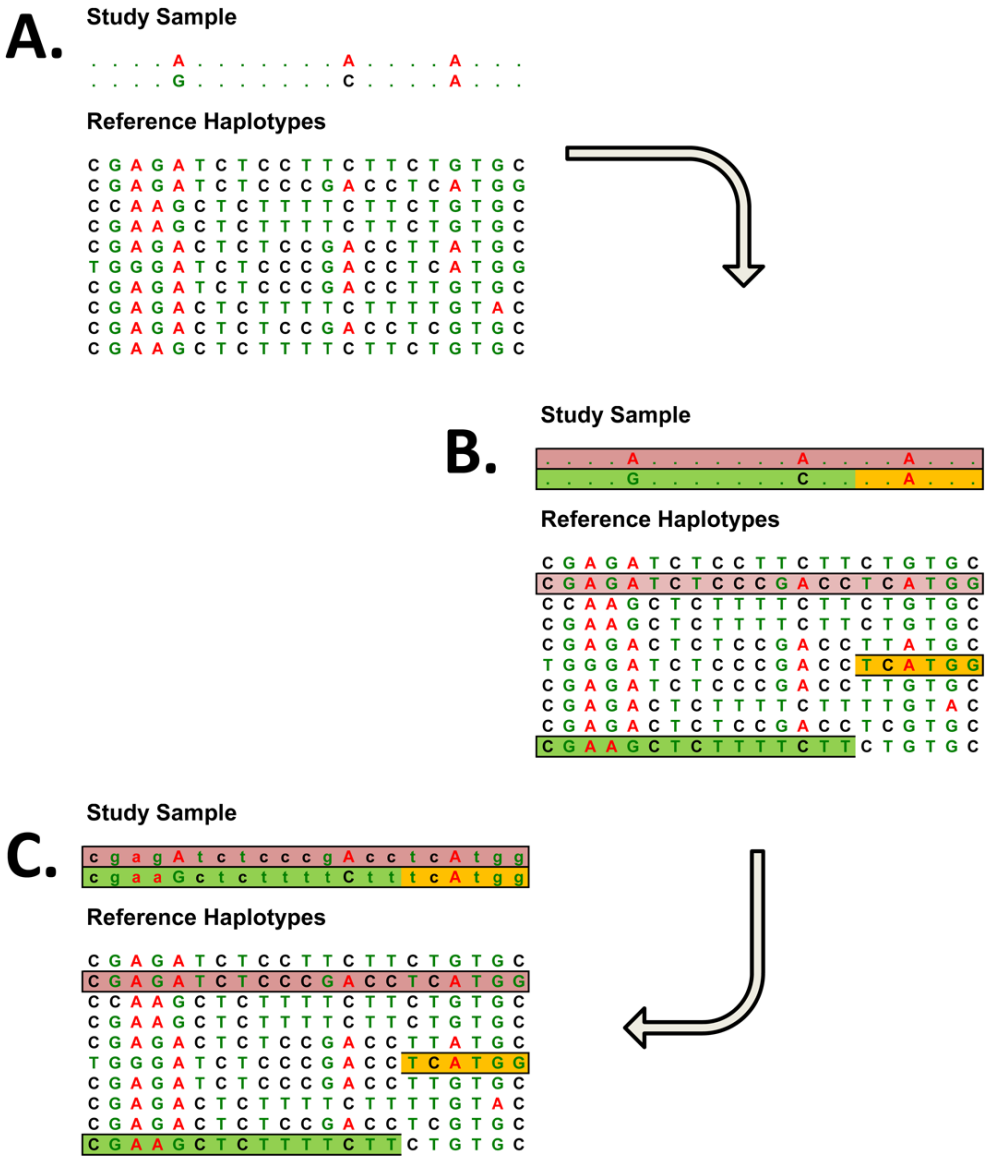


***Figure 3.3**: Genotype imputation in a sample of apparently unrelated individuals. **A:** the observed data which consists of genotypes at a modest number of genetic markers in each sample being studied and of detailed information on genotypes (or haplotypes) for a reference sample. **B:** the process 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.

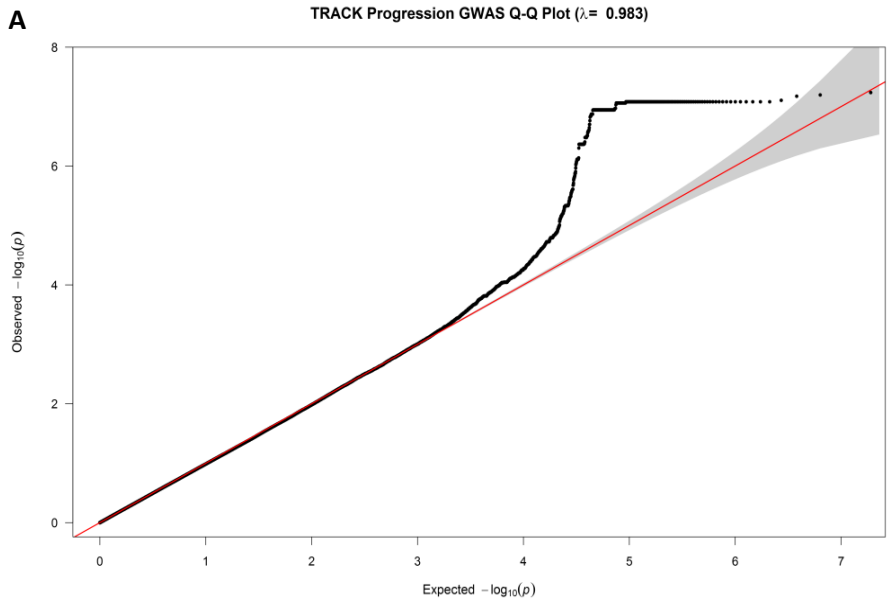
In order to transform the results into independent GWAS signals, PLINK was again used to perform linkage disequilibrium (LD) clumping ($r^2 = 0.1$, $p < 1x10^{-4}$; window size < 3 Mb). Due to the relatively small size of the TRACK-HD and REGISTRY samples, analyses were restricted to SNPs with minor 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.

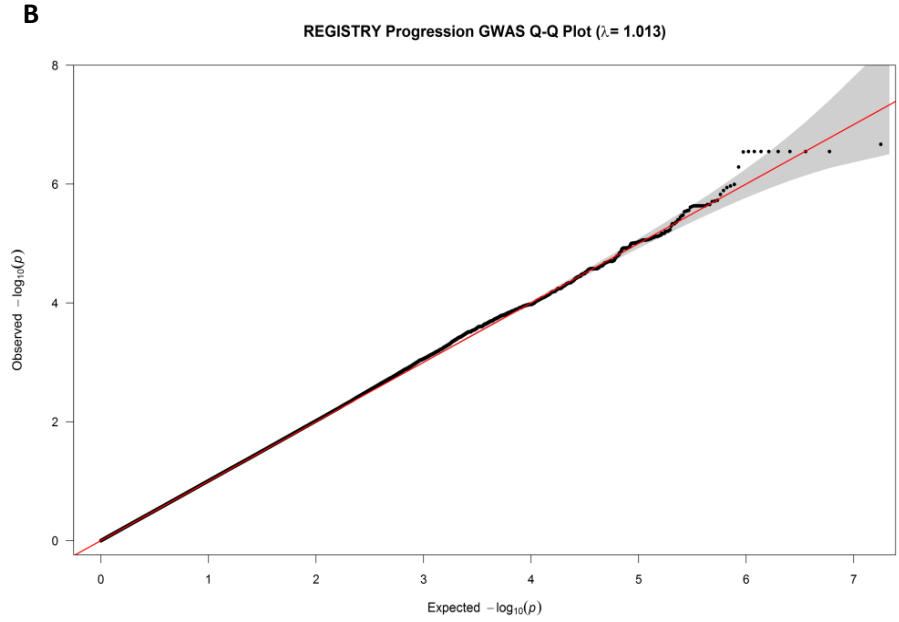
QQ plots of observed log p-values (sorted by value) for each SNP versus their expected values in the absence of association are shown for TRACK-HD, REGISTRY and the meta-analysis are shown in **Figure 3.4**. If there is no association, and no systematic inflation in the test statistics (for example, 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 λ (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 λ 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.

A


TRACK Progression GWAS Q-Q Plot (λ= 0.983)
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
1
2
3
4
5
6
7
8

B


REGISTRY Progression GWAS Q-Q Plot (λ= 1.013)
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
1
2
3
4
5
6
7
8

C

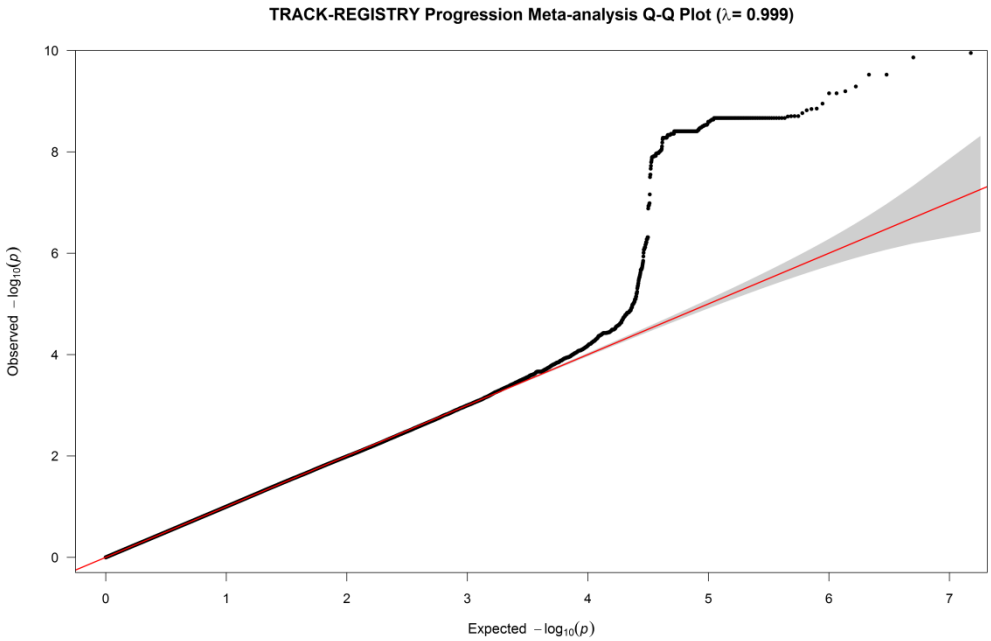


Figure 3.4*: QQ plots of the (A) TRACK-HD and (B) REGISTRY genome wide association studies. And (C) meta-analysis. λ close to 1 shows there is no systematic inflation of test statistics. (Figure prepared by Dr Pandiñas).*

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**)

Variable	Effect of one unit change in subject's progression score on this variable	Standard Error
Raw TMS rate	0.71	0.19
Adjusted TMS rate	0.57	0.18
Raw TFC rate	0.21	0.047
Adjusted TFC rate	0.20	0.044

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.

	brainf1	brainf2	brainf3	cogf1	cogf2	cogf3	cogf4	motf1	motf2	motf3	motf4
brainf1	1	0	0	-0.355	0.077	0.146	-0.068	0.43	0.096	-0.065	-0.139
p	0	1	1	<.0001	0.26	0.03	0.32	<.0001	0.16	0.34	0.04
brainf2	0	1	0	-0.097	-0.055	0.12	-0.016	0.005	-0.149	-0.043	0.041
p	1	0	1	0.15	0.42	0.08	0.81	0.94	0.03	0.53	0.55
brainf3	0	0	1	0.016	0.064	0.12	-0.009	0.15	0.05	-0.108	-0.161
p	1	1	0	0.81	0.35	0.08	0.89	0.03	0.46	0.11	0.02
cogf1				1	0	0	0	-0.434	-0.154	0.035	0.112
p				0	1	1	1	<.0001	0.02	0.6	0.09
cogf2				0	1	0	0	0.035	0.07	-0.12	-0.163
p				1	0	1	1	0.59	0.29	0.07	0.01
cogf3				0	0	1	0	0.105	-0.017	-0.092	-0.143
p				1	1	0	1	0.11	0.8	0.16	0.03
cogf4				0	0	0	1	-0.019	-0.05	-0.011	-0.054
p				1	1	1	0	0.77	0.44	0.87	0.42

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.

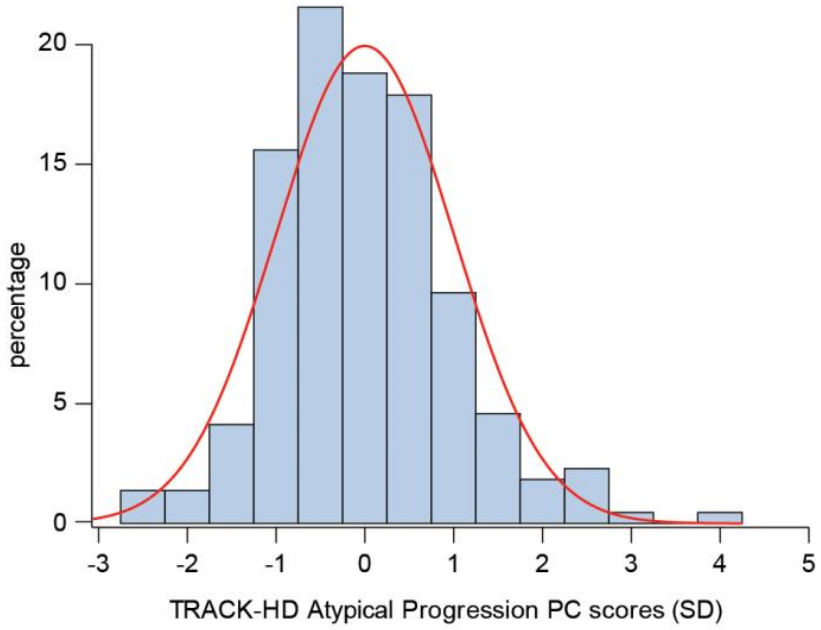


***Figure 3.5**: Distribution of progression measure in 218 members of TRACK-HD cohort. Curve is the normal distribution approximations of the severity score distributions. (Figure prepared by Prof Langbehn, edited by me, version of this figure used in Figure 2 in (Hensman Moss et al., 2017b)).*

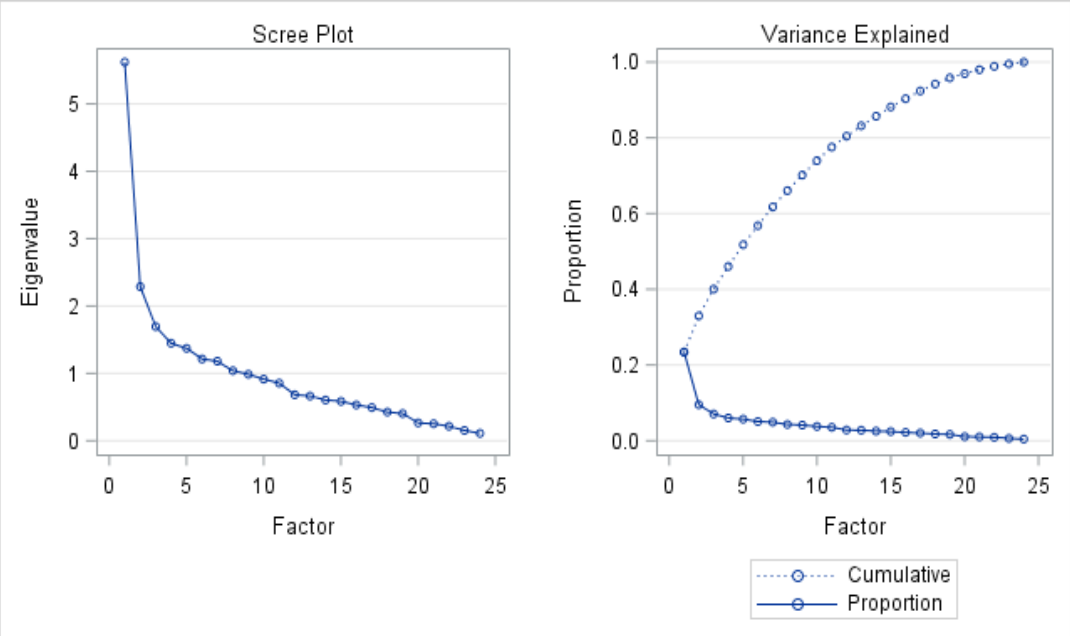


***Figure 3.6**: The first principal component accounts for a high proportion of the variance in the TRACK-HD progression analysis. (A) Scree Plot and (B) Plot showing proportion of variance explained in the TRACK-HD progression principal component analysis: the dominance of the first PC is illustrated. Figure prepared by Antonio Pardiñas.*

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

intertap time)								
Tongue force—heavy (log coefficient of variation)	0.147	0.016	-0.332	0.586	-0.445	0.177	-0.033	0.012
Tongue force—light (log coefficient of variation)	0.247	0.114	-0.399	0.451	-0.407	0.191	0.217	0.066
Grip force, dom. hand, heavy condition (log of mean orientation)	0.615	0.488	0.252	0.009	-0.078	-0.014	-0.336	-0.077
Grip force, dom. hand, heavy condition (log of mean position)	0.568	0.518	0.207	0.033	-0.027	-0.051	-0.381	-0.042
Grip force, nondom. hand, heavy condition (log of coefficient of variation)	0.516	0.400	0.213	0.108	0.003	0.122	0.231	-0.145
Grip force, dom. hand, light condition (log of coefficient of variation)	0.681	0.311	0.250	0.034	0.016	0.140	0.188	0.114
Grip force, nondom. hand, light condition (log of coefficient of variation)	0.647	0.430	0.293	0.071	-0.061	0.071	0.163	-0.055
Variance Explained (%)	23.4	9.5	7.1	6	5.7	5.1	4.9	4.3

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.*

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**.

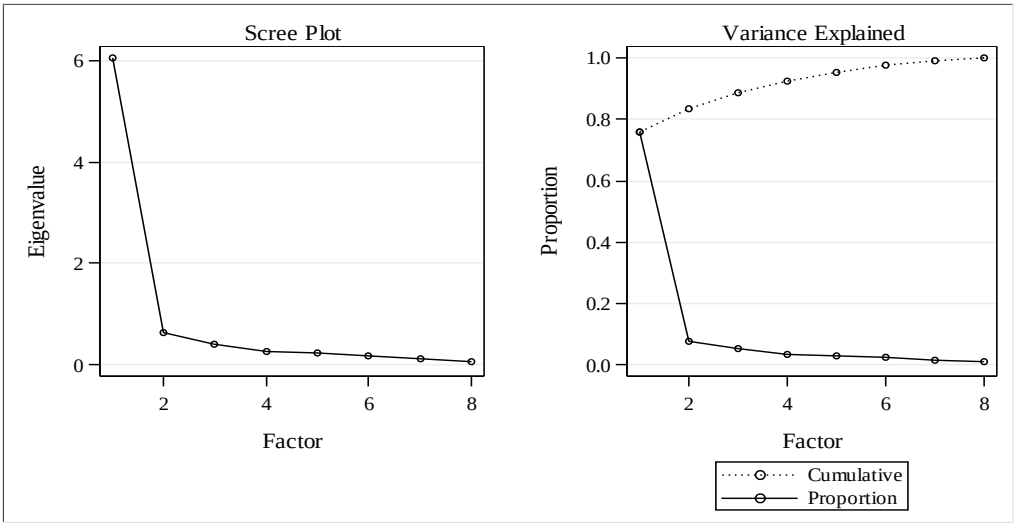


Figure 3.7*: The first principal component accounts for a high proportion of the variance in the REGISTRY progression analysis. (A) Scree Plot and (B) Plot showing proportion of variance explained in the REGISTRY progression principal component analysis: the dominance of the first PC is illustrated.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

Variable	Variable explanation	Factor1	Factor2
sqrtmotor	Square root of the UHDRS total motor score	-0.84233	0.30062
verfl	UHDRS verbal fluency	0.79108	0.24136
sdmt	UHDRS symbol digit score	0.89833	0.1522
scnt	UHDRS Stroop colour naming	0.89596	0.25872
swrt	UHDRS Stroop word reading	0.88978	0.2109
sit1	UHDRS Stroop interference score	0.87684	0.21789
tfc	UHDRS total functional capacity	0.8746	-0.39367
fasscore	UHDRS functional assessment scale	0.88355	-0.38555

***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.

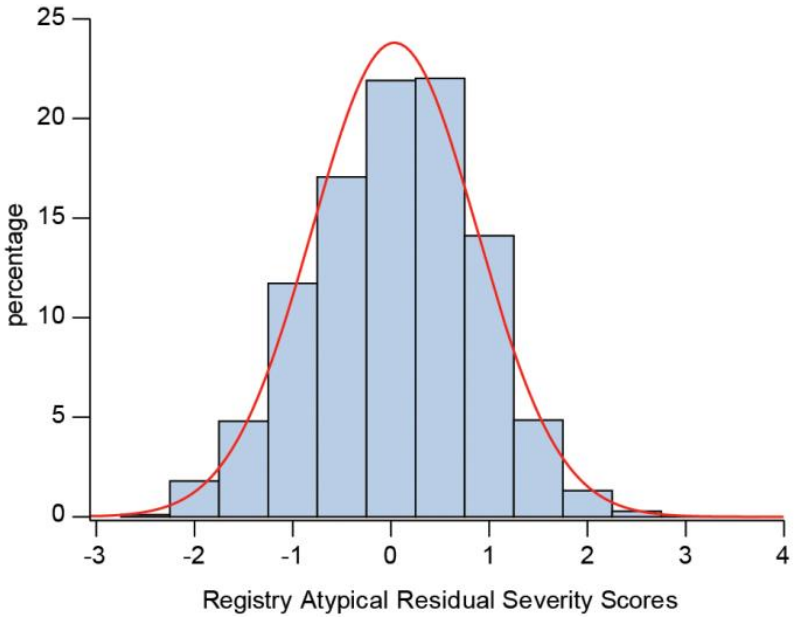


***Figure 3.8:** Distribution of atypical severity (compared to predicted severity at final visit) in 1835 members of the REGISTRY cohort. The curve is the normal distribution approximations of the severity score distributions. (Figure prepared by Prof Langbehn, edited by me, version of this figure used in Figure 2 in (Hensman Moss et al., 2017b)).*

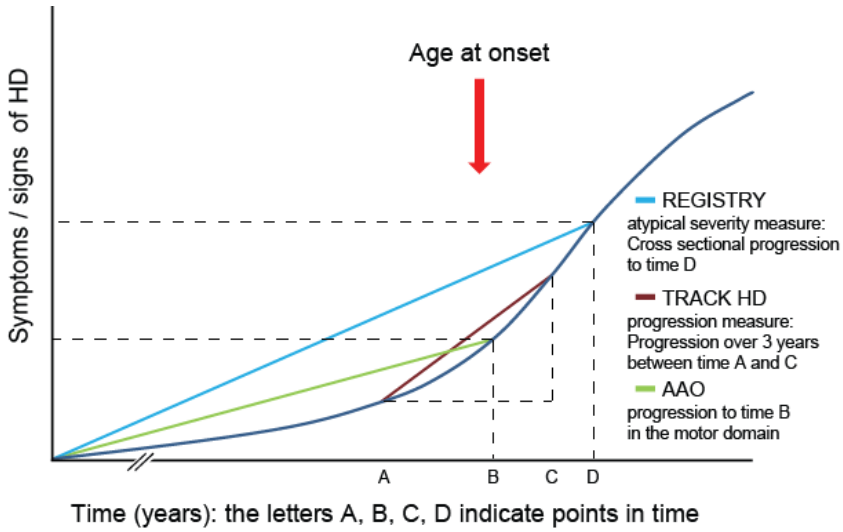


***Figure 3.9**: Assessing progression in Huntington's disease. Graphical illustration of the trajectory of HD symptoms and signs over time, annotated to show what time period the different measures of onset and progression discussed in this paper cover. The TRACK-HD progression score uses longitudinal 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.

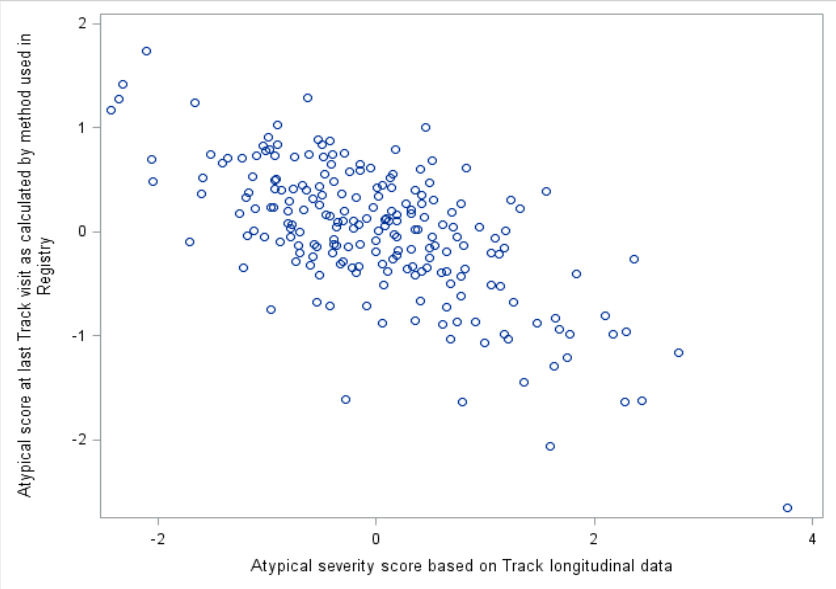


***Figure 3.10**: TRACK-HD and REGISTRY progression scores are correlated. Linear relationship between the longitudinal atypical severity scores used for the TRACK-HD analysis and cross-sectional atypical severity scores at the last TRACK visit when calculated using the method employed for the REGISTRY 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 ($p < 0{\cdot}0001$). 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 ($r=0{\cdot}315$; $p = 0{\cdot}002$).

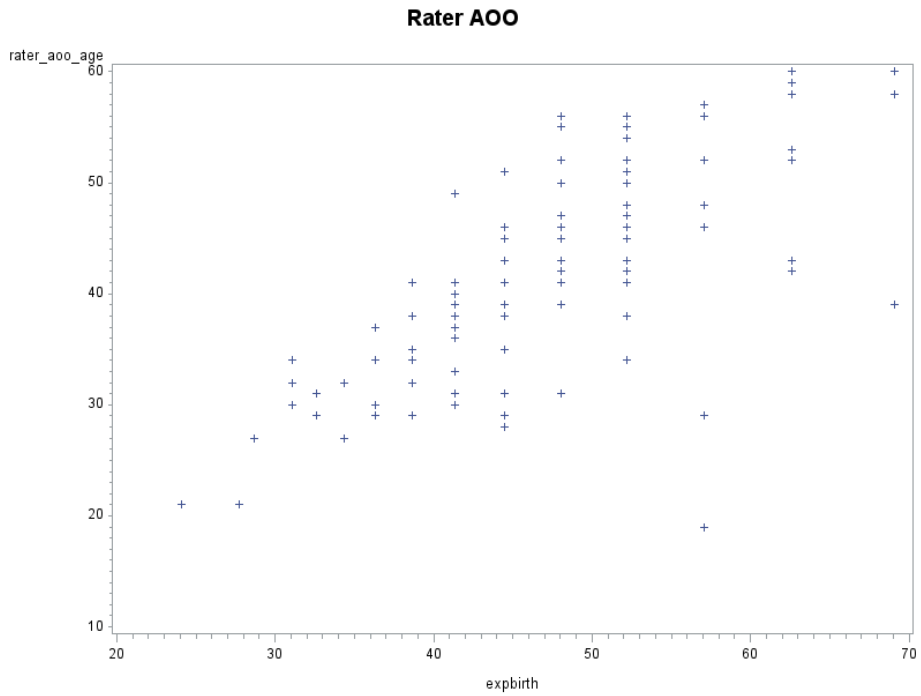


Figure 3.11*: Observed versus Expected Age of Onset among those who have Experienced Onset in the TRACK-HD analysis: amongst these 96 subjects who had experienced onset, the rater AAO showed the expected relation with predicted AAO based on CAG length. Earlier than predicted onset age was correlated 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.

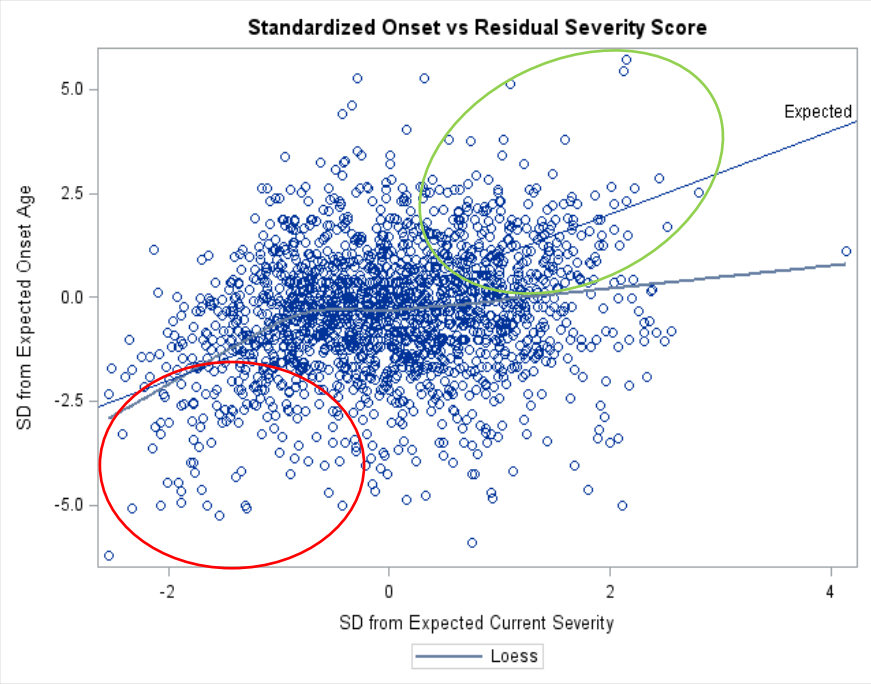


***Figure 3.12**: REGISTRY progression measure (Residual severity score) and atypical onset age (Standardised onset) are modestly correlated in REGISTRY. Note bias for very late expected onset for those with low CAG repeats. SD = Standard deviation. (Figure from Prof Langbehn)*

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* ($p = 5{\cdot}8 \times 10^{-8}$; 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 $P = 1{\cdot}2 \times 10^{-7}$ (Sveinbjornsson et al., 2016). The genes in this locus were the only ones to reach genome-wide gene-wide significance (i.e. $p<2{\cdot}5 \times 10^{-6}$ (Kiezun et al., 2012)) in a MAGMA analysis (de Leeuw et al., 2015) (*MTRNR2L2* $p=2{\cdot}14 \times 10^{-9}$; *MSH3* $p=2{\cdot}94 \times 10^{-8}$; *DHFR* $p=8{\cdot}37 \times 10^{-7}$; **Table 3.7**).

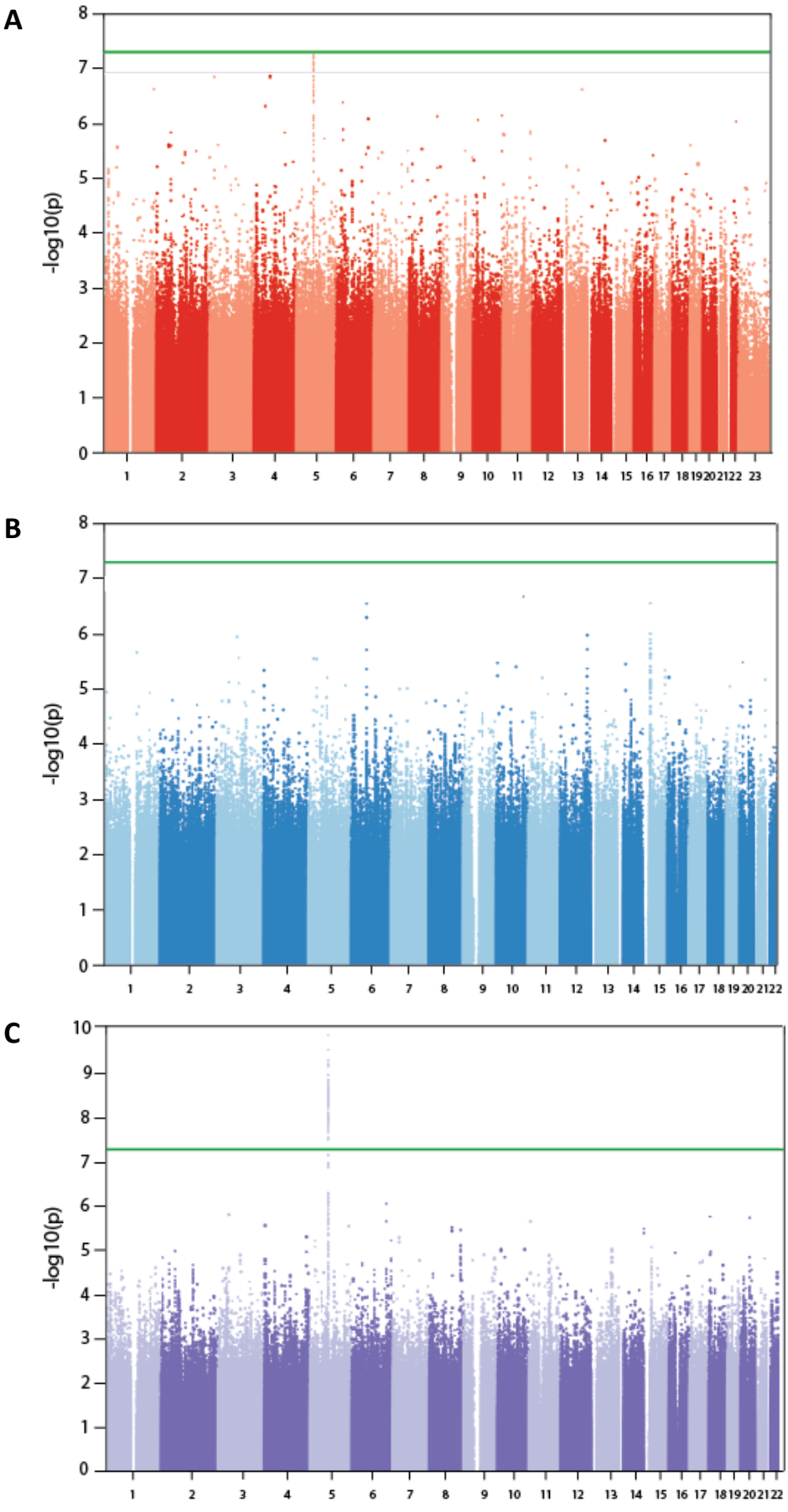

A
B
C
-log10(p)
0
1
2
3
4
5
6
7
8
9
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

***Figure 3.13**: Genome-wide Association Analysis of Progression Score. Green line in A-C: 5.8x10-8. (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))*

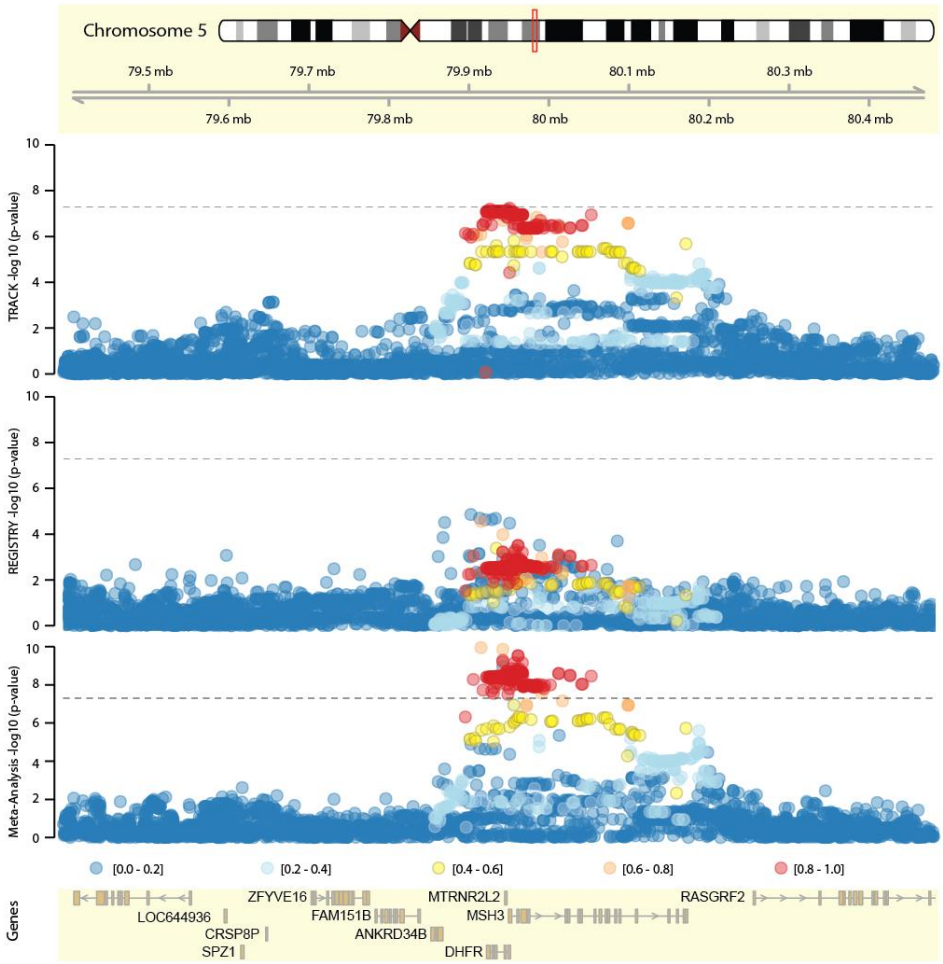


***Figure 3.14**: Locus zoom plot of the TRACK-HD (top), REGISTRY (middle) and meta-analysis (bottom) data showing the structure of linkage disequilibrium (LD) and –log^{10}(p-value) of the significant locus on chromosome. The top image shows the chromosome; the red square shows the region which is zoomed 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: $5x10^{-8}$. (Plots produced by Dr Pardiñas, these plots were also adapted for publication as Figure 3 in (Hensman Moss et al., 2017b))*

Chr	Start (BP)	End (BP)	Index SNP	A1	A2	MAF	INFO score	Beta	Standard Error	P-value	No. of SNPs	Length (KB)	Gene(s) tagged (+/- 20 KB)
5	79895438	80196258	rs557874766	G	C	0.238	1.000	-0.581	0.107	5.80E-08	380	300.82	DHFR, MSH3, MTRNR2L2
4	74064920	74362359	rs16849472	T	C	0.019	1.000	1.677	0.318	1.34E-07	10	297.44	AFM, AFP, ALB, ANKRD17, LOC728040
3	20860340	20919615	rs111902872	T	C	0.012	0.920	2.419	0.460	1.47E-07	2	59.276	none
1	239493679	239917976	rs115206404	A	G	0.009	0.805	2.598	0.503	2.46E-07	2	424.3	CHRM3, CHRM3-AS2
13	89829918	89856005	rs546753686	A	G	0.009	0.949	2.610	0.506	2.50E-07	2	26.088	none
6	31892827	31895971	rs188144048	G	C	0.016	1.000	-1.923	0.380	4.30E-07	2	3.145	C2, CFB, LOC102060414
4	52815077	52815077	rs151302971	C	T	0.060	0.998	0.963	0.192	4.98E-07	1	0.001	none
10	132818509	132881313	rs150136271	T	C	0.007	0.845	2.881	0.582	7.38E-07	3	62.805	TCERG1L
8	128074135	128092501	rs76712904	T	A	0.009	1.000	2.532	0.512	7.68E-07	13	18.367	PCAT2, PRNCR1
6	147033320	147049507	rs76605780	G	A	0.009	1.000	2.524	0.512	8.42E-07	4	16.188	ADGB

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/)

Entrez	Gene Symbol	Chr	Start	End	p(TRACK)	p(REGISTRY)	p(META)	p(GeM)
4437	MSH3	5	79950467	80172634	2.94E-08	9.52E-04	8.89E-11	1.98E-02
1719	DHFR	5	79922045	79950800	8.37E-07	8.45E-04	1.04E-09	6.46E-02
100462981	MTRNR2L2	5	79945819	79946854	2.15E-09	1.20E-03	1.88E-09	N/A
7852	CXCR4	2	136871919	136875725	3.96E-04	6.46E-03	4.40E-06	8.22E-02
54893	MTMR10	15	31231144	31283807	3.01E-01	3.49E-07	1.42E-05	2.74E-11
10873	ME3	11	86152150	86383678	5.07E-03	5.81E-02	2.19E-05	4.78E-01
118	ADD1	4	2845584	2931803	5.84E-02	2.82E-03	2.95E-05	1.16E-03
8605	PLA2G4C	19	48551100	48614109	3.53E-03	1.90E-01	6.73E-05	5.82E-02
9209	LRRFIP2	3	37094117	37217851	5.37E-02	3.16E-04	6.98E-05	3.19E-04
8690	JRKL	11	96123158	96126727	4.37E-05	5.29E-02	8.39E-05	8.91E-01
2788	GNG7	19	2511218	2702746	1.02E-01	3.62E-03	1.11E-04	7.83E-02
22909	FAN1	15	31196055	31235311	5.30E-01	2.16E-06	1.15E-04	1.68E-09
4292	MLH1	3	37034841	37092337	6.98E-02	3.97E-04	1.28E-04	3.91E-04
79780	CCDC82	11	96085929	96123083	2.34E-03	5.99E-02	1.30E-04	7.39E-02
9852	EPM2AIP1	3	37027357	37034795	7.94E-02	4.29E-04	1.53E-04	1.39E-03
4308	TRPM1	15	31293264	31453476	4.78E-01	1.77E-05	1.83E-04	8.33E-04
115509	ZNF689	16	30614686	30621682	1.86E-02	7.52E-03	1.85E-04	9.53E-01
23167	EFR3A	8	132916356	133025889	3.90E-02	1.26E-02	2.32E-04	2.19E-01
146540	ZNF785	16	30591994	30597092	1.17E-01	1.53E-02	2.45E-04	9.39E-01

909	CD1A	1	158223927	158228059	1.60E-01	5.96E-04	3.27E-04	4.85E-01

***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/*)*

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)**.

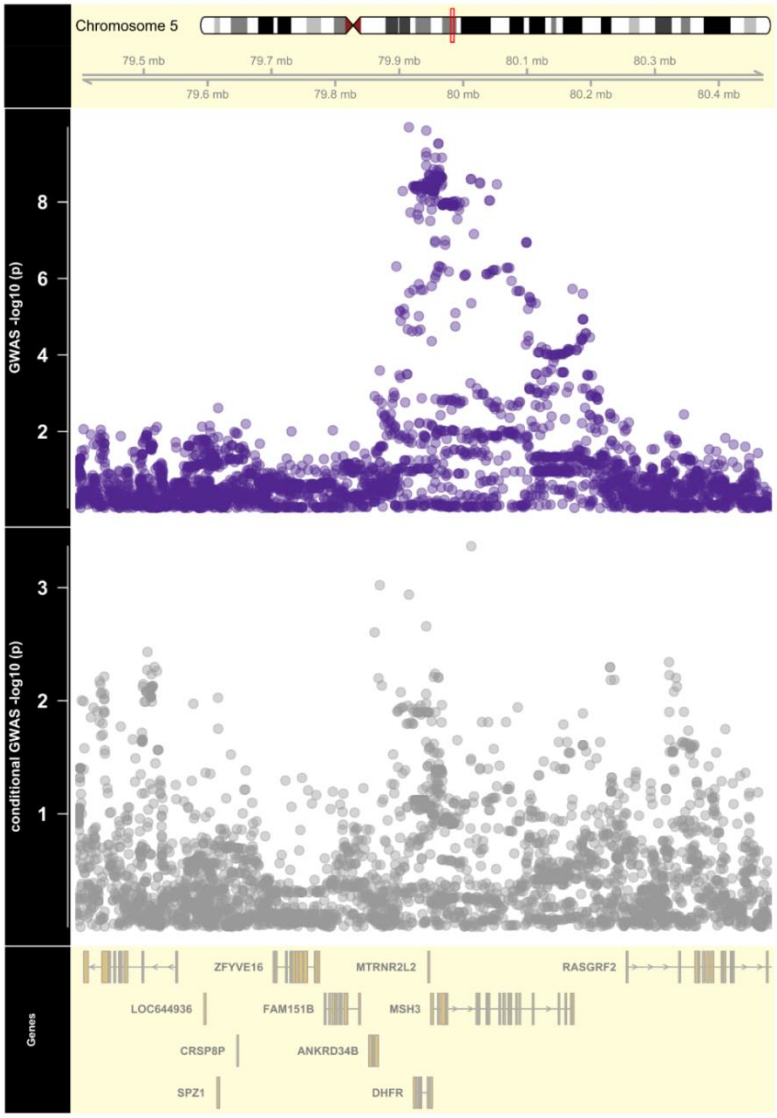


Figure 3.15*: Regional plot of TRACK-HD and REGISTRY meta-analysis GWAS signal in the MSH3-DHFR region before (top) and after (bottom) conditioning on the most significant SNP in TRACK-HD (rs557874766). The lack of significant association after conditioning on this SNP is consistent with here being only 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 ($p = 1{\cdot}39 \times 10^{-5}$) on a narrower locus (chr5:79902336-79950781), 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=8·45x10^{-4}, *MSH3* p=9·36x10^{-4}, *MTRNR2L2* p=1·20x10^{-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)**.

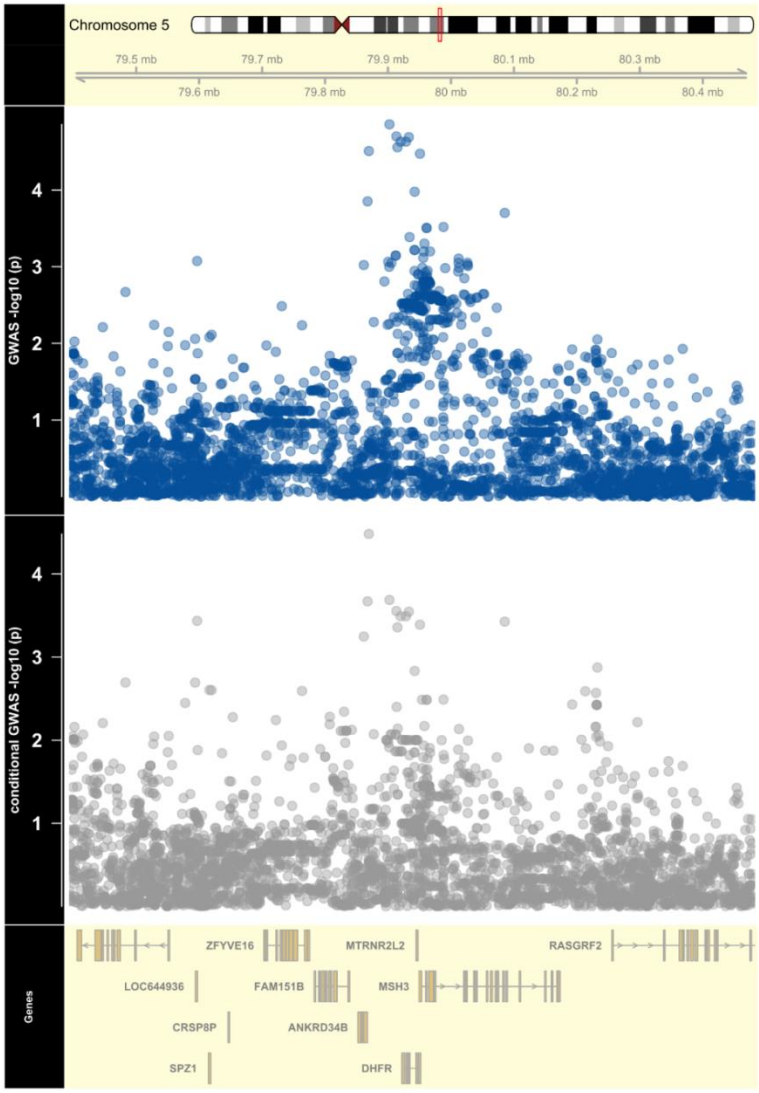


***Figure 3.16**: Regional plot of REGISTRY GWAS signal in the MSH3-DHFR region before (top) and after (bottom) conditioning on the most significant SNP in TRACK-HD (rs557874766).* The significance of association is largely unaffected by conditioning on this SNP. This indicates that rs557874766 does 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=1.12x10^{-10}), with the p-value of rs557874766 being

$1.58x10^{-8}$. 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.

Index SNP	P-value	Clump coordinates	Clump size (KB)	Gene(s) tagged
rs1232027	1.12E-10	chr5:79895438..80198404	302.967	DHFR, MSH3, MTRNR2L2
rs73786719	8.53E-07	chr6:147034576..147037984	3.409	ADGB
rs114688092	1.51E-06	chr3:47026101..47315538	289.438	CCDC12, KIF9, KIF9-AS1, KLHL18, NBEAL2, NRADDP, SETD2
rs79029191	1.67E-06	chr18:8053863..8080538	26.676	PTPRM
rs932428	1.79E-06	chr20:37518361..37876772	358.412	DHX35, FAM83D, LOC339568, PPP1R16B
rs3889139	2.13E-06	chr11:6885429..6917038	31.61	OR2D2, OR10A2, OR10A4, OR10A5
rs114643193	2.65E-06	chr4:2844682..2939191	94.51	ADD1, MFSD10, NOP14, NOP14-AS1, SH3BP2
rs6882169	2.72E-06	chr5:167668230..167668230	0.001	CTB-178M22.2, TENM2
rs80260687	2.92E-06	chr8:97232364..97304966	72.603	MTERFD1, PTDSS1, UQCRB
rs28406206	3.13E-06	chr14:105680474..105688082	7.609	BRF1
rs4736525	3.37E-06	chr8:132924474..133030989	106.516	EFR3A, OC90
rs78621558	4.44E-06	chr5:80012735..80012735	0.001	MSH3
rs72715653	4.80E-06	chr4:178641337..178730329	88.993	LINC01098, LINC01099
rs4720024	4.94E-06	chr7:30941255..30942312	1.058	AQP1, FAM188B, INMT-FAM188B
rs117933444	5.75E-06	chr6:167362873..167410443	47.571	FGFR1OP, MIR3939, RNASET2
rs116220136	5.82E-06	chr5:23353255..23436446	83.192	none
rs8031584	8.15E-06	chr15:31185616..31292023	106.408	FAN1, MTMR10, TRPM1
rs3013648	9.10E-06	chr13:85296644..85374146	77.503	none
rs11197481	9.12E-06	chr10:117708803..117708803	0.001	ATRNL1
rs117440785	9.15E-06	chr10:17411451..17531334	119.884	ST8SIA6, ST8SIA6-AS1

Table 3.8*: Independent association signals from the meta-analysis of TRACK-HD and REGISTRY Progression GWAS (at p-value < 10-5).*

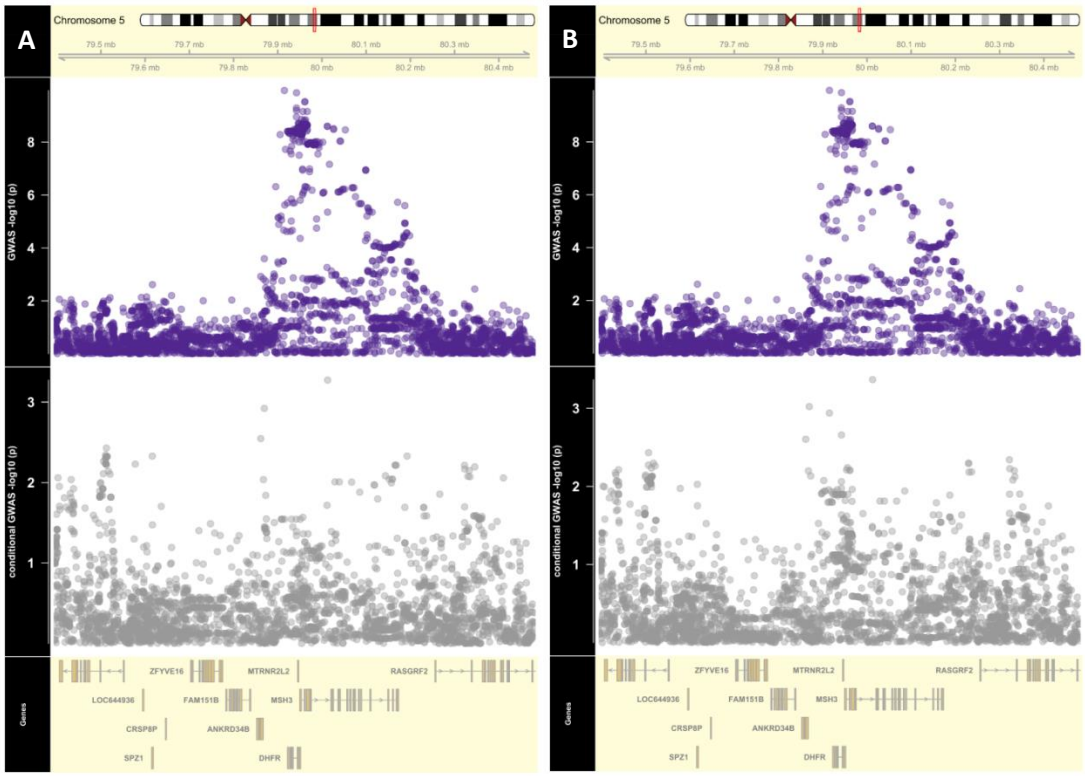


***Figure 3.17**: Conditional analysis. (A) Regional plot of TRACK-HD and REGISTRY meta-analysis GWAS signal in the MSH3-DHFR region before(top) and after (bottom) conditioning on the most significant SNP in the meta-analysis (rs1232027). The lack of significant association after conditioning on this SNP is consistent with here 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 here 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**).

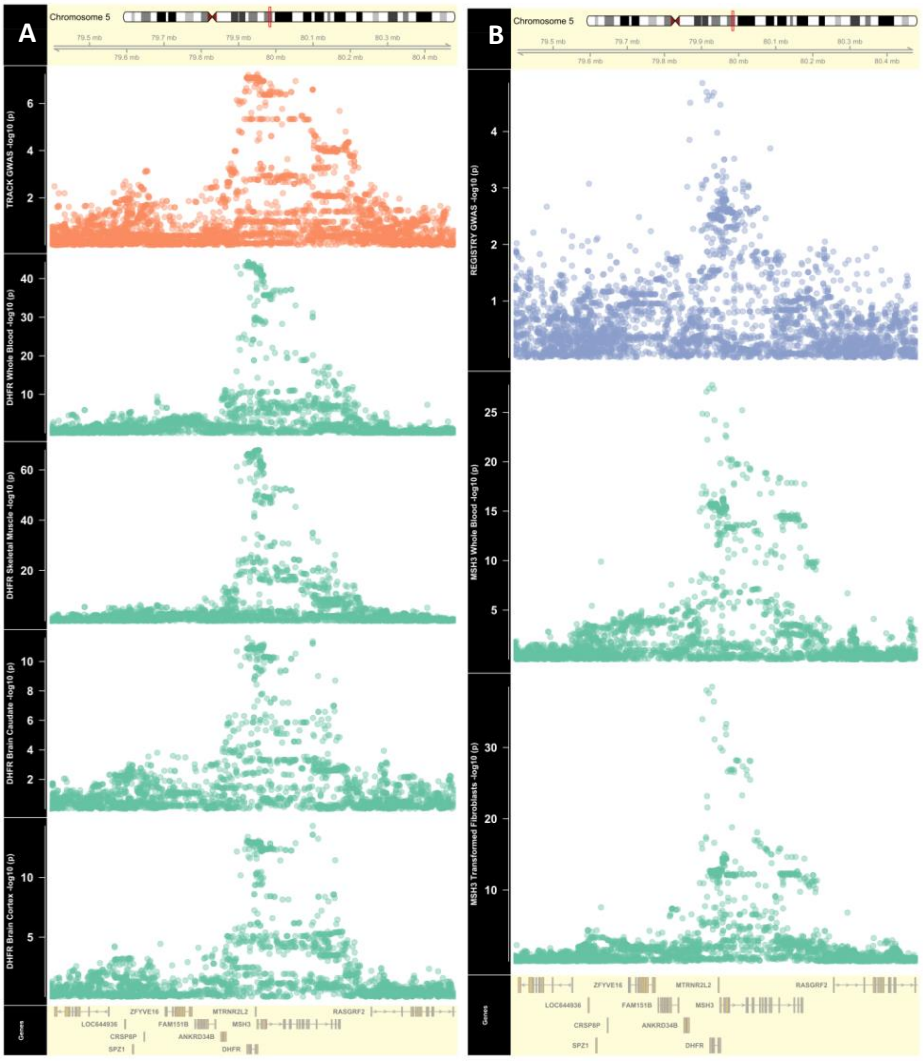


Figure 3.18*: Expression analysis.* ***A:*** *Regional plot of TRACK-HD GWAS signal in the MSH3-DHFR region (top, red), along with GTeX eQTL associations with DHFR expression in (top-bottom) whole blood, skeletal muscle, cerebellum, cortex.* ***B****: Regional plot of REGISTRY GWAS signal in the MSH3-DHFR 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.

SNP	BP	GWAS.ref	GWAS.alt	GWAS.p	GWAS.beta	MSH3.blood.eQTL.p	MSH3.blood.eQTL.beta	MSH3.Fibroblasts.eQTL.p	MSH3.Fibroblasts.eQTL.beta	MSH3.Cerebellum.eQTL.p	MSH3.Cerebellum.eQTL.beta	MSH3.Caudate.eQTL.p	MSH3.Caudate.eQTL.beta	MSH3.Cortex.eQTL.p	MSH3.Cortex.eQTL.beta	GTEx.ref	GTEx.alt
rs863215	79948005	T	C	8.29E-08	-0.5441	1.12E-15	-0.4645	1.43E-13	-0.5248	0.0487	-0.1887	0.00857	-0.2471	0.00042	-0.4224	T	C
rs1478834	79949575	A	C	8.29E-08	-0.5441	1.12E-15	-0.4645	1.43E-13	-0.5248	0.0487	-0.1887	0.00857	-0.2471	0.00042	-0.4224	A	C
rs1382539	79952154	A	G	8.70E-08	-0.5432	1.12E-15	-0.4645	1.43E-13	-0.5248	0.0487	-0.1887	0.00857	-0.2471	0.00042	-0.4224	A	G
rs1677703	79957737	T	C	1.13E-07	-0.5342	6.12E-17	-0.4823	4.06E-15	-0.5518	0.0487	-0.1887	0.00857	-0.2471	0.00042	-0.4224	T	C
rs1650667	79962226	T	C	1.13E-07	-0.5342	3.55E-16	-0.4642	3.58E-15	-0.543	0.0487	-0.1887	0.00705	-0.247	0.00063	-0.3989	T	C
rs1650666	79962439	A	G	1.13E-07	-0.5342	3.55E-16	-0.4642	3.58E-15	-0.543	0.0487	-0.1887	0.00705	-0.247	0.00063	-0.3989	A	G
rs857287	80098957	C	G	2.63E-07	-0.5296	1.92E-11	0.40096	8.39E-10	0.443	0.6642	0.0396	0.05648	0.18319	0.01774	0.28212	G	C
rs863214	79984714	G	A	3.09E-07	-0.5302	6.00E-14	-0.44	6.09E-13	-0.5053	0.1665	-0.1406	0.01627	-0.2326	0.00142	-0.4023	G	A
rs1222809	79917517	G	A	3.14E-07	-0.5355	3.02E-16	-0.4728	1.08E-12	-0.5022	0.1065	-0.16	0.01615	-0.2273	0.00052	-0.4288	G	A
rs836794	80012251	A	C	4.01E-07	-0.5367	2.63E-14	-0.4445	7.40E-13	-0.509	0.1628	-0.1404	0.03621	-0.1998	0.00132	-0.3996	A	C

***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/*

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.

Entrez	Gene Symbol	Chr	Start	End	p(TRACK)	p(TRACKcond)	p(REG)	p(REGcond)	p(META)	p(METAcond)
100462981	MTRNR2L2	5	79945819	79946854	2.15E-09	6.97E-07	1.20E-03	5.51E-02	1.88E-09	3.24E-06
4437	MSH3	5	79950467	80172634	2.94E-08	4.98E-06	9.52E-04	6.24E-02	8.89E-11	6.86E-07
1719	DHFR	5	79922045	79950800	8.37E-07	3.74E-05	8.45E-04	4.42E-02	1.04E-09	2.20E-06
8339	HIST1H2BG	6	26216428	26216872	1.56E-05	4.40E-03	5.10E-01	7.88E-01	2.31E-02	1.14E-01
387638	C10orf113	10	21414692	21435488	2.45E-05	1.46E-05	6.00E-01	3.42E-01	1.65E-01	2.22E-01
8690	JRKL	11	96123158	96126727	4.37E-05	1.30E-03	5.29E-02	4.40E-02	8.39E-05	1.06E-03
55269	PSPC1	13	20248892	20357159	4.80E-05	4.79E-04	5.80E-01	8.84E-01	7.95E-02	4.28E-02
3007	HIST1H1D	6	26234440	26235216	6.67E-05	2.82E-03	5.51E-01	8.13E-01	9.04E-03	6.40E-02
1553	CYP2A13	19	41594356	41602100	7.81E-05	4.06E-05	6.62E-01	6.74E-01	1.96E-03	4.41E-03
8369	HIST1H4G	6	26246839	26247205	8.49E-05	2.36E-03	5.83E-01	8.08E-01	1.18E-02	7.36E-02

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/

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* ($7{\cdot}21x10^{-4}$) show association in TRACK, with *MSH3* ($9{\cdot}52x10^{-4}$) and *MLH1* ($3{\cdot}97x10^{-4}$) showing association in REGISTRY (**http://hdresearch.ucl.ac.uk/data-resources/).**

Pathway	p(TRACK)	p(REGISTRY)	P(META)	p(GeM)	Description
GO: 32300	3·46E-09	8·34E-04	1.14E-11	3·82E-05	mismatch repair complex
KEGG 3430	2·79E-07	4·80E-02	1.34E-16	6·65E-06	KEGG_MISMATCH_REPAIR
GO: 30983	6·66E-07	4·20E-04	3.17E-11	7·43E-06	mismatched DNA binding
GO: 6298	3·53E-06	4·59E-02	6.54E-09	3·25E-06	mismatch repair
GO:	1·82E-02	1·10E-01	6.40E-04	5·74E-	MutSalpha complex binding

32407				05	
GO: 32389	2·25E-02	4·69E-02	5.23E-04	1·66E-05	MutLalpha complex
GO: 33683	8·01E-02	5·87E-04	6.74E-03	1·69E-06	nucleotide-excision repair, DNA incision
GO: 90141	3·32E-01	5·93E-02	7.87E-01	2·30E-06	positive regulation of mitochondrial fission
GO: 1900063	4·10E-01	7·29E-01	6.93E-01	8·39E-05	regulation of peroxisome organization
GO: 90200	4·58E-01	5·44E-01	5.28E-01	8·89E-08	positive regulation of release of cytochrome c from mitochondria
GO: 90140	5·39E-01	3·32E-01	8.10E-01	1·57E-05	regulation of mitochondrial fission
GO: 10822	6·21E-01	6·28E-01	8.53E-01	7·63E-05	positive regulation of mitochondrion organization
GO: 4748	9·64E-01	6·97E-01	9.79E-01	2·66E-05	ribonucleoside-diphosphate reductase activity, thioredoxin disulfide as acceptor
GO: 16728	9·64E-01	6·97E-01	9.79E-01	2·66E-05	oxidoreductase activity, acting on CH or CH2 groups, disulfide as acceptor

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

Gene Set	p(TRACK)	p(REGISTRY)	p(META)	p (GeM)	Description1	Description 2	Description 3	Description 4
2071015	9.05E-07	4.43E-03	2.93E-11	2.01E-02	Repair pathway	SSR	MMR	Mismatch & loop recognition factors
2071000	2.43E-06	6.85E-02	1.49E-14	5.15E-04	Repair pathway	SSR	MMR	
2070000	5.77E-03	4.76E-02	3.32E-07	1.42E-02	Repair pathway	SSR		
2071017	1.95E-02	2.44E-02	5.84E-05	8.92E-08	Repair pathway	SSR	MMR	MutL homologs
2111513	4.71E-02	2.55E-01	8.12E-01	2.86E-03	Repair pathway	Associated process	TLS	DNA polymerases
2070600	5.02E-02	7.99E-01	1.10E-01	2.92E-01	Repair pathway	SSR	NER	
2070607	5.18E-02	7.61E-01	3.02E-02	2.26E-01	Repair pathway	SSR	NER	TCR (Transcription coupled repair)
2071104	5.35E-02	3.90E-01	2.07E-02	5.37E-02	Repair pathway	SSR	BER	LONG PATCH-BER factors
2022100	6.69E-02	3.19E-02	7.21E-04	7.29E-02	Repair pathway	DSR	Alt-NHEJ	
1100000	7.52E-02	6.14E-01	1.94E-01	6.13E-01	Associated process	DNA replication		
1080700	8.99E-02	8.35E-01	2.82E-01	4.92E-01	Associated process	Checkpoint factors	S-CC phase	
1051930	1.02E-01	5.68E-01	1.30E-01	7.62E-01	Associated process	Ubiquitin response	Ubiquitin- conjugating enzymes (E2)	UBL-conjugating enzymes
2000000	1.13E-01	2.60E-01	1.03E-03	1.11E-02	Repair pathway			
2070605	1.14E-01	5.00E-01	8.14E-01	4.64E-01	Repair pathway	SSR	NER	DNA polymerase epsilon
1030000	1.59E-01	1.90E-01	3.59E-01	2.63E-01	Associated process	Telomere maintenance		
2070606	1.60E-01	9.56E-01	6.55E-01	5.49E-01	Repair pathway	SSR	NER	DNA polymerase kappa
2071020	1.73E-01	3.14E-01	9.86E-03	7.97E-02	Repair pathway	SSR	MMR	Other MMR factors

1051900	1.97E-01	7.69E-01	1.71E-01	8.19E-01	Associated process	Ubiquitin response	Ubiquitin- conjugating enzymes (E2)	
2071023	2.15E-01	1.73E-01	7.67E-02	5.90E-01	Repair pathway	SSR	MMR	RPA (replication factor A)

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.*

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

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).

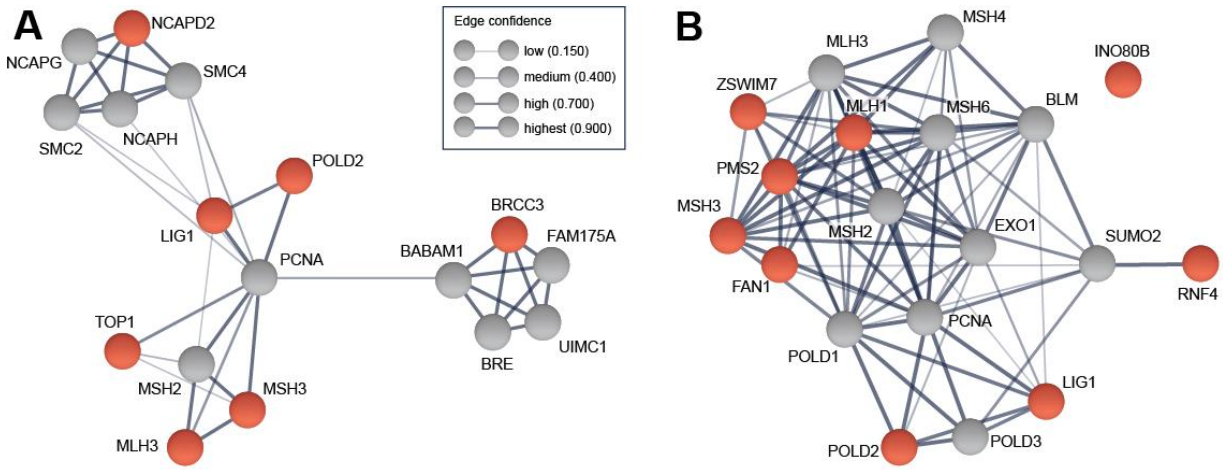


Figure 3.19*: Significant genes are functionally linked and may cause somatic expansion of the HTT CAG repeat tract. STRING diagram showing all proteins from the Pearl et al (20) dataset with gene-wide p-values for association with Huntington's disease progression < 0.02 in* ***A****: the TRACK-HD dataset 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.

Pathway	p(TRACK)	p(TRACK no MSH3)	p(REGISTRY)	p(REGISTRY no MSH3)	p(META)	p(META no MSH3)	Description
GO: 32300	3.455E-09	0.04127	0.0008336	0.07162	1.13E-11	0.001024	mismatch repair complex
KEGG 3430	2.794E-07	0.04521	0.04795	0.1471	1.34E-16	0.000107	KEGG_MISMATCH_REPAIR
GO: 30983	6.661E-07	0.1001	0.0004195	0.009264	3.17E-11	0.000274	mismatched DNA binding
GO: 6298	0.000003533	0.2446	0.04589	0.1839	6.54E-09	0.0729	mismatch repair
GO: 32407	0.01818	0.01818	0.1101	0.1101	0.000640	0.000640	MutSalpha complex binding
GO: 32389	0.02249	0.02249	0.04688	0.04688	0.000523	0.000523	MutLalpha complex
GO: 33683	0.08014	0.08014	0.0005874	0.0005874	0.00675	0.00675	nucleotide-excision repair, DNA incision
GO: 90141	0.3318	0.3318	0.05934	0.05934	0.7872	0.7872	positive regulation of mitochondrial fission
GO: 1900063	0.4103	0.4103	0.7287	0.7287	0.6926	0.6926	regulation of peroxisome organization
GO: 90200	0.4582	0.4582	0.544	0.544	0.5280	0.5280	positive regulation of release of cytochrome c from mitochondria

***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.

Pathway	**p(TRACK)**	**q(TRACK)**	**p(REGISTRY)**	**Description**
GO: 32300	3.46E-09	1.22E-05	8.34E-04	mismatch repair complex
GO: 43570	8.02E-09	1.41E-05	3.20E-03	maintenance of DNA repeat elements
GO: 32135	2.13E-08	2.50E-05	6.99E-03	DNA insertion or deletion binding
GO: 710	4.59E-08	4.05E-05	1.19E-02	meiotic mismatch repair
GO: 51095	9.01E-08	6.35E-05	1.52E-02	regulation of helicase activity
GO: 404	1.14E-07	6.70E-05	3.38E-03	loop DNA binding
KEGG 3430	4.04E-07	2.03E-04	4.80E-02	KEGG MISMATCH REPAIR
GO: 32138	5.69E-07	2.31E-04	1.28E-02	single base insertion or deletion binding
GO: 19237	5.91E-07	2.31E-04	5.97E-03	centromeric DNA binding
GO: 30983	6.66E-07	2.35E-04	4.20E-04	mismatched DNA binding
GO: 32142	9.05E-07	2.90E-04	4.43E-03	single guanine insertion binding
GO: 403	1.87E-06	5.14E-04	7.03E-03	Y-form DNA binding
REACTOME 1234	1.90E-06	5.14E-04	2.66E-02	REACT:TETRAHYDROBIOPTERIN (BH4) SYNTHESIS
GO: 32139	2.85E-06	7.16E-04	2.45E-04	dinucleotide insertion or deletion binding
GO: 6298	3.53E-06	8.30E-04	6.74E-03	mismatch repair
REACTOME 656	4.76E-06	1.05E-03	8.13E-02	REACT:METABOLISM OF FOLATE AND PTERINES
GO: 217	7.69E-06	1.59E-03	2.91E-02	DNA secondary structure binding

GO: 16447	1.13E-05	2.18E-03	3.66E-02	somatic recombination of immunoglobulin gene segments
GO: 51096	1.18E-05	2.18E-03	5.93E-03	positive regulation of helicase activity
GO: 16445	1.36E-05	2.39E-03	4.55E-02	somatic diversification of immunoglobulins
REACTOME 452	3.99E-05	6.70E-03	9.11E-02	REACT:G1 S- SPECIFIC TRANSCRIPTION
GO: 45910	5.44E-05	8.71E-03	5.43E-03	negative regulation of DNA recombination
KEGG 790	6.00E-05	9.19E-03	1.93E-02	KEGG FOLATE BIOSYNTHESIS
GO: 16444	7.07E-05	9.96E-03	8.28E-02	somatic cell DNA recombination
GO: 2562	7.07E-05	9.96E-03	8.28E-02	somatic diversification of immune receptors via germline recombination within a single locus
GO: 2200	1.15E-04	1.55E-02	1.02E-01	somatic diversification of immune receptors
REACTOME 659	7.20E-04	9.07E-02	1.16E-01	REACT:METABOLISM OF NITRIC OXIDE
REACTOME 367	7.20E-04	9.07E-02	1.16E-01	REACT:ENOS ACTIVATION AND REGULATION
GO: 35825	8.81E-04	1.03E-01	7.04E-02	reciprocal DNA recombination
GO: 7131	8.81E-04	1.03E-01	7.04E-02	reciprocal meiotic recombination

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 ($p=1.12x10^{-10}$) (Hensman Moss et al., 2017b). The lead SNP in TRACK-HD, rs557874766, is a coding variant in MSH3 (5.80×10^{-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.

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

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% CI=0.16,0.66) units per year in the rate of change of TMS, and a reduction of approximately 0.12 (95% CI=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).

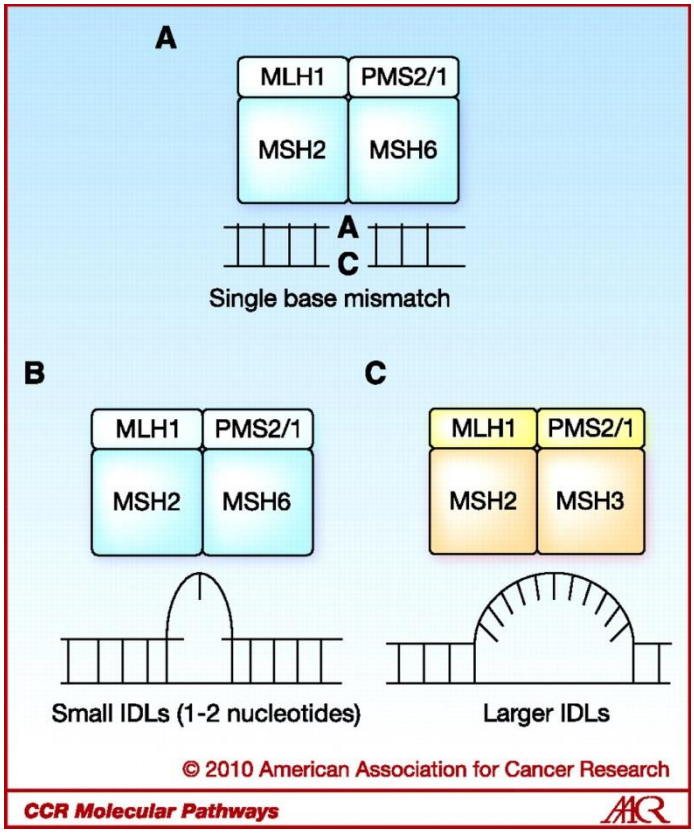


Figure 3.20*: Schematic of DNA damage recognized by the MMR pathway.* ***A****, the MutSα (MSH2/MSH6) heterodimer recognizes base-base mismatches and* ***B****, small insertion-deletion loops (IDL). The MutSβ (MSH2/MSH3) heterodimer recognizes single nucleotide IDLs and* ***C****, longer IDLs (10-nucleotide 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).

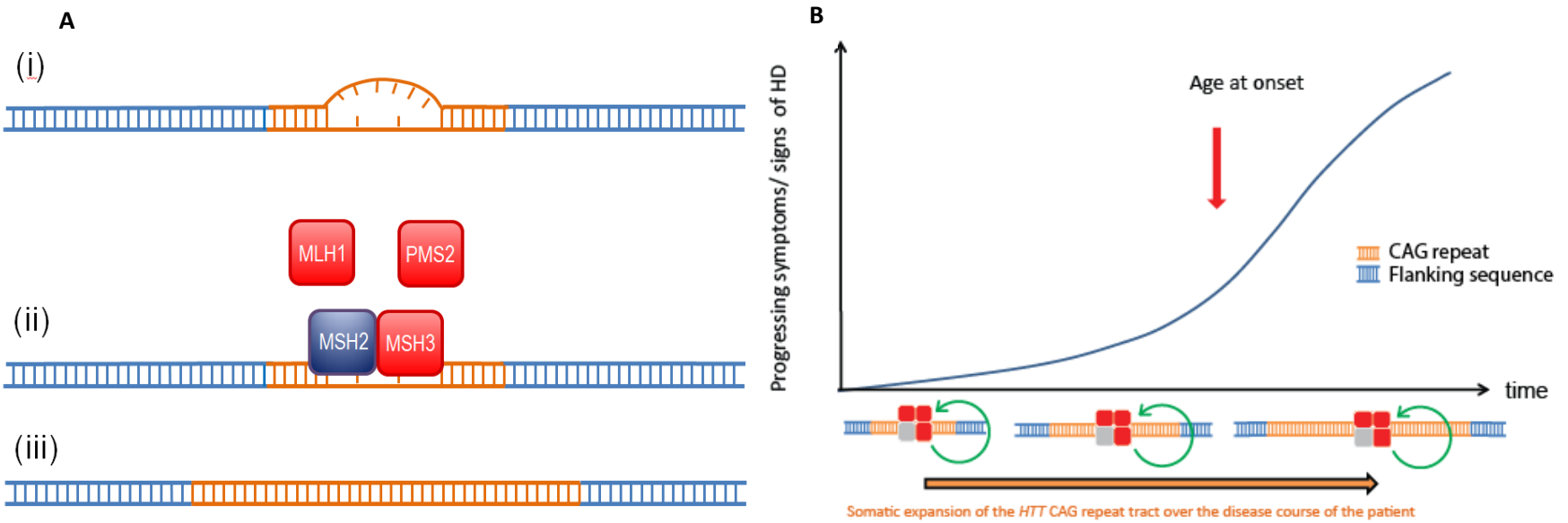


Figure 3.21*: A Schematic diagram showing how DNA mismatch repair proteins may be involved in somatic expansion of the CAG tract. Proteins with p<0.01 in the meta-analysed progression GWAS are coloured red. (i) The CAG repeat DNA is partly unwound by lesions, constraints of the CAG tract structure or 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. *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). *MSH2* forms a heterodimeric complex with *MSH3* or *MSH6* (**Figure 3.21**). We did not find any association between 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 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 1.28×10^{-4} in the meta-analysis; PMS2 had a gene-wide p-value of 8.84×10^{-3} in the meta-analysis; by contrast MLH3 had a gene-wide p-value of 1.47×10^{-1} 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 MutLα heterodimer which associates with the MutSα or MutSβ 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 MutLβ and MutLγ, 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 *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 5.35×10^{-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 *LIG1*, and its interaction with *PCNA* (Lopez Castel et al., 2009). *POLD2*, which encodes the catalytic subunit of DNA polymerase delta, had a gene-wide p-value of 2.75×10^{-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 ($p=2{\cdot}35x10^{-6}$). 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.

Repeat disorder	**Gene**	**Prevalence (per 100,000 European population)**	**Phenotype**	**Variance in AAO explained by repeat length (heritability of residual variability)**	**Normal range**	**Pathogenic range**	**Somatic instability**
HD	HTT	3-10 (Wood, 2012, Rawlins, 2010b, Bates et al., 2014, Warby et al., 2014)	Involuntary movements, cognitive impairment	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)	6-35	40-121	Yes
SCA1	ATXN1	0.16 (Durr, 2010)	Ataxia, ophthalmoplegia, pyramidal and extrapyramidal features (Subramony, 2012)	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))	6-38	45-83	Yes (Hashida et al., 1997)

SCA2	ATXN2	0.2 (Durr, 2010)	Ataxia, neuropathy, ophthalmoplegia, extrapyramidal features (Kawai et al., 2009, Burk, 1999, Storey et al., 1999, Sokolovsky et al., 2010)	50-80% (Giunti et al., 1998, Tezenas du Montcel et al., 2014b, Velazquez Perez et al., 2009, Globas et al., 2008, Hayes et al., 2000, Geschwind et al., 1997, Pulst et al., 2005, van de Warrenburg et al., 2002, van de Warrenburg et al., 2005, Lorenzetti et al., 1997) (17-59%) (Pulst et al., 2005, van de Warrenburg et al., 2005)	15-31	33-500	Yes (Matsuura et al., 1999)
SCA3	ATXN3	0.4 (Durr, 2010)	Ataxia, pyramidal signs, neuropathy, extrapyramidal features, ophthalmoplegia (Durr et al., 1996)	45-80% (Bettencourt and Lima, 2011, Durr et al., 1996) (46%) (van de Warrenburg et al., 2005)	12-44	52-87	Yes (Hashida et al., 1997)
SCA6	CACNA1A	0.04 (Durr, 2010)	Ataxia (Rub et al., 2013)	26-52% (Tezenas du Montcel et al., 2014a, van de Warrenburg et al., 2005) (no significant heritable component) (van de Warrenburg et al., 2005)	4-18	20-33	Unknown

SCA7	ATXN7	0.12 (Durr, 2010)	Ataxia, macular degeneration, ophthalmoplegia, pyramidal and extrapyramidal features (Sokolovsky et al., 2010)	71-88% (Rub et al., 2013, Tezenas du Montcel et al., 2014a) (no significant heritable component) (van de Warrenburg et al., 2005)	3-19	37-460	Yes (Trang et al., 2015)
SCA17	TBP	<0.02 (Durr, 2010)	Ataxia, pyramidal signs, dementia, seizures, extrapyramidal features (Schneider et al., 2006, Rolfs et al., 2003)	Unknown	25-40	49-66	Unknown
SCA12	PPP2R2	<0.02	Ataxia, tremor, neuropathy	Unknown	4-32	40-78	Yes
DRPLA	ATN1	0.005-0.04	Myoclonus, epilepsy, ataxia, dementia (Tsuji, 1999)	50-68% (Wardle et al., 2009, Potter, 1996)	6-35	48-93	Yes (Hashida et al., 1997)
SBMA	AR	0.65-2 (Udd et al., 1998, Spada, 2014)	Limb and bulbar weakness, neuropathy, endocrine features (Atsuta et al., 2006, Kennedy et al., 1968, Rhodes et al., 2009)	29% (Sinnreich et al., 2004)	9-34	38-72	Yes (Tanaka et al., 1999)

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), SPATAX 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.

Cohort	Disease								Gender			Ethnicity		
	HD	SCA1	SCA2	SCA3	SCA6	SCA7	SCA17	Total	Male	Female	Unavailable	Caucasian	Other	Unavailable
Athens, Greece	351	0	0	0	0	0	0	351	174	177	0	351	0	0
Azores, Portugal	0	0	0	91	0	0	0	91	48	43	0	91	0	0
London, UK	0	30	66	45	69	7	1	218	103	82	33	109	72	37
Mexico	0	0	113	0	0	66	6	185	91	94	0	0	185	0
Paris, France	0	147	115	261	0	0	0	523	279	244	0	463	42	18
TRACK-HD, Europe	94	0	0	0	0	0	0	94	46	48	0	90	4	0
Total	445	177	294	397	69	73	7	1462	741	688	33	1104	303	55
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 *RRM2B* and *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, *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**.

SNP ID	Chr: position (bp) (GRCh37/hg19)	Gene symbol	Functional annotation	P (GeM-HD)	MAF*	Genotype call rate*	P (HWE)*
rs1800937	2:48025764	MSH6	Stop gained	4.30E-03	0.074	0.973	0.840
rs4150407	2:128049631	ERCC3	Intron variant	4.60E-04	0.479	0.964	0.003
rs5742933	2:190649316	PMS1	NMD transcript variant	9.49E-04	0.205	0.972	1.000
rs1799977	3:37053568	MLH1	Missense variant	7.16E-07	0.28	0.966	0.354
rs6151792	5:80056961	MSH3	Intron variant	2.09E-04	0.117	0.978	0.706
rs115109737	5:80102444	MSH3	Intron variant	4.50E-04	0.041	0.980	0.489
rs71636247	5:80118976	MSH3	Intron variant	2.55E-04	0.034	0.976	1.000
rs1805323	7:6026942	PMS2	Missense variant	3.04E-02	0.043	0.975	0.736
rs12531179	7:6028687	PMS2	Intron variant	3.84E-05	0.169	0.971	0.925
rs3735721	8:103217695	RRM2B	3’_UTR_variant	5.68E-07	0.083	0.971	0.058
rs1037700	8:103250775	RRM2B	Intron variant	5.03E-08	0.094	0.973	0.002
rs5893603	8:103250839	RRM2B	Frameshift variant	4.28E-08	0.093	0.973	0.007
rs1037699	8:103250930	RRM2B	Missense variant	2.70E-08	0.094	0.976	0.002
rs16869352	8:103306033	UBR5	Synonymous variant	4.01E-07	0.08	0.975	0.030
rs61752302	8:103311153	UBR5	Synonymous variant	3.03E-03	0.026	0.977	0.621
rs72734283	14:75495059	MLH3	Intron variant	4.32E-03	0.089	0.971	0.623
rs175080	14:75513828	MLH3	Missense variant	7.72E-03	0.435	0.971	0.447
rs146353869	15:31126401	FAN1	Intron variant	4.30E-20	0.017	0.973	1.000

rs114136100	15:31197976	FAN1	Synonymous variant	8.49E-16	0.019	0.976	0.423
rs150393409	15:31202961	FAN1	Missense variant	9.34E-18	0.013	0.975	1.000
rs3512	15:31235005	FAN1	3'_UTR_variant	5.28E-13	0.283	0.973	1.000
rs20579	19:48668830	LIG1	NMD transcript variant	6.65E-03	0.134	0.942	0.732

***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.

SNPs	HGVS Names	SNP to Chromosome	Seed sense sequences for KASP assay design
rs1800937	NC_000002.11:g.48025764C>T	Forward	TTGCCTGGCAGGTAGGCACAACTTA**[C>T]**GTAACAGATAAGAGTGAAGAAGATA
rs4150407	NC_000002.11:g.128049631T>C	Reverse	AGTACACAATGGGAAGGTGGTCCAT**[A>G]**GACAAGAGCCTTCACCAGAAACTGA
rs5742933	NC_000002.11:g.190649316G>C	Forward	GTAATTGCCTGCCTCGCGCTAGCAG**[G>C]**AAGGTAGTGTGGTGTGACTAACGGG
rs1799977	NC_000003.11:g.37053568A>G	Forward	CTCAACCGTGGACAATATTCGCTCC**[A>G]**TCTTTGGAAATGCTGTTAGTCGGTA
rs6151792	NC_000005.9:g.80056961C>T	Forward	TCACACAGCCATGTAAAATTAGGCC**[C>T]**GCAGACAATTCGAAGGAGGAGAAAA
rs115109737	NC_000005.9:g.80102444G>A	Forward	GAATCACACAAGCTTATTTGCTATA**[G>A]**CATTATAATAACTTTTTACATCTGT
rs71636247	NC_000005.9:g.80118976A>G	Forward	TGTATAAATATATGTGGAGAAAACC**[A>G]**TCTAGATAGAAGGCTTATTCCAAAA
rs1805323	NC_000007.13:g.6026942G>T	Reverse	TCCAGTCACGGACCCAGTGACCCTA**[C>A]**GGACAGAGCGGAGGTGGAGAAGGAC
rs12531179	NC_000007.13:g.6028687C>T	Forward	ATTTTTAGTAGAGACAGAGTTTCAC**[C>T]**GTGTTAGATAGTCTCGATCTCCTGA
rs3735721	NC_000008.10:g.103217695A>G	Forward	GCTGGGGCCAGCTTAGTTGTAAGAA**[A>G]**AACTATTATTGTATATAATTGGACA
rs1037700	NC_000008.10:g.103250775G>C	Reverse	GGCCTCAGGCCGGGGTGAGACTTAC**[C>G]**CCTGCGTTTATCCGCCTCACGCTCT
rs5893603	NC_000008.10:g.103250839_103250840insG	Forward	TTGGCTGGCCCCGGGGCAGAGCAGC**[->G]**GAGCGGGACGCAAACCCAAAGTCAG
rs1037699	NC_000008.10:g.103250930C>T	Reverse	AGGACAGGCCTGTCCGCCCGCCCTC**[G>A]**CCGCAGCCTGGCTTCGTCGTTGCGA
rs16869352	NC_000008.10:g.103306033T>C	Forward	CAGCGTAAGGTAGCAATGCTTGGAA**[T>C]**ACACGCTTGCATTTTCCAATTGGCT
rs61752302	NC_000008.10:g.103311153C>T	Forward	ACAATTTCAATATAAAATGAGCATT**[C>T]**GCCTTTCGATCCTTGGATTCTACTA
rs72734283	NC_000014.8:g.75495059A>G	Forward	ATTATTTTATGATTTGACCTTGACA**[A>G]**CCCATCTAGCCAACTCCCATCCAGT

rs175080	NC_000014.8:g.75513828G>A	Forward	GGTCATAGGACTTTCTCTCAAACTA**[G>A]**GCATCTGTTGTTCTAAACAATCTTC
rs146353869	NC_000015.9:g.31126401C>A	Forward	AATGGTATGTATTAAAATGTGAATC**[C>A]**CAAGAGTGATGTGTCACTGTGCACT
rs114136100	NC_000015.9:g.31197976C>T	Forward	GCTGCAATGGTCCTGGTCAAACAAC**[C>T]**GGTCATCCTTACTACCTTCGGAGTT
rs150393409	NC_000015.9:g.31202961G>A	Forward	GCCTTTCTCAAATTGGCCAAACAGC**[G>A]**TTCAGTCTGCACTTGGGGCAAGAAT
rs3512	NC_000015.9:g.31235005G>C	Reverse	ACAGAGAGCGTTAAAAGTAAAGGCA**[C>G]**TTCCAAGAGTAACACTGCTAATGCG
rs20579	NC_000019.9:g.48668830G>A	Reverse	GCTGGACAGGAAGGGAGAATTCTGA**[C>T]**GCCAACATGCAGCGAAGTATCATGT

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.

Disease	Sample N	A	B	P
HD	445	6.119939	-0.052966	<2e-16
SCA1	177	5.682974	-0.043694	<2e-16
SCA2	294	5.799343	-0.056682	<2e-16
SCA3	397	7.137211	-0.049477	<2e-16
SCA6	69	5.96740	-0.08686	0.00268
SCA7	73	4.643231	-0.026023	2.94e-5
SCA17	7	2.38659	0.01716	0.70

Table 4.5*: Effects of repeat length of the expanded allele on the age at onset.*
Results of fitting a linear regression ln(AAO) = A + B(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 X^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 ($p=1.43x10^{-5}$), 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.

Disease Group	**GeM-HD concordance?**	**P (All SNPs)**	**P (High LD SNPs removed)**	**P (rs3512 removed)**
ALL (HD+SCAs)	non directional	$4.74x10^{-4}$	$2.26x10^{-4}$	0.00492
	Same as GeM-HD	$1.43x10^{-5}$	$6.98x10^{-6}$	$2.26x10^{-4}$
HD	non directional	0.0226	0.00775	0.0364
	Same as GeM-HD	0.00194	$4.63x10^{-4}$	0.00394
SCAs	non directional	0.0188	0.0236	0.0771
	Same as GeM-HD	0.00107	0.00142	0.00667
SCA1	non directional	0.376	0.386	0.444
	Same as GeM-HD	0.416	0.287	0.524
SCA2	non directional	0.0230	0.0629	0.0233
	Same as GeM-HD	0.00350	0.0138	0.00442
SCA3	non directional	0.176	0.114	0.355
	Same as GeM-HD	0.0809	0.0381	0.205

SCA6	non directional	0.00588	0.0735	0.00506
	Same as GeM-HD	0.00162	0.0340	0.00163
SCA7	non directional	0.155	0.217	0.297
	Same as GeM-HD	0.0447	0.0885	0.113

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 (r2>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 ($p=5.28x10^{-13}$, 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.

SNP	Chr	Pos	A1 (GeM-HD)	A2 (GeM-HD)	MAF (GeM-HD)	Beta (GeM-HD)	P (GeM-HD)	A1(All)	A2(All)	MAF(All)	Beta(All)	P(All)	Beta(HD)	p(HD)	Beta(SCA1)	P(SCA1)	Beta(SCA2)	P(SCA2)	Beta(SCA3)	P(SCA3)	Beta(SCA6)	P(SCA6)	Beta(SCA7)	P(SCA7)	Beta(AllSCA)	P(AllSCA)
rs1800937	2	48025764	T	C	0.092	0.820	4.30E-03	T	C	0.074	0.490	4.75E-01	0.520	6.21E-01	-0.571	6.51E-01	-0.459	8.18E-01	2.455	4.47E-02	0.614	8.25E-01	-10.050	5.34E-01	0.438	6.13E-01
rs4150407	2	128049631	C	T	0.444	0.575	4.60E-04	G	A	0.479	0.064	8.50E-01	-0.585	2.53E-01	-0.574	3.91E-01	1.384	1.03E-01	-0.013	9.85E-01	-2.129	2.55E-01	-2.702	3.83E-01	0.260	5.48E-01
rs5742933	2	190649316	C	G	0.206	-0.699	9.49E-04	C	G	0.205	-0.725	9.59E-02	-0.732	2.49E-01	1.102	2.19E-01	-2.333	3.69E-02	-1.005	2.19E-01	0.939	6.76E-01	0.551	8.77E-01	-0.714	2.03E-01
rs1799977	3	37053568	G	A	0.319	0.847	7.16E-07	G	A	0.280	-0.359	3.55E-01	0.531	3.39E-01	-0.241	7.58E-01	-2.555	2.20E-02	1.081	1.31E-01	-1.424	4.17E-01	-5.899	1.44E-01	-0.698	1.68E-01
rs6151792	5	80056961	T	C	0.099	-1.049	2.09E-04	T	C	0.117	-0.662	2.16E-01	-1.395	7.99E-02	-0.436	7.30E-01	0.495	6.79E-01	-1.347	2.21E-01	-1.577	5.07E-01	-0.116	9.77E-01	-0.350	6.09E-01
rs115109737	5	80102444	A	G	0.060	-1.289	4.50E-04	A	G	0.041	-2.095	1.81E-02	-3.014	2.10E-02	1.321	5.13E-01	-4.651	8.59E-02	0.100	9.50E-01	-5.197	1.41E-01	-5.756	2.87E-01	-1.726	1.28E-01
rs71636247	5	80118976	G	A	0.054	-1.398	2.55E-04	G	A	0.034	-2.208	2.63E-02	-1.917	1.89E-01	-1.974	4.39E-01	-6.400	4.86E-02	0.324	8.52E-01	-3.813	2.82E-01	-7.123	2.49E-01	-2.329	6.76E-02
rs1805323	7	6026942	T	G	0.038	-0.950	3.04E-02	A	C	0.043	-3.605	**3.14E-05**	-3.890	**3.14E-04**	-5.677	**1.67E-03**	-1.835	3.94E-01	-2.307	2.70E-01	-2.123	5.50E-01	-17.190	1.44E-01	-3.305	6.62E-03
rs12531179	7	6028687	T	C	0.147	0.938	3.84E-05	T	C	0.169	0.579	2.16E-01	1.070	1.23E-01	0.039	9.67E-01	1.137	3.08E-01	0.083	9.32E-01	-0.320	8.83E-01	-0.798	8.07E-01	0.367	5.39E-01
rs3735721	8	103217695	G	A	0.085	-1.529	5.68E-07	G	A	0.083	-0.389	5.25E-01	0.354	6.56E-01	0.692	5.13E-01	-3.278	6.32E-02	1.308	3.11E-01	-15.150	2.35E-03	-3.035	5.89E-01	-0.790	3.47E-01
rs1037700	8	103250775	C	G	0.097	-1.541	5.03E-08	G	C	0.094	-0.817	1.54E-01	-0.012	9.87E-01	1.046	2.46E-01	-4.132	2.11E-02	0.863	4.72E-01	-14.250	**5.47E-04**	-8.021	1.55E-01	-1.235	1.11E-01
rs5893603	8	103250839	G	-	0.097	-1.548	4.28E-08	G	-	0.093	-0.983	8.89E-02	-0.092	9.05E-01	0.914	3.13E-01	-4.189	1.84E-02	0.537	6.59E-01	-11.770	**2.13E-03**	-9.077	1.24E-01	-1.441	6.45E-02
rs1037699	8	103250930	T	C	0.096	-1.570	2.70E-08	A	G	0.094	-0.819	1.53E-01	-0.006	9.94E-01	0.758	4.13E-01	-3.519	3.97E-02	0.896	4.55E-01	-14.260	**4.86E-04**	-9.077	1.24E-01	-1.228	1.11E-01
rs16869352	8	103306033	C	T	0.083	-1.528	4.01E-07	C	T	0.080	-0.464	4.57E-01	0.691	3.98E-01	0.756	4.36E-01	-2.854	1.25E-01	0.681	6.27E-01	-10.850	3.24E-02	-7.745	1.64E-01	-1.067	2.09E-01
rs61752302	8	103311153	T	C	0.023	-1.671	3.03E-03	T	C	0.026	-0.150	8.92E-01	-0.520	7.46E-01	0.567	7.10E-01	-1.045	6.76E-01	4.882	1.18E-01	-8.015	1.69E-01	NA	NA	0.019	9.89E-01
rs72734283	14	75495059	G	A	0.099	0.858	4.32E-03	G	A	0.089	0.898	1.40E-01	2.057	1.14E-02	1.585	1.82E-01	-1.099	5.41E-01	-0.650	5.88E-01	-1.686	5.59E-01	10.770	3.82E-02	0.318	6.98E-01
rs175080	14	75513828	A	G	0.466	-0.434	7.72E-03	A	G	0.435	-0.671	5.66E-02	-1.245	1.61E-02	0.279	7.16E-01	-0.090	9.23E-01	0.397	5.62E-01	-0.927	5.84E-01	-4.356	1.66E-01	-0.405	3.70E-01
rs146353869	15	31126401	A	C	0.017	-6.107	4.30E-20	A	C	0.017	-2.362	8.17E-02	-1.804	3.28E-01	1.980	5.64E-01	-8.999	3.81E-02	-1.537	4.94E-01	-3.496	5.52E-01	7.338	6.60E-01	-2.610	1.48E-01
rs114136100	15	31197976	T	C	0.018	-5.073	8.49E-16	T	C	0.019	-2.101	9.20E-02	-1.188	4.88E-01	1.609	6.00E-01	-1.168	7.89E-01	-3.519	8.25E-02	-3.464	5.55E-01	6.909	6.73E-01	-2.521	1.27E-01
rs150393409	15	31202961	A	G	0.016	-5.765	9.34E-18	A	G	0.013	-2.735	7.03E-02	-2.909	1.39E-01	-0.354	9.28E-01	-4.224	4.88E-01	-3.176	1.92E-01	-0.912	8.99E-01	7.443	6.57E-01	-2.551	2.17E-01
rs3512	15	31235005	C	G	0.309	1.325	5.28E-13	G	C	0.283	1.680	**1.52E-05**	1.297	2.94E-02	1.388	8.70E-02	1.020	3.03E-01	2.156	2.36E-03	0.886	6.37E-01	9.647	5.00E-03	1.809	**2.22E-04**
rs20579	19	48668830	A	G	0.124	0.769	6.65E-03	T	C	0.134	0.427	4.09E-01	0.119	8.82E-01	1.244	2.84E-01	0.412	7.55E-01	1.099	2.17E-01	-7.791	2.19E-02	-0.216	9.54E-01	0.515	4.28E-01

***Table 4.7**: Single SNP associations.*

*Beta denotes the effect size – that is, the number of years added to or subtracted from the expected age at onset for each copy of the minor allele (A1). MAF 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 Bonferoni correction for 22 SNPs; those highlighted red satisfy Bonferoni 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.

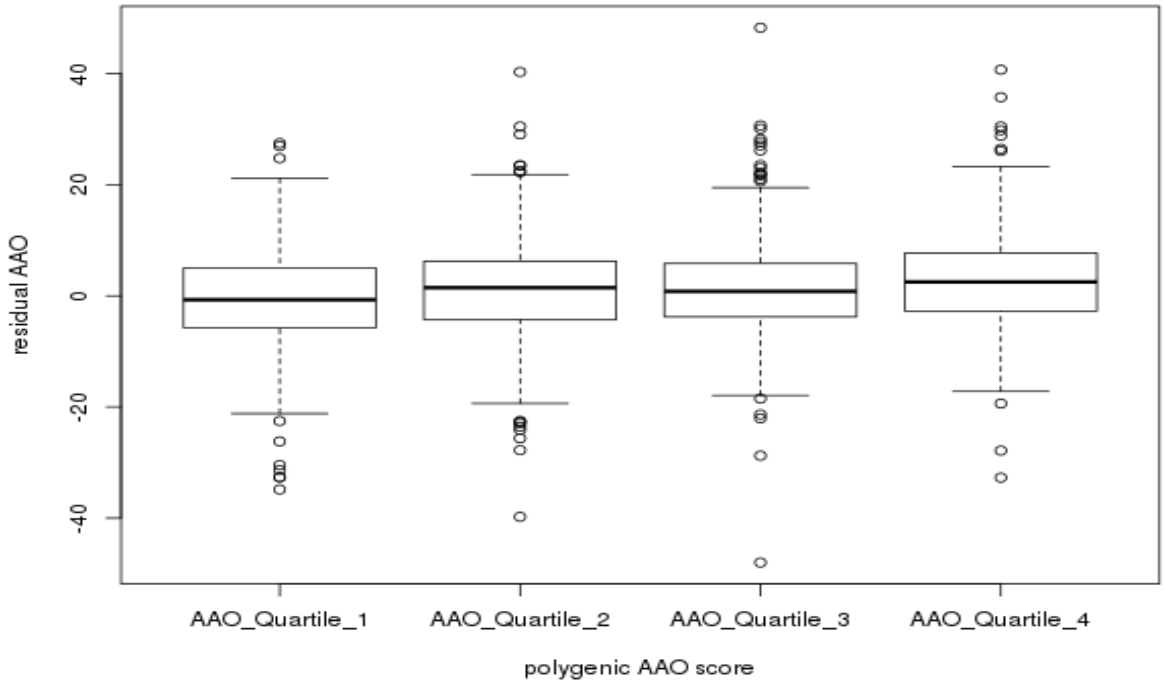


Figure 4.1: *Boxplot of residual AAO (across all samples) by quartiles of polygenic age at onset score. Polygenic score calculated by summing the number of minor alleles (weighted by their effect on age at onset in the GeM GWAS) across the 22 SNPs. Note that lower scores correspond to earlier than expected AAO, and thus smaller residuals. Figure devised and produced by Professor Peter Holmans for publication in (Bettencourt et al., 2016).*

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.

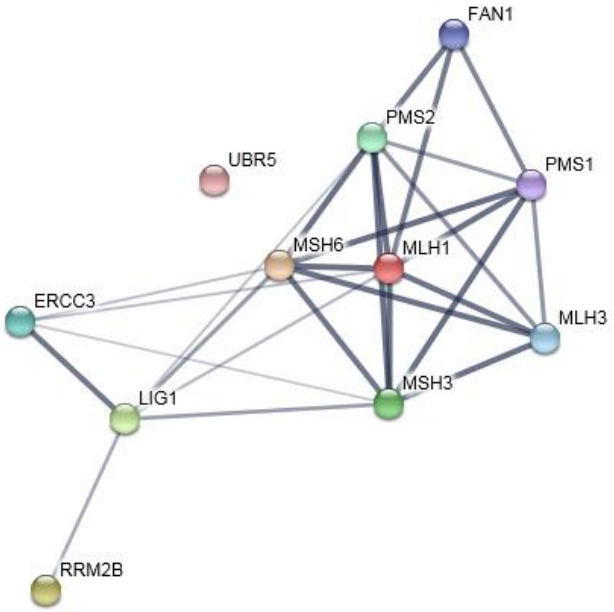


***Figure 4.2**: String diagram illustrating the functional connection between the proteins included in this study. Nodes represent proteins while edges represent protein-protein interactions and are the intensity of the lines reflect low (0.150), medium (0.400), high (0.700) and highest (0.900) confidence). 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**).

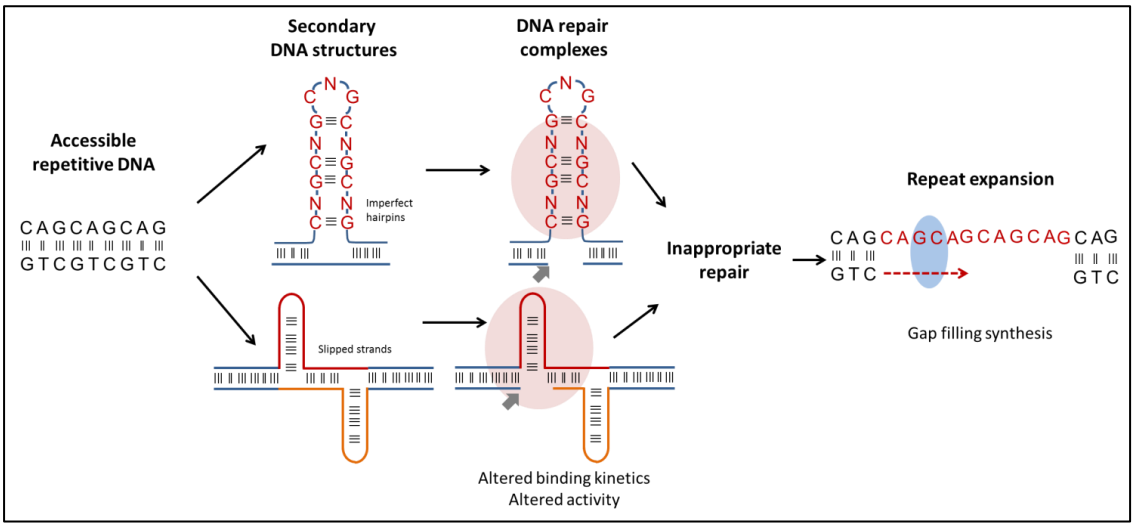


***Figure 4.3**: Potential mechanism by which variants in DNA repair could influence somatic expansion of CAG repeats. Hypothesised mechanism of somatic expansion of the CAG repeats in polyglutamine diseases due to variation in genes encoding DNA repair proteins. The accessibility of repetitive DNA sequences 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.

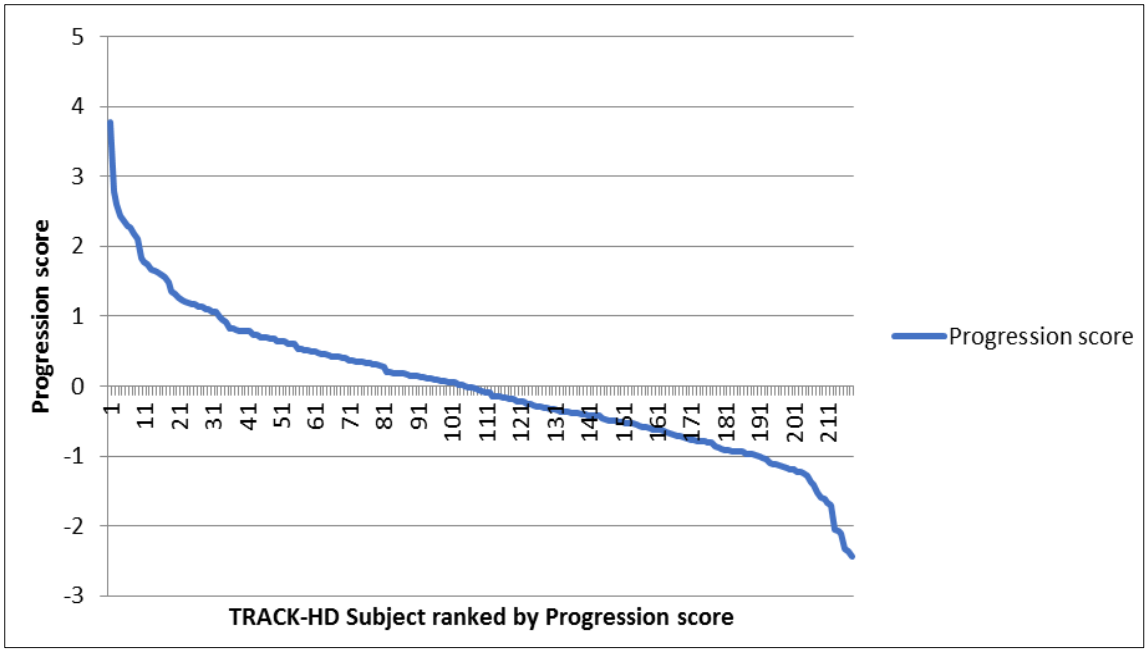


***Figure 5.1**: Distribution of the Progression scores in the TRACK-HD cohort*, showing that there were more fast progressors than slow progressors who were clearly atypical.

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 re-calibrated 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^{0C}.

- Left primer: TTGCCCTGCCATGTCTCG
- Right primer: TCCCACCTTCCCCTTCTTCA

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) 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. 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 ddH20, 0.75µl sequencing primer (at 5µM concentration) and 1µl of PCR product. They were run with the following cycling conditions 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. 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.

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

KEGG 53	24	0.0214	0.06268	KEGG ASCORBATE AND ALDARATE METABOLISM
GO: 5954	4	0.0228	0.06342	calcium- and calmodulin-dependent protein kinase complex
GO: 5578	320	0.0244	0.02156	proteinaceous extracellular matrix
GO: 32138	**4**	**0.0248**	**0.00154**	**single base insertion or deletion binding**
GO: 217	**11**	**0.0278**	**0.01587**	**DNA secondary structure binding**
GO: 51297	43	0.0308	0.07898	centrosome organization
GO: 10595	40	0.0314	0.07721	positive regulation of endothelial cell migration
GO: 30198	274	0.0334	0.06379	extracellular matrix organization
GO: 33013	54	0.0366	0.01982	tetrapyrrole metabolic process
REACTOME 580	77	0.0370	0.00302	REACT: INTEGRIN CELL SURFACE INTERACTIONS
GO: 43062	275	0.0378	0.06438	extracellular structure organization
GO: 403	**5**	**0.0402**	**0.05498**	**Splayed Y-form DNA binding**
GO: 32389	**4**	**0.0418**	**0.00010**	**MutLalpha complex**
GO: 51153	66	0.0426	0.07427	regulation of striated muscle cell differentiation
REACTOME 387	142	0.0456	0.09237	REACT: EXTRACELLULAR MATRIX ORGANIZATION
GO: 51096	**4**	**0.0466**	**0.02983**	**positive regulation of helicase activity**
GO: 6281	**308**	**0.0488**	**0.09386**	**DNA repair**
GO: 43566	**166**	**0.0492**	**0.04560**	**structure-specific DNA binding**
GO: 51567	10	0.0500	0.06351	histone H3-K9 methylation

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**).

Gene/ pseudogene	Number of variants
GOLGA8J	0
GOLGA8T	0
KFZP434L187	0
CHRFAM7A	0
GOLGA8R	0
GOLGA8H	0
ARHGAP11B	0
OC100288637	0
HERC2P10	0
FAN1	5
MTMR10	1
MIR211	0
TRPM1	2
RP11-16E12.2	0
KLF13	0
OTUD7A	0
CHRNA7	0
GOLGA8K	0
ULK4P3	0
ULK4P1	0
ULK4P2	0
GOLGA8O	0
ARHGAP11A	0
SCG5	0

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.

Group	**Variant**			**Progression rank of subject with variant**	**AAO residual** (onset minus expected onset)	**Protein domain**	**MAF**	**p in GeM GWAS**	**Prediction of functional effects**		
	Amino acid	cDNA	SNP ID						**SIFT** (Kumar et al., 2009).	Polyphen2 (Adzhubei et al., 2010)	Mutation Assessor (Reva et al., 2007)
Fast	p.R145H	c.G434A	rs146408181	9	-10.05	UBZ-SAP	0.0002	>1E-5 or not present	Tolerated	Benign	Neutral
Fast	p.E240K	c.G718A	rs150748572	5	-10.22	UBZ-SAP	0.0012	>1E-5 or not present	Tolerated	Benign	Low
Fast	p.R507H	c.G1520A	rs150393409	10	-7.22	SAP	0.0028	9.339E-18	Damaging	Possibly damaging	Low
Fast	p.Q829H	c.G2487C	Novel	7	-9.34	TPR-VRR	-	>1E-5 or not present	Damaging	Possibly damaging	
Fast	p.F762F	c.C2286T	rs200756403	16	No onset	TPR-VRR	0.0002	>1E-5 or not present	Tolerated	-	
Slow (n=2)	p.P894S	c.C2680T	rs80120912	170, 209	+1.5	VRR	0.008	>1E-5	Tolerated	Benign	

								or not present			
Slow	p.R377W	c.C1129T	rs151322829	204	No onset	UBZ-SAP	0.0014	>1E-5 or not present	Damaging	Damaging	Medium

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 = 1.06×10^{-4} in this study (Hensman Moss et al., 2017b).

	Fast progressors	**Slow progressors**
Frequency of G allele	46	27
Frequency of A allele	4	19

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**).

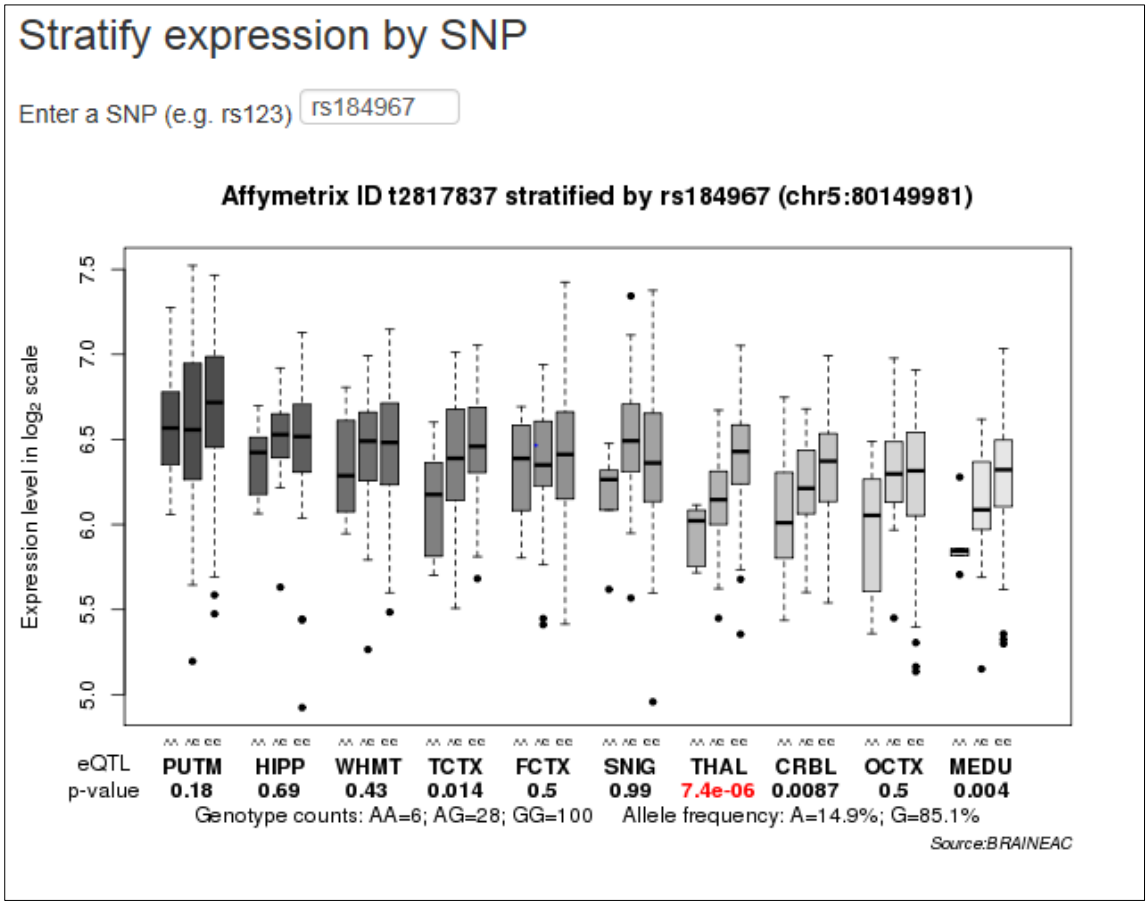


***Figure 5.2**: Influence of rs184967 allele status on brain expression of MSH3.* eQTL analysis of rs184967 using Braineac database (UKBEC, 2015). Genotype (AA/AC/CC) shown on x-axis; expression levels in log2 scale shown on y axis.

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.

	Fast progressors	Slow progressors
GCAGCGGCTGCAGCGGCC	20	15
- (deletion)	2	19

***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 r^2=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.

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

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**).

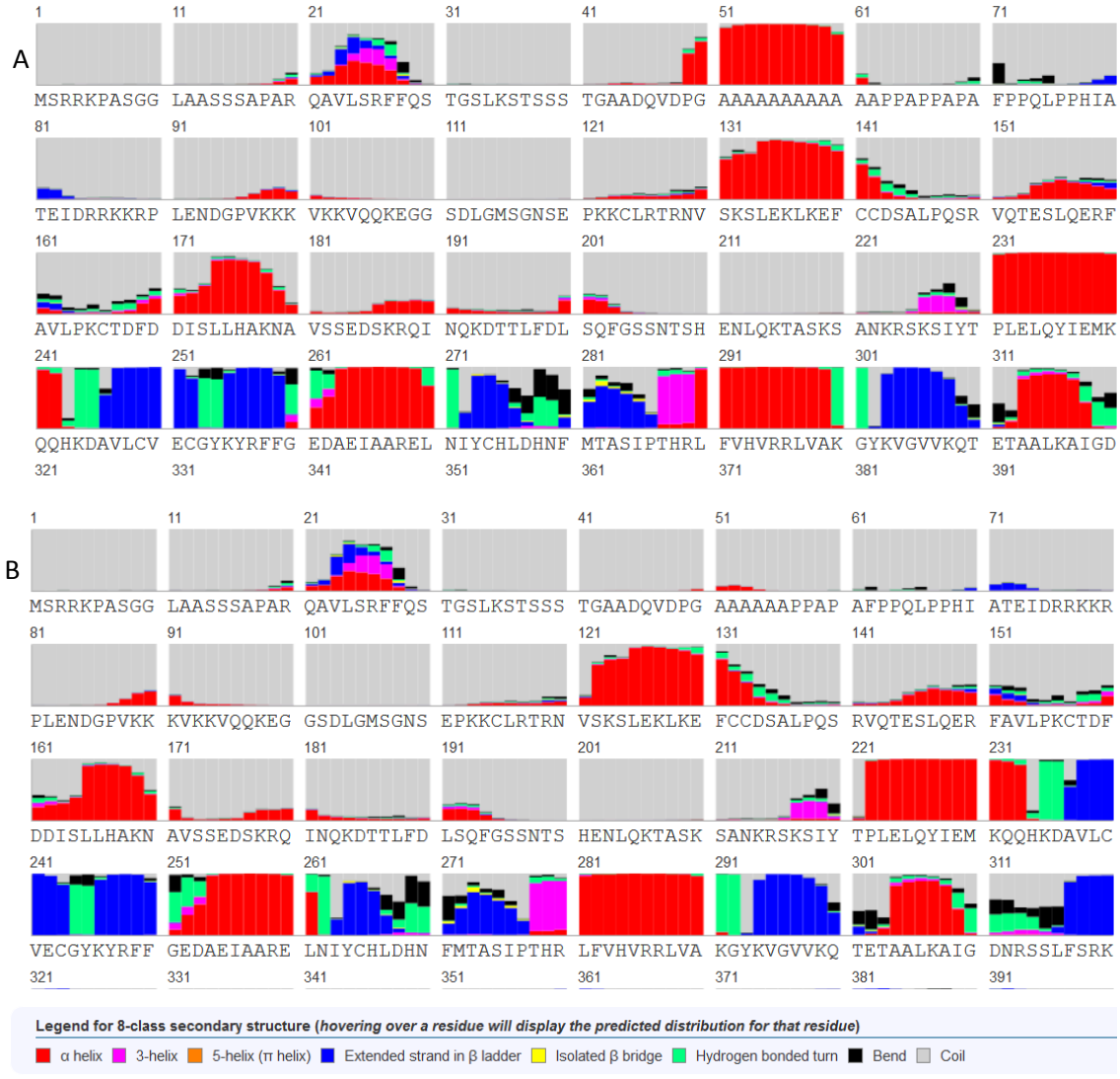


***Figure 5.3**: Secondary structure predictions for MSH3 in the wild type form (**A**) and with the deletion (**B**) which is seen in the slow progressing subjects who drive the progression GWAS*

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.

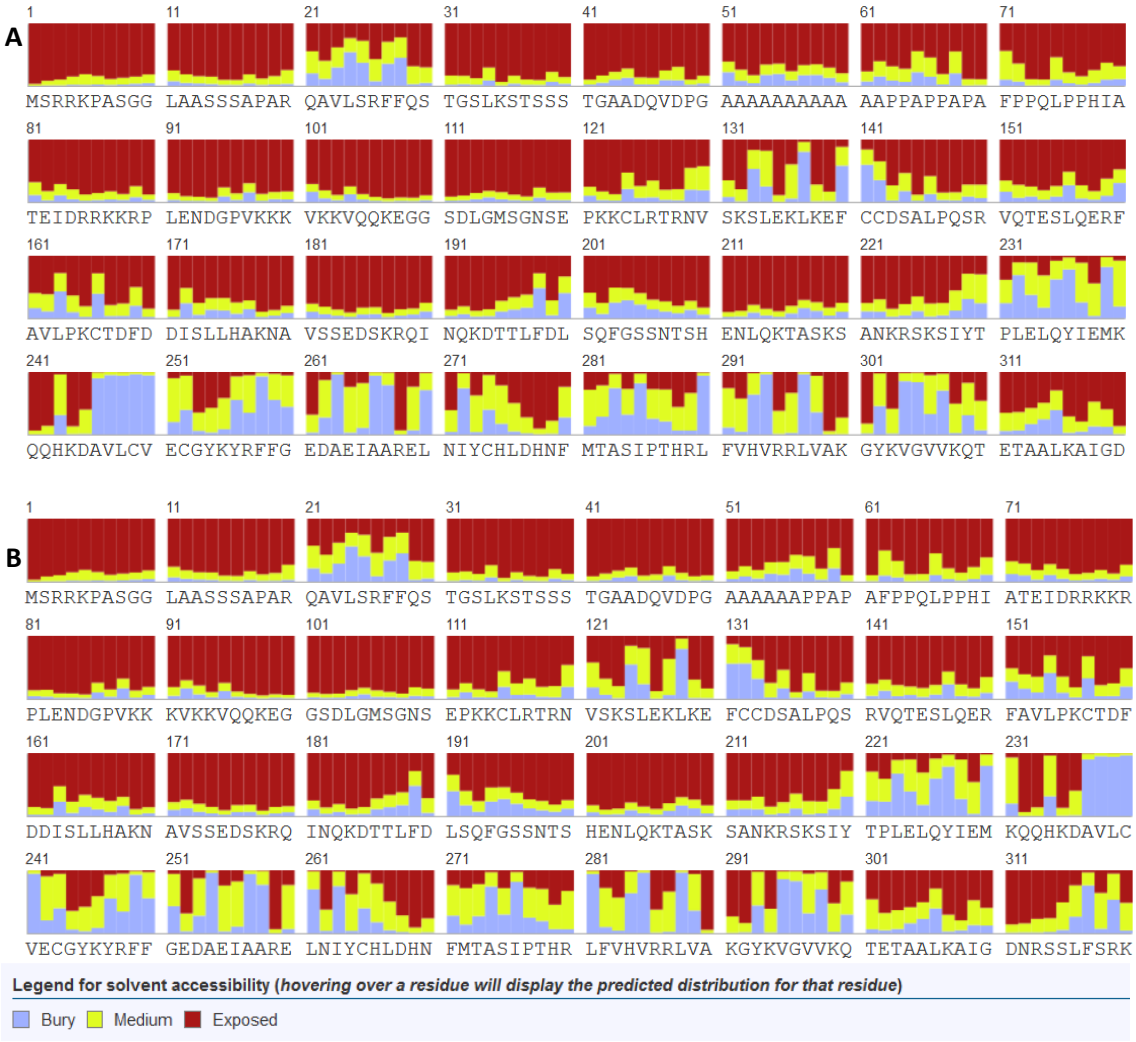


***Figure 5.4**: Predicted solvent exposure for MSH3 wild type (**A**) and with deletion (**B**) using Raptor (Källberg et al., 2012) (accessed 22/06/17) showing that the region containing the deletion is predicted to be exposed both with and without the deletion present. Only residues 1 – 320 are shown 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**.

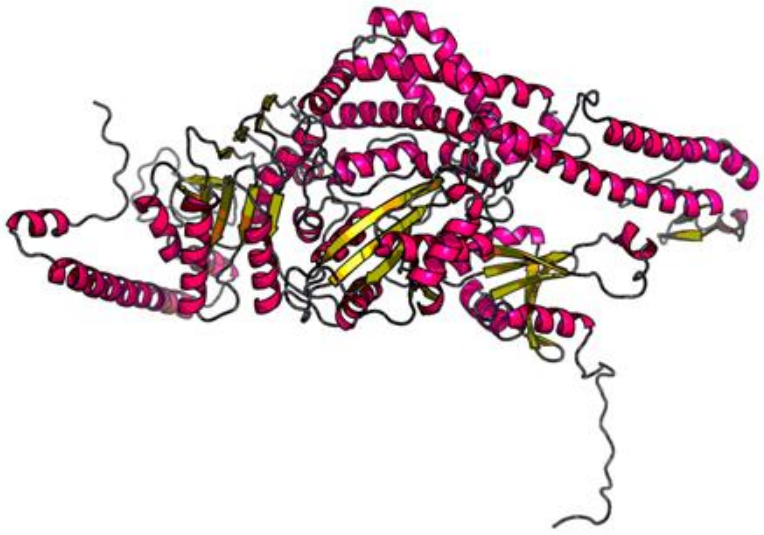


Figure 5.5*: Tertiary structure predictions of MSH3 from residue 211 to 1128 generated by Raptorx (Källberg et al., 2012) accessed 22/06/17 .*

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**.

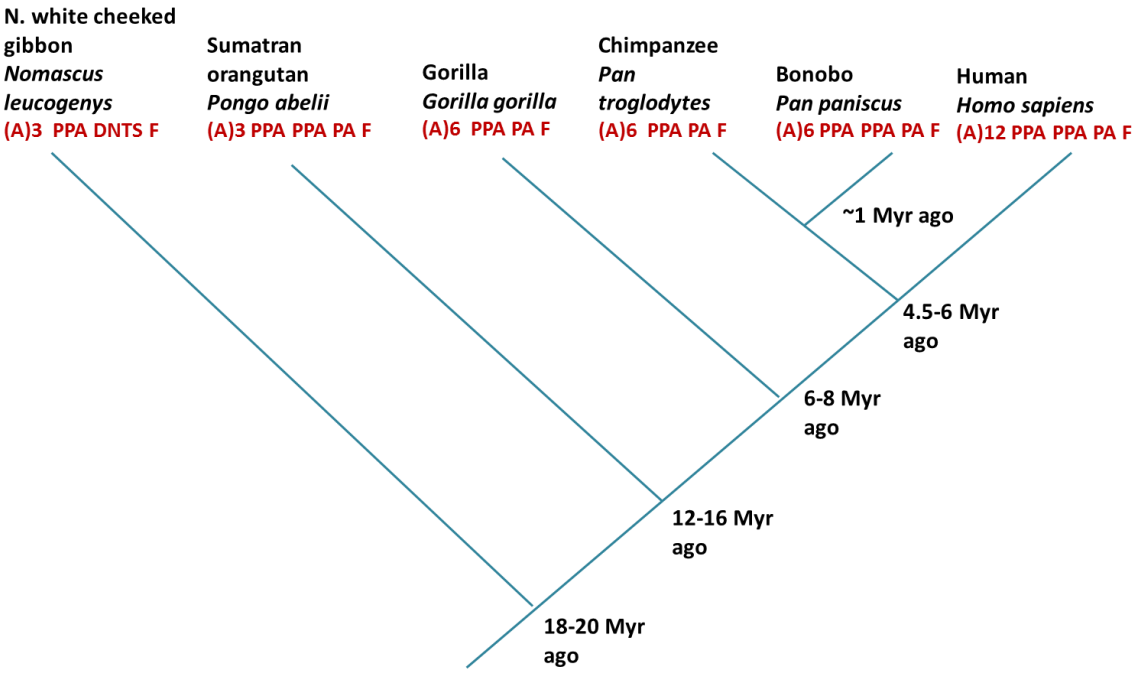


Figure 5.6*: Cladogram of the apes showing the MSH3 protein sequence at the repetitive region of interest in six different ape species, showing that the presence of 12 alanines is novel to the human lineage. Gibbons, lower apes which diverged earlier from the lineage have just three consecutive alanines 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**).

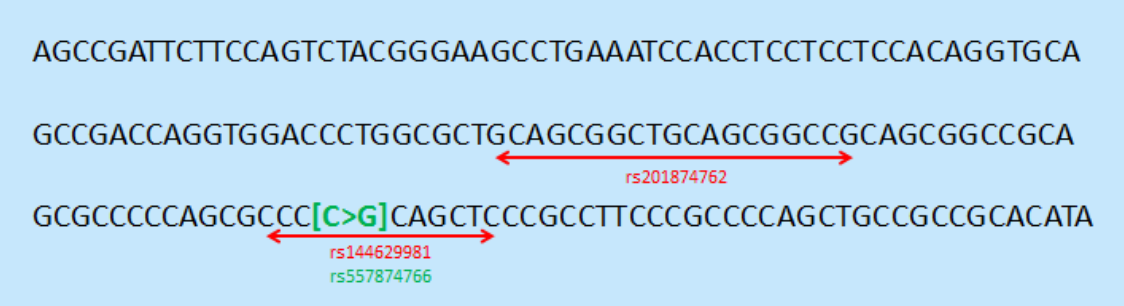


***Figure 5.7**: Excerpt of the MSH3 exon 1 sequence showing the positions of the rs557874766, rs201874762 and rs144629981 variants*.

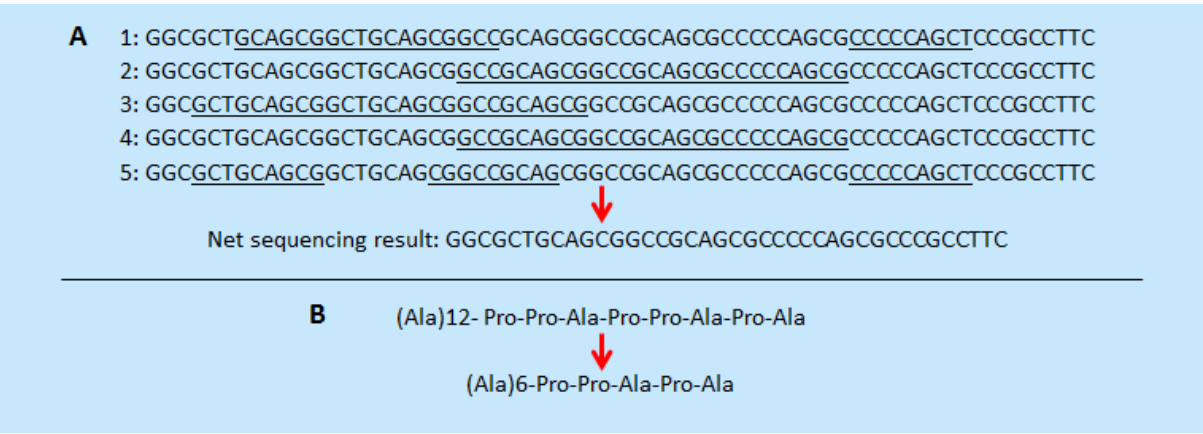


***Figure 5.8**: Alternative MSH3 deletions achieve the same protein sequence result. **A**: Short excerpt of the MSH3 exon 1 coding sequence showing alternative deletions which generate the same net sequence at the DNA sequence level, the option shown in Figure 5.3 above is listed as option 1. **B**: Short excerpt of the amino acid sequence that results from the deletions shown in (A): p.A57_p.65del. Ala: alanine; Pro: proline.*

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).

Repeats	**Sizes (bp)**	**N (total 116)**	**Frequencies**
3	171	18	0.155
4	180	5	0.043
5	189	1	0.009
6	198	70	0.603
7	207	22	0.190

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**).

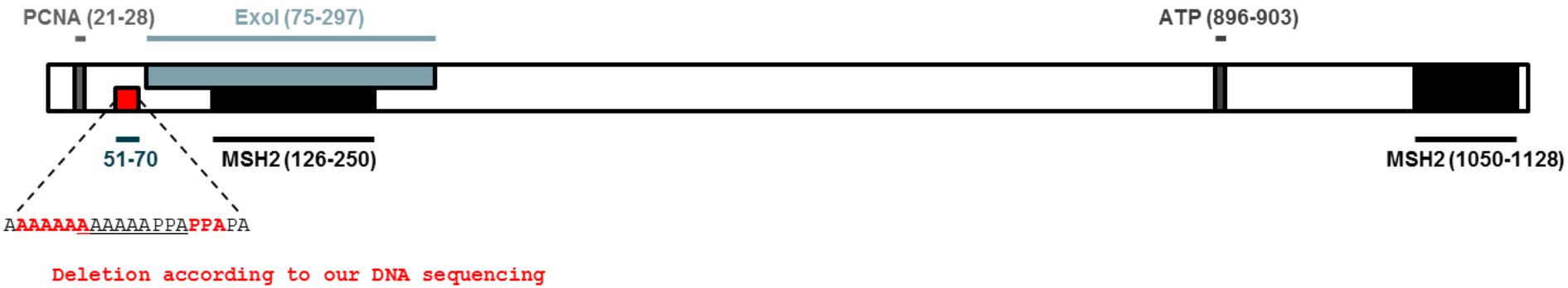


Figure 5.9*: Figure of MSH3 showing the putative binding domains for proteins with which it interacts, and the ATP binding site. The location on the deletion we identified is also shown, illustrating that the Exo1 and PCNA binding sites are very close to the deletion.*

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's 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 Spinocerebellar 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.

Score	Description of family history structure

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

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^{O}C to 56 ^{O}C in 2 ^{O}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 ^{o}C, 30 s annealing at 60 ^{o}C, and 1 min elongation at 72^{o}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 $(GGGGCC)_5$ 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 $\log_{10}$ 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).

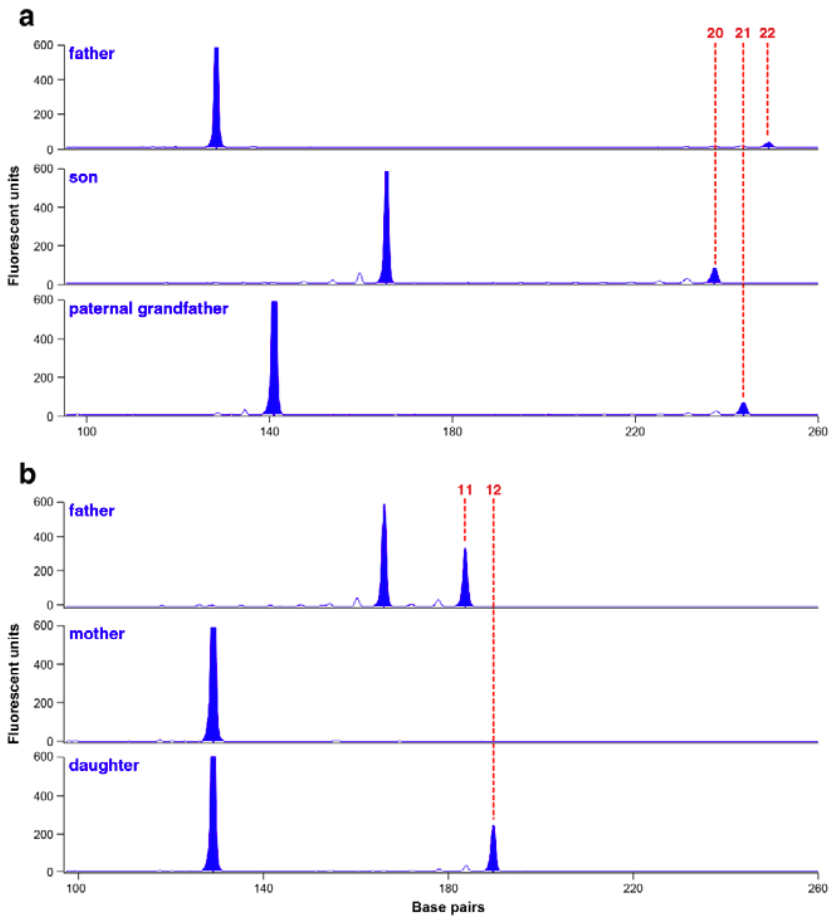


***Figure 6.1**: Fragment analysis of CEPH families with inter-generational repeat slippage. Data from fluorescent labelled PCR followed by fragment length analysis on an ABI 3730xl automated sequencer from 2 CEPH families showing evidence of inter-generational repeat slippage. For clarity, the numbers 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.

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

***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.

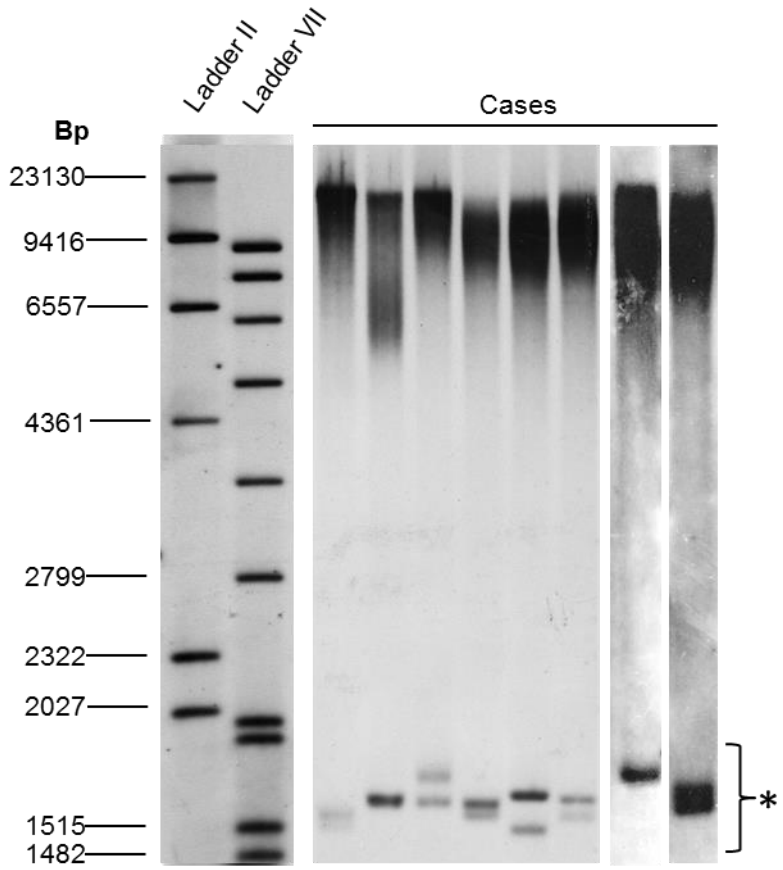


***Figure 6.2**: Southern Blot of eight HD phenocopy patient DNAs, blot produced by Mark Poulter. Southern Blot of eight HD phenocopy patient DNAs, showing that C9orf72 repeat expansions can be seen in all cases. The asterisk indicates a GGGGCC containing a short-tandem-repeat genome motif unrelated 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 *C9or72* 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).

	Case									
Clinical feature	1	2	3	4	5	6	7	8	9	10
Chorea					√		√		√	
Myoclonus	√			√	√		√			
Dystonia				√	√		√		√	
Tremor					√			√		√
Rigidity			√	√	√		√	√		√
Bradykinesia				√	√		√	√		√
Torticollis				√	√			√		
UMN signs				√	√			√		√
Depression			√	√	√			√		
Anxiety	√	√								
Apathy				√	√					
Executive dysfunction	√	√	√	√				√		√
Impaired memory		√	√	√	√		√			√
Impaired face recognition		√	√				√			
Impaired verbal fluency	√				√					√

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.

	Number in C9orf72 negative cases (N=504) (Percentage)	**Number in C9orf72 positive cases** (N=10) (Percentage)	**Number in whole HD phenocopy cohort** (N=514) (Percentage)	**P value** (Fisher's exact test)
All movement disorder features	394 (78%)	8 (80%)	402 (78%)	1
Chorea	154 (31%)	3 (30%)	157 (31%)	1
Dystonia	53 (11%)	4 (40%)	57 (11.1%)	0.017
Bradykinesia/ rigidity	78 (15%)	6 (60%)	84 (16%)	0.002
Tremor	39 (8%)	3 (30%)	42 (8%)	0.041
Ataxia	72 (14%)	1 (10%)	73 (14%)	1
Myoclonus	31 (6%)	4 (40%)	35 (7%)	0.003
UMN features	18 (4%)	4 (40%)	24 (5%)	<0.001
LMN features	8 (1.6%)	0 (0%)	8 (2%)	1
Psychiatric problems	53 (11%)	7 (70%)	60 (12%)	<0.001
Depression	17 (3%)	4 (40%)	21 (4%)	0.035
Anxiety	4 (0.8%)	2 (20%)	6 (1%)	0.005
Cognitive impairment	167 (33%)	9 (90%)	176 (34%)	<0.001
Executive dysfunction	19 (4%)	6 (60%)	25 (5%)	<0.001
Memory problems	29 (6%)	9 (90%)	176 (34%)	<0.001
Family history	98 (19%)	7 (70%)	105 (20%)	0.001

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).

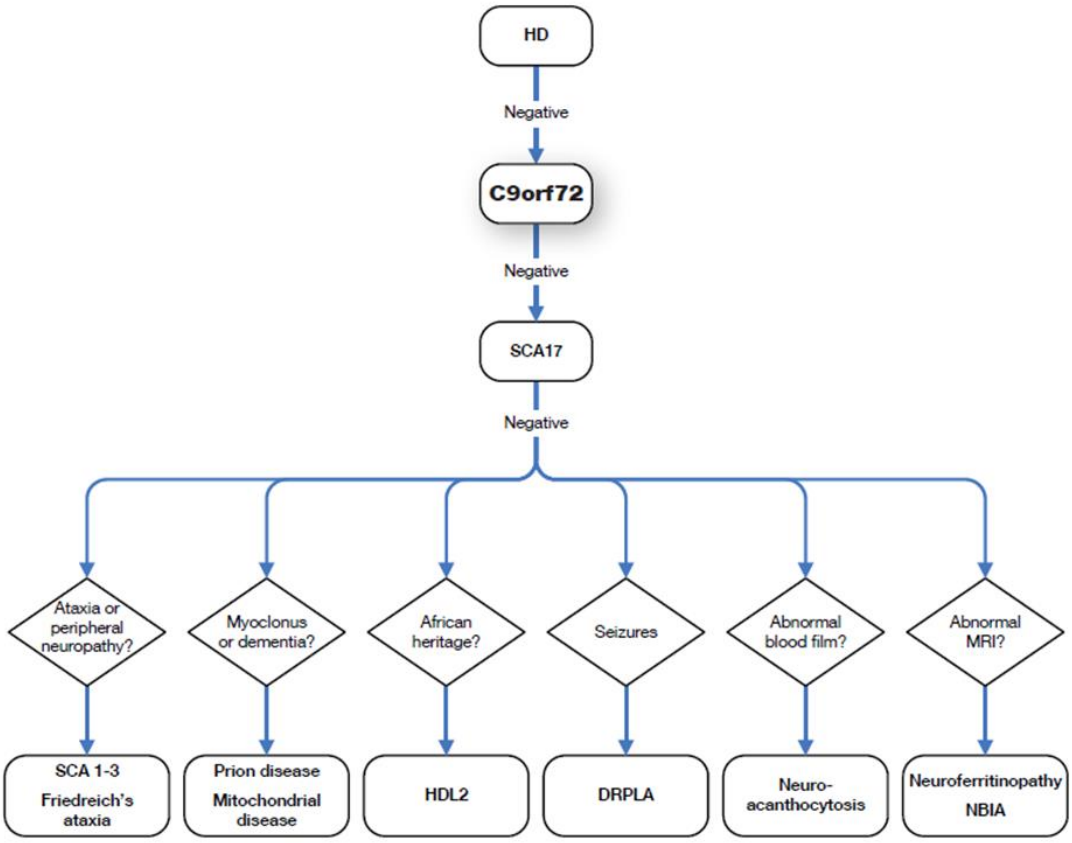


Figure 6.3*: Algorithm for the investigation of HD phenocopy cases. Proposed clinico-genetic algorithm for the work-up of Huntington's disease phenocopy patients, highlighting key diagnoses to be considered. SCA, Spinocerebellar ataxia; HDL2, Huntington's disease-like 2, DRPLA, dentatorubral-pallidoluysian 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 *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 *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 *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 *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 *C9orf72* expansion given that the population frequency of *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 *C9orf72* mutations.

Among the ten HD phenocopy *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 *C9orf72* expansion with these motor symptoms manifests with earlier onset.

Incomplete penetrance has been previously suggested in *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 *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 *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 Scientic 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.

Cohort	Group	N	Mean age, y ± SD (range)	Gender (male/ female)	Mean (CAG)n length ± SD (range)	Mean TMS ± SD (range)	Mean TFC ± (range)	Rate of progression (n)		
								Fast	Average	Slow
TRACK-HD	HD	112	46 ± 10 (22-64)	50/62	44 ± 3 (39-59)	14 ± 13 (0-45)	12 ± 2 (7-13)			
	Premanifest	50	42 ± 9 (22-64)	24/26	43 ± 3 (39-52)	2 ± 2 (0-8)	13 ± 0 (12-13)	16	14	20
	Manifest	62	48 ± 10 (23-64)	26/36	44 ± 3 (39-59)	23 ± 11 (5-45)	11 ± 2 (7-13)	27	13	22
	Control	22	45 ± 5 (34-53)							
LUMC	HD	74	53 ± 11 (29-79)	34/40	44 ± 3 (39-53)	32 ± 31 (0-102)	8 ± 5 (0-13)			
	Premanifest	18	46 ± 10 (29-63)	5/13	42 ± 2 (39-47)	3 ± 2 (0-5)	12 ± 1 (10-13)			
	Manifest	56	55 ± 11 (35-79)	29/27	44 ± 3 (39-53)	42 ± 30 (6-102)	7 ± 5 (0-13)			
	Control	27	43 ± 11 (26-65)	13/14						
Combined	HD	186	48 ± 11 (22-79)	84/102	44 ± 3 (39-59)	21 ± 24 (0-102)	10 ± 4 (0-13)			
	Control	49	44 ± 9 (26-65)	22/27						

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^{o}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^{TM} 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^{TM} 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 - is the total number of genes

B - is the total number of genes associated with a specific GO term

n - is the number of genes in the top of the user's input list or in the target set when appropriate

b - is the number of genes in the intersection

Enrichment = (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:

1. A set of 124 HD brain expression modules derived by Neueder and Bates (2014), who applied weighted gene correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) to the Hodges et al. (2006) microarray brain expression data set of 44 human HD and 36 matched control brains. They generated networks for four brain

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**).

Gene ID	**Basic p-value**	**Adjusted p-value**	**z-score**	**Condition Premanifest**	**Average count: Manifest**	**Average count: Premanifest**
PHGDH	5.79E-06	0.1044	4.53381	0.4688	237.78	325.33
SPAG1	4.66E-04	1	-3.49935	-0.4093	422.76	320.67
GATSL1	6.10E-04	1	-3.42713	-2.0503	1.54	0.38
IGF1	1.19E-03	1	3.24199	1.2629	4.69	11.56
IGSF23	1.46E-03	1	-3.18292	-0.6685	20.53	13.20
OTOGL	1.53E-03	1	-3.16988	-4.4884	0.49	0.02
MIR29B2	1.57E-03	1	-3.16142	-0.3314	125.69	100.47
ANKUB1	1.69E-03	1	-3.13974	-0.5821	37.86	25.07
HIST1H3J	2.02E-03	1	3.08728	1.0005	2.08	4.02

ARMC9	2.58E-03	1	3.01333	0.5361	13.81	19.33
GP5	3.39E-03	1	2.92997	0.3355	140.15	174.27
HIST1H2AG	3.56E-03	1	2.91473	0.4295	20.10	26.93
TREM2	3.83E-03	1	-2.89183	-0.6085	9.71	6.36
RPGRIP1	3.86E-03	1	-2.88954	-0.2727	720.69	604.40
FOXJ1	4.37E-03	1	2.84985	0.5063	19.95	27.87
ADAMTS4	4.47E-03	1	2.84315	0.3578	33.71	42.89
C11orf93	4.86E-03	1	-2.81617	-0.7836	5.51	3.24

***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.*

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**.

Entrez gene ID	**Gene Symbol**	**p (diffexp)**	**q (diffexp)**	**log2(FC)**
722	C4BPA	7.81E-06	1.41E-01	1.371
2297	FOXD1	9.09E-05	7.02E-01	-0.785
3805	KIR2DL4	1.93E-04	7.02E-01	0.651
196394	AMN1	2.11E-04	7.02E-01	0.208
94137	RP1L1	2.47E-04	7.02E-01	-1.350
158248	TTC16	2.67E-04	7.02E-01	-0.347
100422824	MIR3128	2.86E-04	7.02E-01	0.930
5797	PTPRM	3.12E-04	7.02E-01	-0.359
84692	CCDC54	4.79E-04	9.58E-01	2.532
889	KRIT1	7.30E-04	9.58E-01	-0.081
54221	SNTG2	7.51E-04	9.58E-01	-0.689
22979	EFR3B	8.17E-04	9.58E-01	0.494
56934	CA10	8.42E-04	9.58E-01	2.036
8763	CD164	9.53E-04	9.58E-01	0.098
597	BCL2A1	1.06E-03	9.58E-01	0.423
4940	OAS3	1.12E-03	9.58E-01	0.688
49	ACR	1.13E-03	9.58E-01	1.237
9262	STK17B	1.19E-03	9.58E-01	0.132

54407	SLC38A2	1.19E-03	9.58E-01	0.124
285590	SH3PXD2B	1.26E-03	9.58E-01	0.414
60370	AVPI1	1.49E-03	9.58E-01	0.273
6425	SFRP5	1.49E-03	9.58E-01	0.634
387849	REP15	1.49E-03	9.58E-01	1.526
283726	FAM154B	1.53E-03	9.58E-01	0.748
143502	OR52I2	1.54E-03	9.58E-01	2.469
1999	ELF3	1.58E-03	9.58E-01	-0.343
54957	TXNL4B	1.65E-03	9.58E-01	0.088
23446	SLC44A1	1.65E-03	9.58E-01	0.116
693213	MIR628	1.65E-03	9.58E-01	-0.365
375757	SWI5	1.68E-03	9.58E-01	0.114
728340	GTF2H2C	1.69E-03	9.58E-01	0.206
146713	RBFOX3	1.86E-03	9.58E-01	-0.434
26834	RNU4-2	1.89E-03	9.58E-01	1.028
79725	THAP9	1.93E-03	9.58E-01	0.149
164668	APOBEC3H	1.98E-03	9.58E-01	-0.323

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

p (diffexp) – p value for differential expression between HD and controls; q (diffexp) – q value shows correction for multiple testing in the combined dataset; Log2(FC) – log2 of the ratio of the mean counts in HD and controls. Transcripts with P<0.002 are shown.

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.

Reference dataset	**Comparison dataset**	**Direction of dysregulation in HD**	**Number of pathways significant in both datasets (p value)**		
			Generic pathways	HD brain modules	Control brain modules
LUMC	TRACK-HD	Nondirectional	69 (4.6E-02)	-	-
		Downregulated	139 (<1.0E-03)	4 (1.1E-01)	24 (<1.0E-03)
		Upregulated	219 (<1.0E-03)	9 (<1.0E-03)	23 (<1.0E-03)
LUMC	TRACK-HD	Nondirectional	69 (1.4E-01)	-	-
		Downregulated	130 (1.7E-02)	4 (3.5E-02)	24 (<1.0E-03)
		Upregulated	217 (<1.0E-03)	10 (<1.0E-03)	21 (<1.0E-03)

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 Neueder 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.

Pathway	Number of dysregulated genes	p-value (combined)	q-value (combined)	p-value (Track-HD)	p-value (Leiden)	Description
MGI: 2419	434	3.03E-10	4.32E-06	5.10E-05	3.01E-05	abnormal innate immunity
MGI: 3009	432	5.78E-09	4.13E-05	5.96E-06	1.65E-04	abnormal cytokine secretion
GO: 50792	117	2.59E-08	1.23E-04	1.12E-02	7.24E-05	regulation of viral process
GO: 9615	208	1.22E-07	4.36E-04	3.06E-02	5.34E-06	response to virus
MGI: 2451	278	1.68E-07	4.80E-04	1.26E-02	9.51E-06	abnormal macrophage physiology
GO: 19221	308	2.38E-07	5.45E-04	4.60E-05	1.71E-04	cytokine-mediated signalling pathway
GO: 2252	365	3.10E-07	5.45E-04	7.01E-03	1.14E-04	immune effector process
MGI: 5025	406	3.44E-07	5.45E-04	5.91E-05	2.02E-04	abnormal response to infection
MGI: 1793	372	4.33E-07	5.82E-04	5.93E-05	2.42E-04	altered susceptibility to infection
MGI: 8568	305	4.49E-07	5.82E-04	4.79E-05	6.25E-05	abnormal interleukin secretion
GO: 48525	48	6.05E-07	7.09E-04	1.83E-02	6.41E-05	negative regulation of viral process
MGI: 8250	462	6.46E-07	7.09E-04	4.99E-04	4.84E-03	abnormal myeloid leukocyte morphology
REACTOME 287	264	8.59E-07	8.76E-04	1.24E-03	3.86E-05	REACT:CYTOKINE SIGNALLING IN IMMUNE SYSTEM
GO: 71345	403	1.15E-06	1.08E-03	4.41E-05	4.82E-04	cellular response to cytokine stimulus
GO: 45069	53	1.21E-06	1.08E-03	1.37E-02	2.99E-05	regulation of viral genome replication
GO: 45071	37	2.30E-06	1.93E-03	1.19E-02	1.36E-04	negative regulation of viral genome replication

GO: 1817	409	2.76E-06	2.19E-03	2.64E-03	1.81E-03	regulation of cytokine production
MGI: 8251	387	3.12E-06	2.34E-03	6.32E-04	2.83E-03	abnormal phagocyte morphology
GO: 31347	430	4.73E-06	3.38E-03	7.31E-04	6.90E-04	regulation of defence response
MGI: 2406	317	6.87E-06	4.67E-03	1.11E-03	1.33E-03	increased susceptibility to infection
GO: 50778	403	7.29E-06	4.73E-03	1.50E-02	3.41E-02	positive regulation of immune response
MGI: 8835	258	8.31E-06	5.16E-03	1.25E-02	1.67E-04	abnormal intercellular signalling peptide or protein level
MGI: 2444	438	1.14E-05	6.69E-03	8.89E-03	6.28E-03	abnormal T cell physiology
MGI: 5005	243	1.21E-05	6.69E-03	1.05E-03	1.25E-04	abnormal self-tolerance
MGI: 1844	242	1.28E-05	6.69E-03	8.83E-04	1.07E-04	autoimmune response
MGI: 8713	253	1.35E-05	6.69E-03	1.24E-02	3.66E-04	abnormal cytokine level
GO: 5773	379	1.36E-05	6.69E-03	7.03E-03	9.19E-03	vacuole
GO: 323	318	1.41E-05	6.69E-03	5.94E-03	8.00E-04	lytic vacuole
GO: 5764	318	1.41E-05	6.69E-03	5.94E-03	8.00E-04	lysosome
MGI: 5000	246	1.63E-05	7.52E-03	1.07E-03	1.71E-04	abnormal immune tolerance
MGI: 8195	412	1.94E-05	8.41E-03	3.20E-03	1.90E-04	abnormal antigen presenting cell morphology
MGI: 8469	437	2.68E-05	1.08E-02	2.01E-02	2.07E-04	abnormal protein level
GO: 2253	325	2.92E-05	1.13E-02	3.85E-02	4.33E-02	activation of immune response
REACTOME 589	161	3.02E-05	1.14E-02	2.65E-03	5.75E-04	REACT:INTERFERON SIGNALLING
MGI: 2441	257	3.42E-05	1.25E-02	1.02E-02	9.51E-03	abnormal granulocyte morphology
MGI: 2462	180	4.09E-05	1.40E-02	6.48E-04	2.50E-02	abnormal granulocyte physiology

GO: 71357	61	4.20E-05	1.40E-02	1.17E-02	8.60E-05	cellular response to type I interferon
GO: 60337	61	4.20E-05	1.40E-02	1.17E-02	8.60E-05	type I interferon-mediated signalling pathway
GO: 43903	134	4.28E-05	1.40E-02	1.11E-02	5.23E-04	regulation of symbiosis, encompassing mutualism through parasitism
MGI: 8248	307	4.39E-05	1.40E-02	4.16E-03	6.89E-03	abnormal mononuclear phagocyte morphology
GO: 44437	243	4.50E-05	1.40E-02	6.10E-04	1.03E-02	vacuolar part
GO: 5765	135	4.97E-05	1.51E-02	1.88E-04	7.32E-03	lysosomal membrane
MGI: 2446	240	5.88E-05	1.75E-02	1.30E-03	1.97E-02	abnormal macrophage morphology
REACTOME 587	61	6.00E-05	1.75E-02	6.87E-03	1.32E-04	REACT:INTERFERON ALPHA BETA SIGNALLING
MGI: 2425	192	7.94E-05	2.27E-02	2.34E-03	3.21E-04	altered susceptibility to autoimmune disorder
MGI: 10210	188	8.70E-05	2.44E-02	2.09E-02	7.63E-03	abnormal circulating cytokine level
GO: 34340	62	8.95E-05	2.46E-02	1.33E-02	6.41E-05	response to type I interferon
MGI: 2463	126	9.20E-05	2.48E-02	2.79E-03	1.49E-02	abnormal neutrophil physiology
MGI: 2498	226	1.01E-04	2.66E-02	1.18E-02	1.69E-02	abnormal acute inflammation
MGI: 2459	402	1.11E-04	2.77E-02	2.28E-02	1.08E-03	abnormal B cell physiology
KEGG 5164	158	1.14E-04	2.80E-02	5.13E-02	2.53E-03	KEGG INFLUENZA A
MGI: 8704	106	1.54E-04	3.66E-02	4.76E-03	9.41E-03	abnormal interleukin-6 secretion
MGI: 8705	48	2.00E-04	4.68E-02	4.43E-02	4.94E-02	increased interleukin-6 secretion

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.*

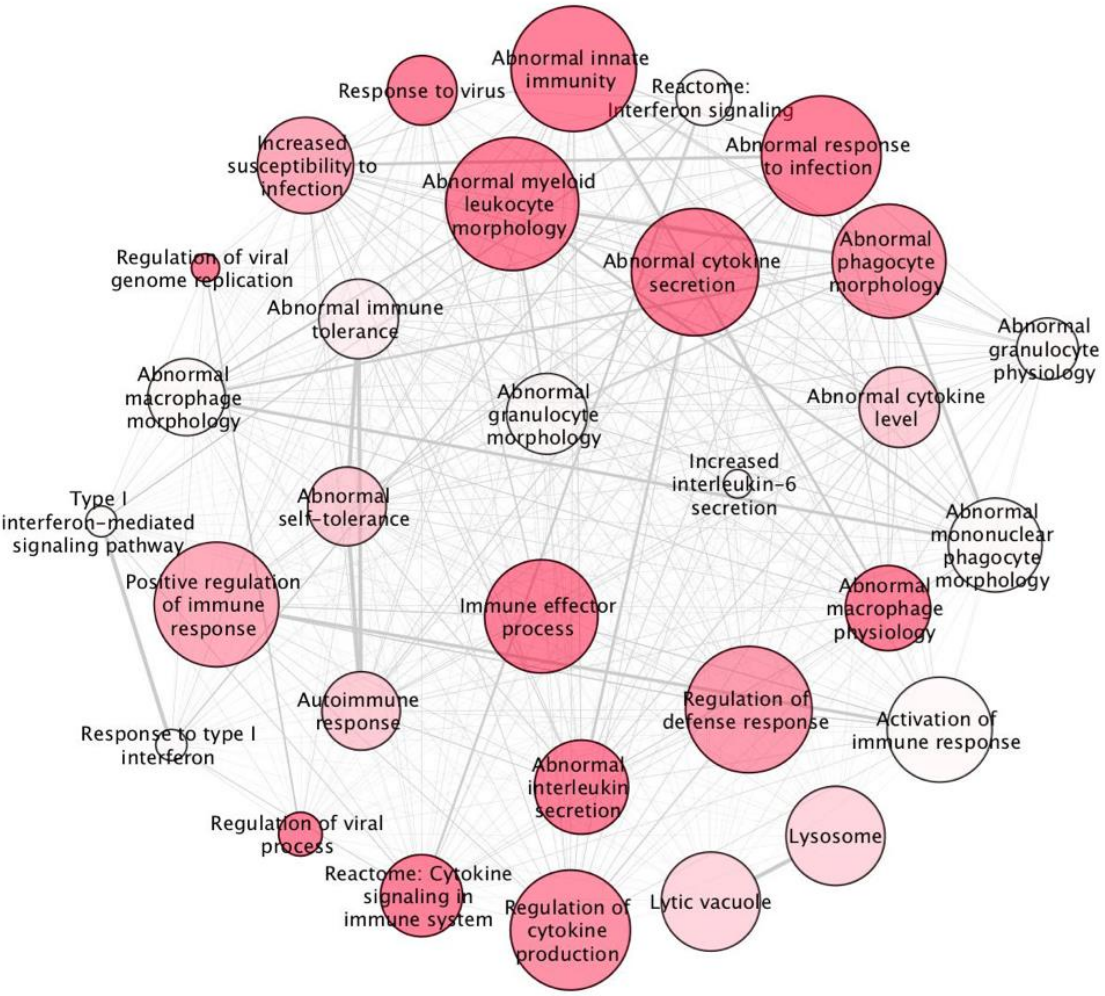


Figure 7.1*: Upregulated pathways in HD versus control blood. Schematic representation of pathways collated from publicly available databases that are significantly upregulated in HD versus controls after correction for multiple testing (q < 0.05). Modules with similar gene content and functional annotation 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).*

Pathway	Number of dysregulated genes	p-value (combined)	q-value (combined)	p-value (Track-HD)	p-value (Leiden)	Description
GO: 8380	282	5.22E-08	7.45E-04	4.25E-05	7.24E-05	RNA splicing
GO: 6397	359	2.38E-07	1.70E-03	1.48E-04	4.14E-04	mRNA processing
GO: 16887	329	1.37E-06	5.48E-03	1.96E-04	3.34E-02	ATPase activity
GO: 6200	333	1.54E-06	5.48E-03	2.42E-04	3.36E-02	ATP catabolic process
GO: 46034	361	5.36E-06	1.53E-02	1.74E-04	4.45E-02	ATP metabolic process
GO: 16607	144	9.06E-06	2.15E-02	4.68E-04	4.61E-03	nuclear speck
GO: 6281	356	1.66E-05	2.75E-02	2.00E-03	1.18E-04	DNA repair
GO: 16604	271	2.08E-05	2.75E-02	5.59E-03	2.46E-03	nuclear body
GO: 4386	135	2.12E-05	2.75E-02	2.83E-02	4.81E-02	helicase activity
GO: 375	184	2.40E-05	2.86E-02	1.14E-03	2.05E-03	RNA splicing, via transesterification reactions
MGI: 5094	219	4.60E-05	3.88E-02	1.87E-02	2.34E-02	Abnormal T cell proliferation
GO: 398	180	6.25E-05	3.88E-02	2.12E-03	2.01E-03	mRNA splicing, via spliceosome
GO: 377	180	6.25E-05	3.88E-02	2.12E-03	2.01E-03	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GO: 5681	143	7.29E-05	4.34E-02	4.66E-03	3.52E-03	spliceosomal complex

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.*

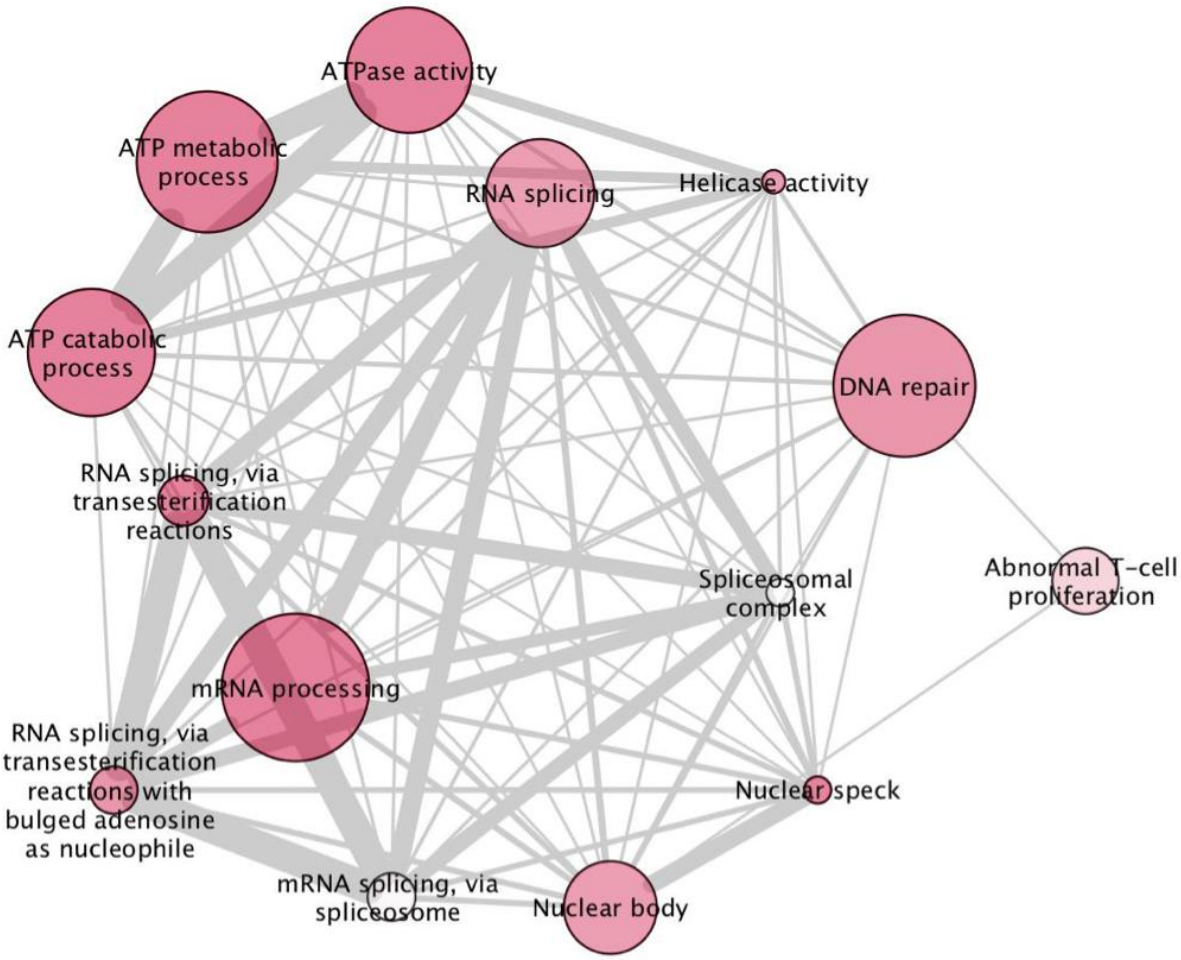


***Figure 7.2**: Downregulated pathways in HD versus control blood. Schematic representation of pathways collated from publicly available databases that are significantly downregulated in HD versus controls after correction for multiple testing ($q < 0.05$). Modules with similar gene content and functional 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)
Genes in upregulated pathways	722	C4BPA	7.81E-06	1.371	1.29E-01	0.437	7.36E-01	0.187	GO:2252, GO:2253, GO:5773, GO:31347, GO:44437, GO:50778
	8763	CD164	9.53E-04	0.098	2.97E-01	0.083	5.57E-03	0.101	GO:323, GO:5764, GO:5765, GO:5773, GO:44437
	597	BCL2A1	1.06E-03	0.423	8.85E-02	0.319	1.20E-02	0.393	MGI:1793, MGI:2419, MGI:2462, MGI:2463, MGI:5025
	4940	OAS3	1.12E-03	0.688	5.14E-02	0.602	6.45E-02	0.455	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
	49	ACR	1.13E-03	1.237	7.54E-03	1.417	1.79E-01	0.768	GO:5773, GO:44437
	9262	STK17B	1.19E-03	0.132	4.42E-02	0.134	3.56E-02	0.136	MGI:1844, MGI:2425, MGI:2444, MGI:3009, MGI:5000, MGI:5005, MGI:8568
	164668	APOBEC3H	1.98E-03	-0.323	1.41E-01	-0.208	4.33E-03	-0.476	GO:2252, GO:9615, GO:43903, GO:45069, GO:45071, GO:48525, GO:50792
	79026	AHNAK	2.12E-03	-0.169	1.48E-02	-0.201	1.27E-01	-0.106	MGI:1793, MGI:2406, MGI:2444, MGI:3009, MGI:5025, MGI:8568
	6614	SIGLEC1	4.39E-03	0.634	3.58E-01	0.291	9.79E-02	0.552	MGI:2459, MGI:8195
	875	CBS	4.42E-03	0.592	1.15E-01	0.439	2.38E-02	0.681	MGI:8469, MGI:8713, MGI:8835

Genes in downregulated pathways	9262	STK17B	1.19E-03	0.132	4.42E-02	0.134	3.56E-02	0.136	MGI:5094
	54957	TXNL4B	1.65E-03	0.088	2.99E-02	0.088	2.67E-02	0.090	GO:5681, GO:6397, GO:8380
	375757	SWI5	1.68E-03	0.114	3.22E-02	0.112	2.67E-02	0.130	GO:6281
	146713	RBFOX3	1.86E-03	-0.434	3.81E-02	-0.396	7.65E-02	-0.357	GO:6397, GO:8380
	79026	AHNAK	2.12E-03	-0.169	1.48E-02	-0.201	1.27E-01	-0.106	MGI:5094
	29890	RBM15B	2.67E-03	-0.055	9.18E-02	-0.048	8.98E-02	-0.044	GO:6397, GO:8380
	9987	HNRNPDL	3.38E-03	-0.078	2.98E-02	-0.088	9.41E-03	-0.098	GO:5681
	23499	MACF1	3.72E-03	-0.120	4.52E-03	-0.172	2.15E-01	-0.068	GO:6200, GO:16887, GO:46034
	146754	DNAH2	4.04E-03	-0.621	1.82E-01	-0.415	2.39E-02	-0.723	GO:6200, GO:16887, GO:46034
	10236	HNRNPR	5.92E-03	-0.069	1.15E-01	-0.053	5.62E-02	-0.074	GO:375, GO:377, GO:398, GO:5681, GO:6397, GO:8380

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.

Direction of dysregulation in HD	**Number of pathways significant in both datasets (p value)**
Nondirectional	132 (0.009)
Downregulated	36 (0.113)
Upregulated	339 (<1.0E-03)

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.*

Direction of dysregulation in HD	Pathway	Number of genes	p (blood: London+ Leiden)	p (myeloid un-stimulated)	p (blood & myeloid combined)	Description
Upregulated	MGI: 2419	434	3.03E-10	3.77E-08	4.44E-16	abnormal innate immunity
	MGI: 3009	432	5.78E-09	4.26E-07	8.54E-14	abnormal cytokine secretion
	GO: 31347	430	4.73E-06	8.96E-09	1.35E-12	regulation of defence response
	GO: 9615	208	1.22E-07	9.68E-07	3.64E-12	response to virus
	MGI: 2451	278	1.68E-07	1.83E-06	9.16E-12	abnormal macrophage physiology
	GO: 2252	365	3.10E-07	1.95E-06	1.76E-11	immune effector process
	MGI: 1793	372	4.33E-07	2.30E-06	2.85E-11	altered susceptibility to infection
	MGI: 8568	305	4.49E-07	3.26E-06	4.13E-11	abnormal interleukin secretion
	MGI: 5025	406	3.44E-07	4.42E-06	4.28E-11	abnormal response to infection
	MGI: 8835	258	8.31E-06	1.92E-07	4.49E-11	abnormal intercellular signalling peptide or protein level
	MGI: 8713	253	1.35E-05	2.72E-07	1.01E-10	abnormal cytokine level
	GO: 51607	138	3.28E-07	1.34E-05	1.19E-10	defence response to virus
	GO: 1817	409	2.76E-06	1.68E-06	1.26E-10	regulation of cytokine production
	REACTOME 287	264	8.59E-07	6.56E-06	1.52E-10	REACT:CYTOKINE SIGNALING IN IMMUNE SYSTEM
	GO: 6954	352	2.72E-05	2.38E-07	1.73E-10	inflammatory response

	GO: 50792	117	2.59E-08	4.23E-04	2.88E-10	regulation of viral process
Downregulated	GO: 43202	66	4.25E-02	2.47E-04	1.31E-04	lysosomal lumen
	GO: 10921	88	1.48E-03	2.41E-02	3.99E-04	regulation of phosphatase activity
	GO: 38024	57	1.91E-03	3.51E-02	7.10E-04	cargo receptor activity
	GO: 16874	450	1.70E-03	4.11E-02	7.38E-04	ligase activity
	MGI: 358	276	2.71E-02	6.37E-03	1.67E-03	abnormal cell morphology
	REACTOME 596	10	3.98E-03	4.49E-02	1.72E-03	REACT:INTERLEUKIN-7 SIGNALING
	GO: 6399	126	4.03E-02	7.24E-03	2.67E-03	tRNA metabolic process
	GO: 2285	55	1.71E-02	2.09E-02	3.20E-03	lymphocyte activation involved in immune response
	GO: 6457	185	3.37E-02	1.10E-02	3.29E-03	protein folding
	GO: 70286	8	2.03E-02	1.83E-02	3.30E-03	axonemal dynein complex assembly
	GO: 6302	99	1.46E-02	3.04E-02	3.88E-03	double-strand break repair
	GO: 30165	75	4.59E-02	1.06E-02	4.18E-03	PDZ domain binding
	MGI: 2419	434	2.06E-02	2.47E-02	4.37E-03	abnormal innate immunity
	MGI: 1701	56	1.05E-02	4.94E-02	4.45E-03	incomplete embryo turning
	GO: 19320	66	2.34E-02	2.67E-02	5.22E-03	hexose catabolic process
	KEGG 4146	77	1.92E-02	4.05E-02	6.35E-03	KEGG PEROXISOME

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**.

Direction	Brain expression gene set	Module	Brain region	Annotation	Number of dysregulated genes	p (Combined)	p (Track-HD)	p (Leiden)	Cor (HD)	BH (HD)
Upregulated	HD	111	FC_BA9	Immune response	514	7.81E-12	1.27E-04	7.53E-05	-	-
	HD	69 (FC4pos1)	FC_BA4	Inflammatory response	712	3.77E-08	3.05E-05	1.32E-03	0.61	3.77E-03
	Control (B)	712	TCTX	Inflammatory response	213	1.41E-07	3.40E-05	8.14E-04	-	-
	HD*	48 (CNpos2)*	CN	Lipid metabolism/regulation of transcription	1785	2.03E-07	3.85E-03	6.33E-03	0.72	2.21E-11
	Control (B)	110	FCTX	Inflammatory response	173	8.94E-07	1.04E-03	2.50E-03	-	-
	Control (B)	909	White Matter	Activation of immune response	265	2.12E-06	1.24E-03	2.48E-02	-	-
	Control (B)	610	Substantia Nigra	Inflammatory response	178	1.21E-05	8.56E-04	5.57E-04	-	-
	Control (B)	811	Thalamus	Inflammatory response	142	1.61E-05	3.94E-03	2.89E-03	-	-
	Control (G)	56	Pons	Lipoprotein/ immune response /GTPase regulator activity	207	1.97E-05	2.44E-04	4.19E-02	-	-
	Control (B)	911	White Matter	Inflammatory response	159	3.00E-05	8.42E-04	1.39E-02	-	-
	HD	28	CB	Immune response	209	3.11E-05	1.07E-02	1.19E-02	-	-
	Control (B)	713	TCTX	Activation of immune	171	4.02E-05	2.39E-02	4.67E-02	-	-

				response						
	HD	33	CB	Immune response	255	4.34E-05	1.08E-02	1.37E-02	-	-
	Control (B)	505	Putamen	Ether lipid metabolism	500	6.28E-05	3.16E-03	2.06E-02	-	-
	HD	68 (CNpos5)	CN	Cilium	1268	1.09E-04	3.05E-02	5.00E-02	0.54	7.74E-06
	Control (B)	516	Putamen	Cellular response to cytokine stimulus	133	3.07E-04	1.44E-02	1.71E-02	-	-
	HD	64 (CNpos6)	CN	Inflammatory response	114	3.13E-04	1.18E-02	3.80E-02	0.46	2.28E-04
	HD	124	FC_BA9	NA	1176	2.91E-03	1.19E-02	2.37E-02	-	-
Downregulated	Control (G)	22	CB	Pro-rich region	831	1.83E-08	2.49E-03	2.06E-02	-	-
	Control (G)	28	FC	Intra-cellular transport/mitochondrion	3178	2.10E-08	6.30E-04	7.66E-05	-	-
	Control (B)	304	Medulla	mRNA metabolic process	1811	2.91E-08	5.00E-15	4.01E-02	-	-
	HD*	66 (CNneg1)*	CN	Synapse/ion channels	2645	2.71E-07	1.51E-04	2.13E-02	-0.80	6.03E-15
	Control (B)	804	Thalamus	Regulation of cell morphogenesis	857	1.31E-06	4.03E-02	4.13E-04	-	-
	Control (B)	522	Putamen	Regulation of RNA splicing	64	4.44E-06	6.26E-03	2.66E-04	-	-
	Control (G)	74	Pons	Transcription/acetylation/protein transport	1183	9.22E-06	3.85E-08	7.44E-04	-	-
	Control (B)	702	TCTX	Antigen processing:	4602	3.87E-04	1.22E-03	2.47E-02	-	-

				ubiquitination and proteasome degradation						
	Control (G)	48	FC	Transcription corepressor/cell morphogenesis	648	4.65E-04	7.83E-03	2.05E-02	-	-
	Control (B)	202	Hippocampus	Mitochondrial membrane	2737	4.75E-04	1.16E-07	1.54E-02	-	-
	HD	19	CB	Protein binding	155	7.44E-04	2.66E-02	2.26E-02	-	-
	Control (B)	906	White Matter	Uridyltransferase activity	416	1.12E-03	2.53E-02	1.12E-02	-	-
	Control (G)	93	Pons	Mitochondrion/nuclear lumen	317	1.30E-03	9.85E-03	8.74E-04	-	-
	Control (B)	812	Thalamus	Transport of mature transcript to cytoplasm	114	1.42E-03	1.99E-02	4.70E-02	-	-
	HD	102	FC_BA9	Cytoplasm	1908	1.47E-03	7.57E-03	1.31E-04	-	-
	Control (B)	706	TCTX	Microtubule organising centre	481	1.93E-03	3.70E-05	3.80E-03	-	-
	Control (G)	52	Pons	Acetylation/fatty acid metabolism	1590	3.28E-03	2.23E-02	1.31E-02	-	-
	HD	3 (CBneg2)	CB	mitochondrion	1164	3.19E-02	2.56E-02	1.29E-05	-0.45	1.66E-03
	Control (G)	25	CB	RNA binding	648	8.02E-01	1.72E-04	3.62E-02	-	-

***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)

Entrez gene ID	Gene Symbol	p (Comb)	log2FC (Comb)	p (Track-HD)	log2FC (Track-HD)	p (Leiden)	log2FC (Leiden)	Module membership
2297	FOXD1	9.09E-05	-0.785	1.10E-02	-0.685	1.69E-03	-1.014	HD 48 (CNpos2), HD 111
3805	KIR2DL4	1.93E-04	0.651	2.57E-03	0.823	1.52E-02	0.533	CTRL (B) 702
196394	AMN1	2.11E-04	0.208	1.87E-02	0.205	9.25E-03	0.195	CTRL (B) 202, CTRL (B) 702
5797	PTPRM	3.12E-04	-0.359	5.26E-03	-0.381	2.82E-03	-0.448	CTRL (B) 202, CTRL (B) 702, CTRL (B) 904, HD 66 (CNneg1)
889	KRIT1	7.30E-04	-0.081	1.59E-02	-0.097	7.86E-02	-0.057	CTRL (B) 304, CTRL (G) 28, HD 102
22979	EFR3B	8.17E-04	0.494	6.02E-03	0.603	2.07E-02	0.496	CTRL (B) 702, CTRL (B) 904, HD 66 (CNneg1)
56934	CA10	8.42E-04	2.036	1.21E-02	2.020	8.42E-02	1.945	CTRL (B) 702, CTRL (B) 902, HD 66 (CNneg1), HD 102
8763	CD164	9.53E-04	0.098	2.97E-01	0.083	5.57E-03	0.101	HD 68 (CNpos5)
597	BCL2A1	1.06E-03	0.423	8.85E-02	0.319	1.20E-02	0.393	CTRL (B) 110, CTRL (B) 217, CTRL (B) 516, CTRL (B) 610, CTRL (B) 712, CTRL (B) 811, CTRL (B) 911,HD 33, HD 68 (CNpos5), HD 69 (FC4pos1), HD 111
4940	OAS3	1.12E-03	0.688	5.14E-02	0.602	6.45E-02	0.455	CTRL (B) 702, CTRL (B) 902

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 – log2 of the ratio of the mean counts in HD and controls; 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).*

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.

Module	**Brain Region**	**Module name**	**Number of genes**	**p (combined)**	**p (TRACK-HD)**	**p (LUMC)**	**cor (HD brain)**	**p (HD brain)**	**Description**
69	FC_BA4	FC4pos1	712	3.77E-08	3.05E-05	1.32E-03	0.610	3.77E-03	Inflammatory response
48	CN	CNpos2	1785	2.03E-07	3.85E-03	6.33E-03	0.724	2.21E-11	Lipid metabolism/regulation of transcription
64	CN	CNpos6	114	3.13E-04	1.17E-02	3.80E-02	0.463	2.28E-04	Inflammatory response
66	CN	CNneg1	2644	2.71E-07	1.51E-04	2.13E-02	-0.800	6.03E-15	Synapse

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(HD brain) is the p-value for that correlation (corrected for multiple testing of modules).

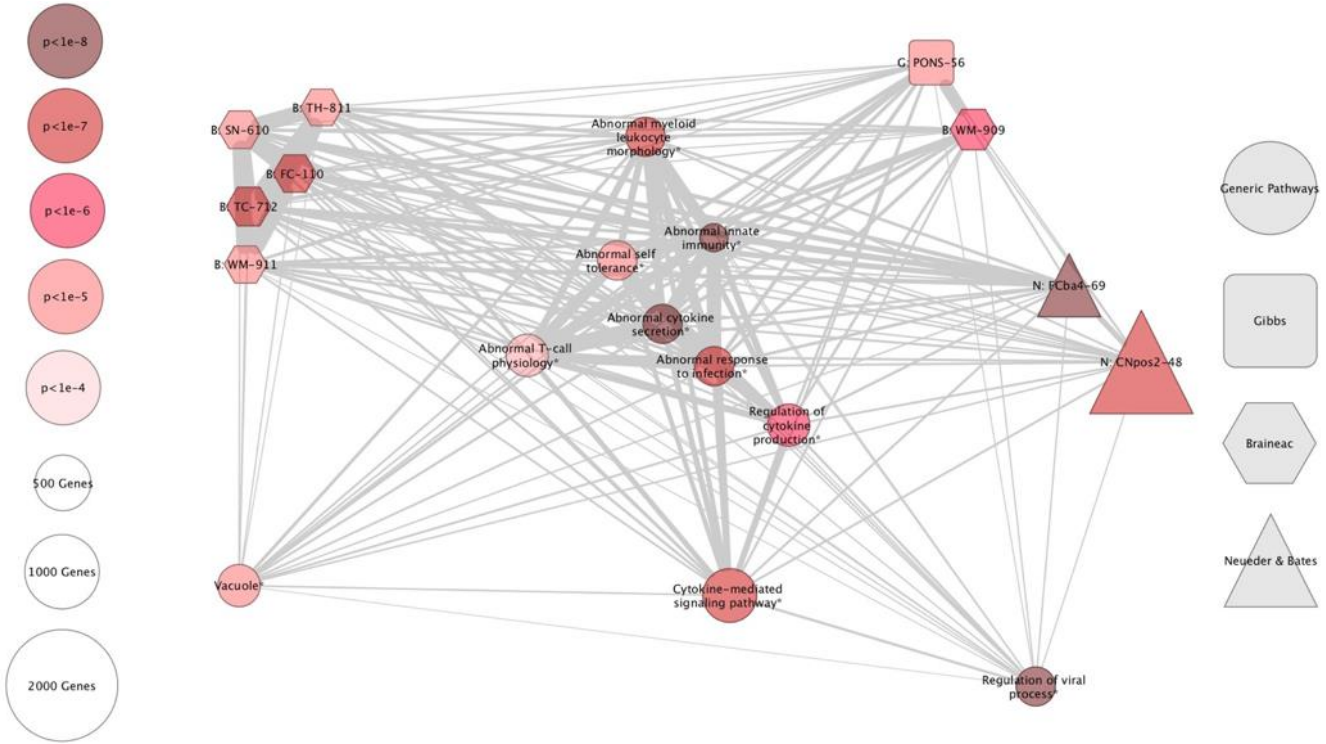


Figure 7.3: *Network diagram of the relationship between significantly (q<0.05) upregulated gene modules (Table 7.10) and generic biological pathways (Table 7.5) based on shared gene membership. The thickness of the edges corresponds to the proportion of overlap from the smaller term to the larger (overlap 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.*

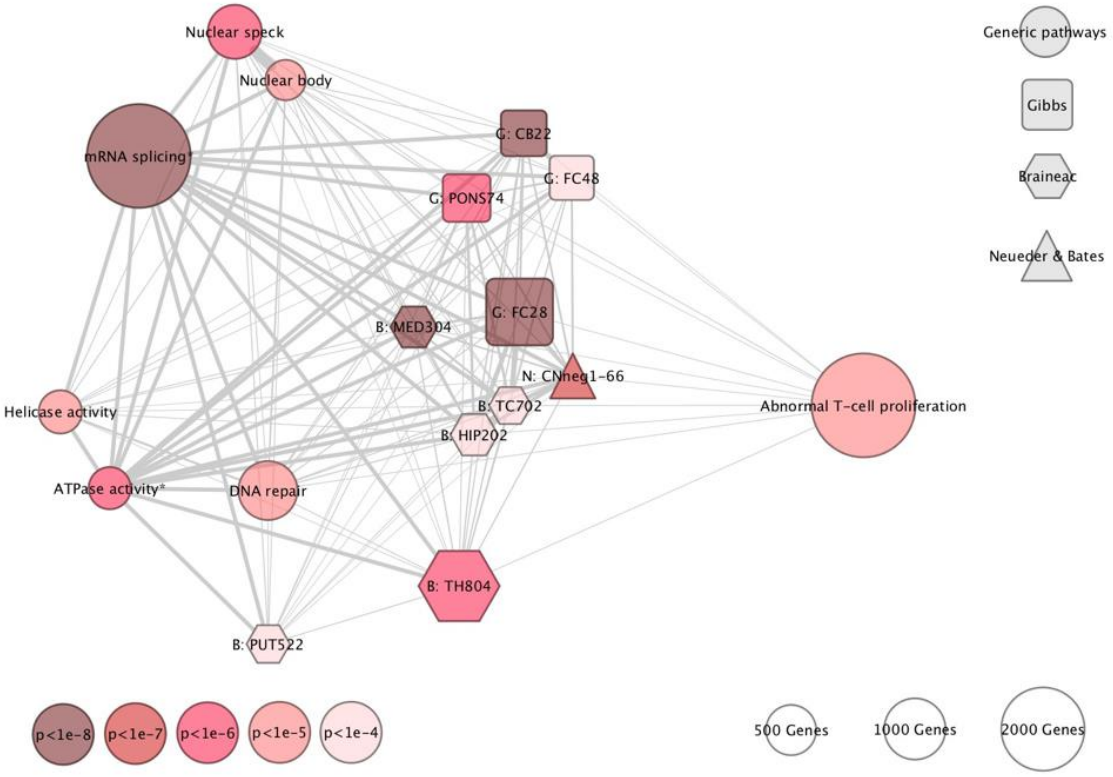


Figure 7.4: *Network diagram of the relationship between significantly (q<0.05) downregulated gene modules (Table 7.10) and generic biological pathways (Table 6) based on shared gene membership. The thickness of the edges corresponds to the proportion of overlap from the smaller term to the larger (overlap 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 ($p = 7.6 \times 10^{-4}$) 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).

Direction of effect	Pathway	Number of dysregulated genes	Blood p (Combined)	Brain p (Labadorf)	Description
Upregulated	MGI: 2459	402	1.11E-04	1.39E-13	abnormal B cell physiology
	MGI: 2419	434	3.03E-10	2.05E-12	abnormal innate immunity
	MGI: 1800	361	5.45E-04	2.58E-12	abnormal humoral immune response
	MGI: 8195	412	1.94E-05	8.78E-12	abnormal antigen presenting cell morphology
	MGI: 2490	333	8.00E-04	3.52E-11	abnormal immunoglobulin level
	MGI: 8250	462	6.46E-07	4.04E-11	abnormal myeloid leukocyte morphology
	MGI: 4939	381	3.31E-03	1.68E-10	abnormal B cell morphology
	GO: 50778	403	7.29E-06	2.11E-10	positive regulation of immune response
	MGI: 8251	387	3.12E-06	3.29E-10	abnormal phagocyte morphology
	MGI: 3009	432	5.78E-09	5.24E-10	abnormal cytokine secretion
Downregulated	GO: 5874	327	4.97E-05	8.10E-05	microtubule
	GO: 86	120	8.11E-03	1.70E-04	G2/M transition of mitotic cell cycle
	GO: 48812	455	3.45E-02	2.20E-04	neuron projection morphogenesis
	PAN-PW 29	120	4.80E-02	2.67E-04	Huntington disease
	GO: 15631	187	4.14E-04	2.84E-04	tubulin binding
	GO: 7017	372	7.94E-04	4.13E-04	microtubule-based process
	MGI: 1828	233	1.66E-04	7.14E-04	abnormal T cell activation

	REACTOME 214	68	3.16E-03	1.12E-03	REACT:CENTROSOME MATURATION
	REACTOME 952	68	3.16E-03	1.12E-03	REACT:RECRUITMENT OF MITOTIC CENTROSOME PROTEINS AND COMPLEXES
	REACTOME 636	59	1.87E-03	1.48E-03	REACT:LOSS OF NLP FROM MITOTIC CENTROSOMES

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 phosphatidylcholine 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).

Entrez gene ID	Gene Symbol	p (corr-TMS)	q (corr-TMS)	log2(FC)
58488	PCTP	1.82E-06	3.25E-02	8.00E-03
51060	TXNDC12	4.42E-05	1.79E-01	5.30E-03
57096	RPGRIP1	4.64E-05	1.79E-01	1.25E-02
9258	MFHAS1	5.12E-05	1.79E-01	-8.40E-03
3667	IRS1	6.73E-05	1.79E-01	-1.37E-02
158293	FAM120AOS	6.88E-05	1.79E-01	3.80E-03
84263	HSDL2	7.01E-05	1.79E-01	6.30E-03
56925	LXN	1.01E-04	2.22E-01	1.05E-02
118881	COMTD1	1.12E-04	2.22E-01	-8.10E-03
597	BCL2A1	1.44E-04	2.35E-01	1.44E-02
23002	DAAM1	1.58E-04	2.35E-01	-7.80E-03
3655	ITGA6	1.58E-04	2.35E-01	-8.00E-03
137835	TMEM71	1.96E-04	2.53E-01	7.00E-03

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).

Direction of effect	Pathway	p (combined-diffexp)	p (TRACK-diffexp)	p (TRACK-TMS)	Description
Positively correlated with TMS	MGI: 2419	3.03E-10	5.10E-05	2.18E-03	Abnormal innate immunity
	GO: 10942	8.79E-02	4.70E-02	3.21E-03	positive regulation of cell death
	MGI: 2462	4.09E-05	6.48E-04	6.39E-03	Abnormal granulocyte physiology
	MGI: 8556	6.85E-04	8.68E-03	7.91E-03	Abnormal tumour necrosis factor secretion
	MGI: 2463	9.20E-05	2.79E-03	8.99E-03	Abnormal neutrophil physiology
	MGI: 8704	1.54E-04	4.76E-03	9.56E-03	abnormal_interleukin-6_secretion
	GO: 5773	1.36E-05	7.03E-03	1.62E-02	vacuole
	GO: 50792	2.59E-08	1.12E-02	1.64E-02	regulation of viral process
	MGI: 5351	6.95E-03	2.20E-02	2.76E-02	Decreased susceptibility to autoimmune disorder
	GO: 44437	4.50E-05	6.10E-04	3.48E-02	vacuolar part
	MGI: 3627	5.45E-02	3.61E-02	3.62E-02	Abnormal leukocyte tethering or rolling
	GO: 50427	1.64E-01	4.26E-02	4.30E-02	3'-phosphoadenosine 5'-phosphosulfate metabolic process
	GO: 34035	1.64E-01	4.26E-02	4.30E-02	purine ribonucleoside bisphosphate metabolic process
	MGI: 2451	1.68E-07	1.26E-02	4.33E-02	Abnormal macrophage physiology
	GO: 6024	1.39E-02	2.77E-02	4.56E-02	glycosaminoglycan biosynthetic process
Negatively correlated with TMS	GO: 45786	8.92E-04	1.88E-02	3.23E-05	negative regulation of cell cycle
	MGI: 706	9.70E-02	1.11E-02	9.09E-05	Small thymus

	MGI: 2364	6.81E-02	1.14E-02	2.57E-04	Abnormal thymus size
	MGI: 5018	6.50E-04	5.94E-03	2.70E-04	Decreased T cell number
	MGI: 2435	1.95E-04	6.87E-03	2.79E-04	Abnormal effector T cell morphology
	MGI: 8081	1.48E-03	5.61E-03	3.83E-04	Abnormal single-positive T cell number
	MGI: 2145	1.19E-03	5.67E-03	8.68E-04	Abnormal T cell differentiation
	MGI: 2444	3.61E-04	7.95E-03	8.74E-04	Abnormal T cell physiology
	MGI: 2432	6.45E-04	3.48E-02	1.01E-03	Abnormal CD4-positive T cell morphology
	MGI: 6387	8.81E-05	4.95E-03	1.31E-03	Abnormal T cell number
	MGI: 8083	7.34E-03	3.36E-02	1.77E-03	Decreased single-positive T cell number
	MGI: 8077	7.89E-03	3.31E-02	2.14E-03	abnormal_CD8-positive_T_cell_number
	MGI: 1823	5.18E-02	4.61E-02	2.15E-03	Thymus hypoplasia
	GO: 6200	1.54E-06	2.42E-04	2.34E-03	ATP catabolic process
	GO: 46034	5.36E-06	1.74E-04	2.56E-03	ATP metabolic process

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.*

Direction	Brain expression gene set	Module	Brain region	Annotation	Number of dysregulated genes	p (Combined-diffexp)	p (TRACK-diffexp)	p (TRACK-TMS)	Cor (HD)	BH (HD)
Upregulated	HD	68 (CNpos5)	CN	Cilium	1268	1.09E-04	3.05E-02	5.52E-07	0.54	7.74E-06
	Control (B)	909	White Matter	Activation of immune response	265	2.12E-06	1.24E-03	8.22E-04	-	-
	Control (B)	713	TCTX	Activation of immune response	171	4.02E-05	2.39E-02	1.69E-03	-	-
	HD	111	FC_BA9	Immune response	514	7.81E-12	1.27E-04	3.75E-03	-	-
	Control (G)	56	Pons	Lipoprotein/ immune response /GTPase regulator activity	207	1.97E-05	2.44E-04	7.72E-03	-	-
	HD	28	CB	Immune response	209	3.11E-05	1.07E-02	8.70E-03	-	-
	Control (B)	505	Putamen	Ether lipid metabolism	500	6.28E-05	3.16E-03	6.43E-02	-	-
	Control (B)	911	White Matter	Inflammatory response	159	3.00E-05	8.42E-04	7.75E-02	-	-
	HD	124	FC_BA9	NA	1176	2.91E-03	1.19E-02	9.14E-02	-	-
	Control (B)	110	FCTX	Inflammatory response	173	8.94E-07	1.04E-03	1.34E-01	-	-
	HD	33	CB	Immune response	255	4.34E-05	1.08E-02	1.52E-01	-	-
	Control (B)	610	Substantia Nigra	Inflammatory response	178	1.21E-05	8.56E-04	2.00E-01	-	-

	HD	64 (CNpos6)	CN	Inflammatory response	114	3.13E-04	1.18E-02	2.22E-01	0.46	2.28E-04
	Control (B)	811	Thalamus	Inflammatory response	142	1.61E-05	3.94E-03	2.28E-01	-	-
	Control (B)	712	TCTX	Inflammatory response	213	1.41E-07	3.40E-05	2.35E-01	-	-
	Control (B)	516	Putamen	Cellular response to cytokine stimulus	133	3.07E-04	1.44E-02	4.16E-01	-	-
	HD	69 (FC4pos1)	FC_BA4	Inflammatory response	712	3.77E-08	3.05E-05	5.22E-01	0.61	3.77E-03
	HD*	48 (CNpos2)	CN	Lipid metabolism/regulation of transcription	1785	2.03E-07	3.85E-03	6.14E-01	0.72	2.21E-11
Downregulated	Control (B)	304	Medulla	mRNA metabolic process	1811	2.91E-08	5.00E-15	6.11E-16	-	-
	Control (B)	702	TCTX	Antigen processing: ubiquitination and proteasome degradation	4602	3.87E-04	1.22E-03	2.04E-13	-	-
	Control (B)	202	Hippocampus	Mitochondrial membrane	2737	4.75E-04	1.16E-07	1.44E-09	-	-
	Control (G)	28	FC	Intra-cellular transport/mitochondrion	3178	2.10E-08	6.30E-04	4.16E-09	-	-
	HD*	66 (CNneg1)	CN	Synapse/ion channels	2645	2.71E-07	1.51E-04	1.05E-07	-0.80	6.03E-15

	Control (G)	52	Pons	Acetylation/fatty acid metabolism	1590	3.28E-03	2.23E-02	1.30E-07	-	-
	Control (G)	74	Pons	Transcription/acetylation/protein transport	1183	9.22E-06	3.85E-08	1.19E-05	-	-
	Control (G)	22	CB	Pro-rich region	831	1.83E-08	2.49E-03	7.72E-05	-	-
	Control (B)	804	Thalamus	Regulation of cell morphogenesis	857	1.31E-06	4.03E-02	8.29E-05	-	-
	Control (B)	706	TCTX	Microtubule organising center	481	1.93E-03	3.70E-05	3.00E-04	-	-
	Control (G)	48	FC	Transcription corepressor/cell morphogenesis	648	4.65E-04	7.83E-03	7.14E-04	-	-
	HD	102	FC_BA9	Cytoplasm	1908	1.47E-03	7.57E-03	9.26E-03	-	-
	Control (B)	906	White Matter	Uridyltransferase activity	416	1.12E-03	2.53E-02	1.34E-02	-	-
	Control (B)	812	Thalamus	Transport of mature transcript to cytoplasm	114	1.42E-03	1.99E-02	1.36E-02	-	-
	HD	19	CB	Protein binding	155	7.44E-04	2.66E-02	2.18E-02	-	-
	HD	3 (CBneg2)	CB	mitochondrion	1164	3.19E-02	2.56E-02	6.17E-02	-0.45	1.66E-03
	Control (G)	93	Pons	Mitochondrion/nuclear lumen	317	1.30E-03	9.85E-03	1.24E-01	-	-
	Control (B)	522	Putamen	Regulation of RNA splicing	64	4.44E-06	6.26E-03	2.52E-01	-	-
	Control (G)	25	CB	RNA binding	648	8.02E-01	1.72E-04	9.99E-01	-	-

***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(combined-diffexp) – enrichment p-value for downregulated genes in the combined Track-HD and Leiden sample. p(TRACK-diffexp) - enrichment p-value for downregulated genes in the Track-HD sample alone. p(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; 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 https://www.nature.com/articles/srep44849#supplementary-information (Hensman Moss et al., 2017a)). 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 Labadorf 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 ($p=4.78x10^{-7}$).

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

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.*

7.3.7 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 ($p = 2.50 \times 10^{-4}$).

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

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**).

Gene ID	log2 Fold Change	Lfc SE	stat	p-value	Adjusted p-value	Selected information from Genecards about protein function (accessed 02/06/2015) (GeneCards)
MSH4	-0.40704	0.102704	19.64606	9.32E-06	0.16819	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.
RRM2	-0.22329	0.060098	17.14964	3.45E-05	0.311739	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)
PBK	-0.37372	0.097875	16.35145	5.26E-05	0.31652	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
ANLN	-0.21363	0.056665	15.57077	7.95E-05	0.358565	GO annotations related to this gene include *actin binding* and *phospholipid binding*.
REN	1.556558	0.421303	14.67092	0.000128	0.436492	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.
PKDCC	0.356707	0.106007	14.36259	0.000151	0.436492	Protein kinase which is required for longitudinal bone growth through regulation

						of chondrocyte differentiation. Involved in protein transport from the Golgi apparatus to the plasma membrane (By similarity)
NCAPG	-0.18097	0.052443	14.14435	0.000169	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
SKA3	-0.23214	0.06639	13.72807	0.000211	0.476607	A microtubule-binding subcomplex of the outer kinetochore that is essential for proper chromosome segregation
DTL	-0.19848	0.060308	12.88651	0.000331	0.653314	E3 ubiquitin-protein ligase complex required for cell cycle control, DNA damage response and translesion DNA synthesis
CCNB1	-0.10543	0.031051	12.71875	0.000362	0.653314	The protein encoded by this gene is a regulatory protein involved in mitosis.
BUB1	-0.15348	0.047313	12.00116	0.000532	0.77395	This gene encodes a serine/threonine-protein kinase that play a central role in mitosis
DLGAP5	-0.19198	0.061511	11.93955	0.00055	0.77395	Potential cell cycle regulator that may play a role in carcinogenesis of cancer cells. Mitotic phosphoprotein regulated by the ubiquitin-proteasome pathway.
CRYZ	-0.14291	0.041116	11.80478	0.000591	0.77395	Binds NADP and acts through a one-electron transfer process.
CCDC152	-0.20601	0.056894	11.77476	0.0006	0.77395	CCDC152 (coiled-coil domain containing 152) is a protein-coding gene.
ZNF684	-0.12772	0.037985	11.3534	0.000753	0.897588	May be involved in transcriptional regulation

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**).

Direction of effect	Pathway	#genes	Enrichment p-value (directional)	q-value (directional)	p (TRACK) GWAS	Description
Negative correlation with HD progression	GO: 280	137	1.11E-15	7.99E-12	6.19E-01	nuclear division
	GO: 7067	7	1.11E-15	7.99E-12	6.19E-01	mitosis
	GO: 48285	150	2.83E-15	1.36E-11	7.95E-01	organelle fission
	GO: 793	481	4.55E-15	1.64E-11	1.51E-01	condensed chromosome
	GO: 51301	211	4.03E-13	1.16E-09	8.07E-01	cell division
	REACTOME 694	3	2.77E-12	6.64E-09	2.64E-01	REACTOME: MITOTIC_M-M_G1_PHASES
	GO: 10564	62	8.48E-12	1.74E-08	9.89E-01	regulation of cell cycle process
	GO: 7059	391	4.12E-11	7.42E-08	9.18E-01	chromosome segregation
	REACTOME 642	25	5.74E-11	9.18E-08	3.20E-01	REACT:M_PHASE
	GO: 775	32	1.11E-10	1.60E-07	3.17E-01	chromosome, centromeric region
	GO: 7346	37	4.31E-10	5.44E-07	9.99E-01	regulation of mitotic cell cycle
	GO: 779	110	4.53E-10	5.44E-07	4.02E-01	condensed chromosome, centromeric region
	GO: 5813	123	8.10E-10	8.53E-07	6.62E-01	centrosome
	GO:1901987	22	8.30E-10	8.53E-07	9.86E-01	regulation of cell cycle phase transition
	GO: 777	130	2.44E-09	2.34E-06	3.35E-01	condensed chromosome kinetochore
	REACTOME 695	169	3.63E-09	3.26E-06	2.34E-01	REACTOME: MITOTIC PROMETAPHASE
	GO: 901988	53	4.33E-09	3.67E-06	9.37E-01	negative regulation of cell cycle phase transition
	GO: 776	24	6.20E-09	4.96E-06	2.02E-01	kinetochore

	GO: 44772	29	7.84E-09	5.64E-06	4.12E-01	mitotic cell cycle phase transition
	GO: 44770	29	7.84E-09	5.64E-06	4.12E-01	cell cycle phase transition
Positive correlation with HD progression	NCI: 26	36	1.31E-04	0.844	6.85E-01	NCI:_VALIDATED TARGETS OF C-MYC TRANSCRIPTIONAL ACTIVATION
	GO: 32320	8	1.54E-04	0.844	7.91E-01	positive regulation of Ras GTPase activity
	GO: 5099	87	1.78E-04	0.844	6.62E-01	Ras GTPase activator activity
	GO: 3001	4	3.12E-04	0.844	9.78E-01	generation of a signal involved in cell-cell signalling
	GO: 23061	4	3.12E-04	0.844	9.78E-01	signal release
	GO: 71013	4	3.52E-04	0.844	4.02E-01	catalytic step 2 spliceosome
	GO: 5681	4	5.01E-04	0.912	2.66E-01	spliceosomal complex
	GO: 32318	144	7.19E-04	0.912	5.99E-01	regulation of Ras GTPase activity
	GO: 71814	160	9.35E-04	0.912	1.12E-01	protein-lipid complex binding
	GO: 71813	113	9.35E-04	0.912	1.12E-01	lipoprotein particle binding

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**).

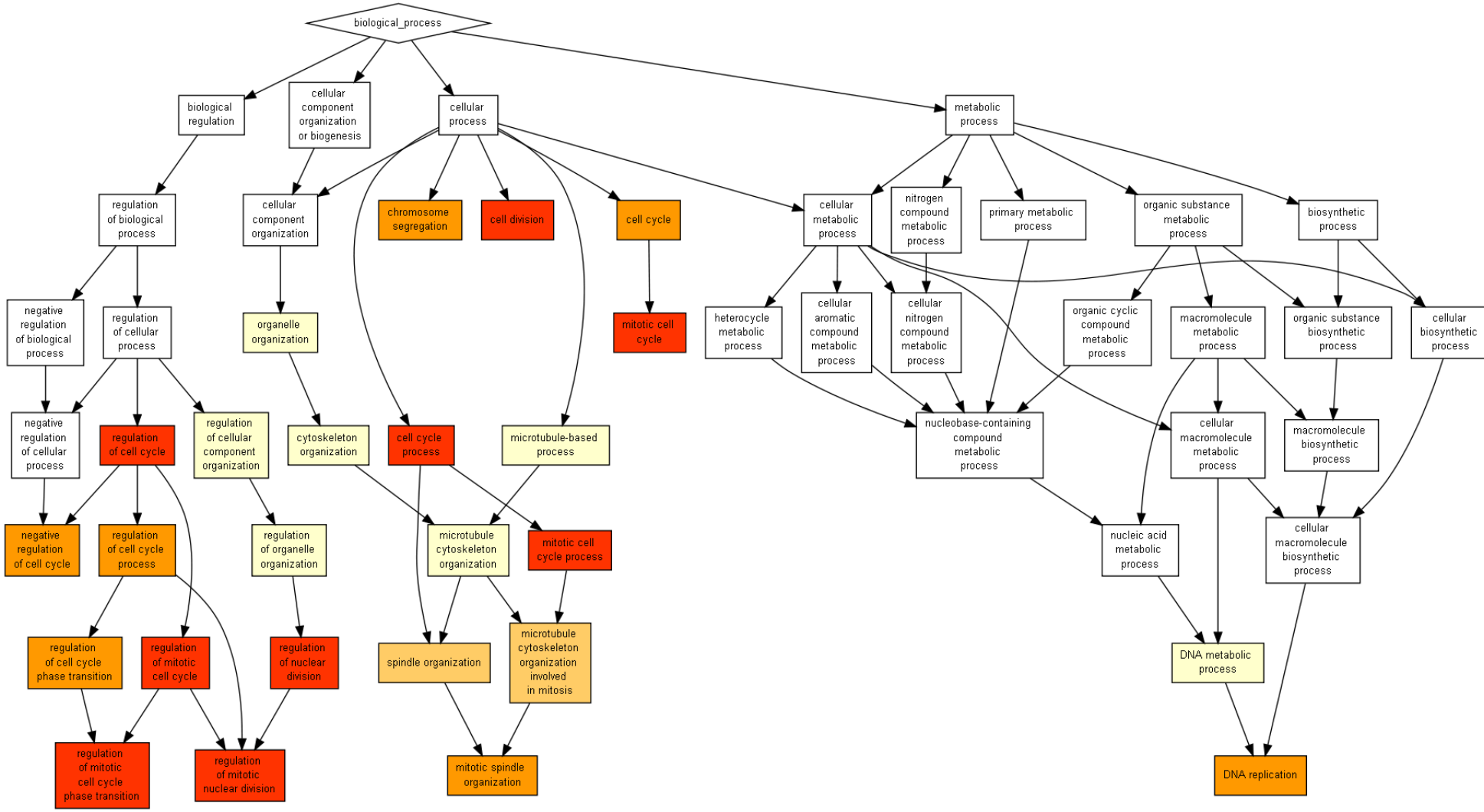


Figure 7.5*: Cell cycle pathways expression is associated with rate of HD progression***.** Diagrammatic illustration of the GO terms enriched in the top vs bottom of the list of transcripts in which transcripts are ranked according to differential expression with rate of HD progression. P-value cut-off for GO terms set at 10^-7. GOrilla accessed 22/11/2017.

GO term	Description	P-value	FDR q-value
GO:1903047	mitotic cell cycle process	3.72E-14	5.61E-10
GO:0022402	cell cycle process	1.25E-13	9.41E-10
GO:0007346	regulation of mitotic cell cycle	5.05E-11	2.54E-7
GO:0051726	regulation of cell cycle	7.23E-11	2.72E-7
GO:0007088	regulation of mitotic nuclear division	1.16E-10	3.5E-7
GO:0051301	cell division	1.46E-10	3.67E-7
GO:0000278	mitotic cell cycle	3.99E-10	8.59E-7
GO:1901990	regulation of mitotic cell cycle phase transition	6.71E-10	1.26E-6
GO:0051783	regulation of nuclear division	7.43E-10	1.24E-6
GO:1901987	regulation of cell cycle phase transition	1.46E-9	2.2E-6
GO:0010564	regulation of cell cycle process	1.73E-9	2.37E-6
GO:0007049	cell cycle	2.55E-9	3.2E-6
GO:0007059	chromosome segregation	2.51E-8	2.91E-5
GO:0006260	DNA replication	2.86E-8	3.08E-5
GO:0007052	mitotic spindle organization	3.28E-8	3.3E-5
GO:0045786	negative regulation of cell cycle	7.45E-8	7.02E-5
GO:0051276	chromosome organization	1.03E-7	9.16E-5
GO:0090068	positive regulation of cell cycle process	1.5E-7	1.26E-4
GO:0045787	positive regulation of cell cycle	1.97E-7	1.56E-4
GO:0000075	cell cycle checkpoint	2.08E-7	1.57E-4
GO:1902850	microtubule cytoskeleton organization involved in mitosis	2.34E-7	1.68E-4
GO:0033044	regulation of chromosome organization	2.67E-7	1.83E-4

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).

Direction of correlation enrichment	**Module**	**Brain Region**	**#genes**	**Enrichment p-value**	**cor(HD)**	**BH (HD)**	**Annotations**
Positive	89	FC_BA4	390	4.14E-06	NA	NA	NA
	119	FC_BA9	696	1.90E-05	NA	NA	NA
	66	CN	2624	5.40E-05	-0.800	6.03E-15	Synapse / ion channels
	25	CB	362	6.86E-05	-0.450	1.66E-03	mitochondrion
	46	CN	1016	5.71E-04	0.744	4.03E-12	Regulation of transcription/ mRNA/ chromatin modification
	26	CB	247	2.55E-03	NA	NA	NA
	100	FC_BA4	78	2.66E-03	NA	NA	NA
	44	CB	628	6.37E-03	0.336	2.90E-02	Zinc finger binding / chromatin modification
	98	FC_BA4	1350	1.82E-02	-0.445	3.35E-02	Glycolysis / protein transport
	75	FC_BA	44	3.75E-	-	1.21	Fibronectin

		4		02	0.506	E-02	
Negative	110	FC_BA 9	752	6.38E-11	NA	NA	
	76	FC_BA 4	956	1.18E-06	-0.570	3.77 E-03	Protein transport
	34	CB	336	2.62E-06	0.327	3.30 E-02	Protein folding/ mRNA/ chromatin assembly
	68	CN	1264	2.44E-05	0.540	7.74 E-06	Cilium
	104	FC_BA 9	186	3.91E-05	NA	NA	NA
	111	FC_BA 9	512	6.28E-05	NA	NA	Immune Response
	49	CN	637	2.12E-04	-0.456	2.66 E-04	Mitochondrion / translation / proteosome / DNA repair
	116	FC_BA 9	183	2.91E-04	NA	NA	NA
	59	CN	100	4.50E-04	NA	NA	NA
	16	CB	512	2.42E-03	NA	NA	NA

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**).

Direction of correlation enrichment	**Module**	**Brain region**	**#genes**	**Enrichment p**	**Annotations**
Positive	113	TCTX	619	<10-16	Proteasome/acetylation/mRNA metabolic process
	23	CB	788	3.16E-15	Protein

					modification/KRAB/translation initiation factor activity
	45	FCTX	1001	1.07E-13	Membrane enclosed lumen
	74	Pons	1182	5.44E-13	Transcription/acetylation/protein folding/histone deacetylase
	26	CB	591	7.42E-13	Protein transport/Golgi apparatus/proteolysis
	115	TCTX	422	2.69E-09	Nucleus
	89	Pons	476	1.88E-08	KRAB/phosphoprotein/GTP binding
	47	FCTX	868	5.57E-08	Small conjugating protein ligase/golgi apparatus
	90	Pons	432	1.30E-05	Phosphoprotein/lytic vacuole
	105	TCTX	1135	7.84E-05	Vesicle-mediated transport/cytoplasm
Negative	22	CB	831	2.71E-12	Compositionally biased region: Pro-rich
	95	TCTX	361	1.57E-08	Golgi membrane/Compositionally biased region: Pro-rich
	48	FCTX	647	2.24E-08	Golgi cisterna/transcription corepressor activity/Pro-rich
	85	Pons	556	1.18E-05	Lipid binding/nucleoside-triphosphatase regulator activity
	66	Pons	136	1.23E-05	Phosphoprotein
	63	Pons	1542	1.40E-05	DNA binding/diencephalon development/protease
	111	TCTX	926	1.34E-04	MAPKinase signalling pathway/chromatin remodelling
	39	FCTX	1325	2.25E-04	DNA binding/hormone/disulphide bond
	103	TCTX	72	3.87E-04	nucleoplasm part
	112	TCTX	731	5.23E-04	Calcium binding/actin binding/tight junction/endocytosis

***Table 7.23**: Correlation enrichment between Gibbs modules(Gibbs et al., 2010) and differential transcription according to progression: top ten modules in each direction shown. FCTX – frontal cortex; CB – cerebellum; TCTX – temporal cortex.*

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**).

Gene ID	log2FoldChange	Standard error of LFC	stat	p-value	Brief function from GeneCards
TME	0.2414876	0.058	17.72	2.55E-	Enhances production of pro-inflammatory

M9		152	727	05	cytokines induced by TNF, IL1B, and TLR ligands.
CHST2	-0.2747902	0.067503	17.32021	3.16E-05	Among its related pathways are Disease and Metabolism. GO annotations related to this gene include *sulfotransferase activity* and *N-acetylglucosamine 6-O-sulfotransferase activity*.
OCM	0.41349425	0.102005	16.9236	3.89E-05	Oncomodulin is a high-affinity calcium ion-binding protein. It belongs to the superfamily of calmodulin proteins, also known as the EF-hand proteins.
PTPRS	-0.3103974	0.074553	16.14137	5.88E-05	PTPs are known to be signalling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation.
TRA2A	-0.1048107	0.025705	16.13811	5.89E-05	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.
NUAK1	-0.6623118	0.168239	15.65251	7.61E-05	Serine/threonine-protein kinase involved in various processes such as cell adhesion, regulation of cell ploidy and senescence, cell proliferation and tumour progression.
LPCAT1	-0.1464862	0.037763	15.08258	0.000103	Lysophosphatidylcholine (LPC) acyltransferase (LPCAT; EC 2.3.1.23) catalyses the conversion of LPC to phosphatidylcholine (PC)
PRDM16	-0.9408971	0.28315	15.0561	0.000104	Binds DNA and functions as a transcriptional regulator.
ZSCAN32	-0.0716239	0.018526	14.88124	0.000114	GO annotations related to this gene include sequence-specific DNA binding RNA

					polymerase II transcription factor activity.
DTHD1	-0.4256999	0.111386	14.80765	0.000119	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 7.24*: Differential transcription of transcripts according to atypical severity score from the LUMC cohort. Ten most significant transcripts are shown.*

7.4.7 Pathway analysis of LUMC progression data

I performed pathway analysis using GOrilla using the ranked list of differentially transcribed genes. The top pathway hit from the TRACK-HD GOrilla analysis, "mitotic cell cycle process" was not present in the list of transcripts with a basic p-value of <1x10-3; indeed no cell cycle pathways feature on this list (**Table 7.25**). 20 pathways were significant (FDR q-value <0.05), many of these relating to cellular response to stimuli (**Table 7.25** and **Figure 7.6**).

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

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.*

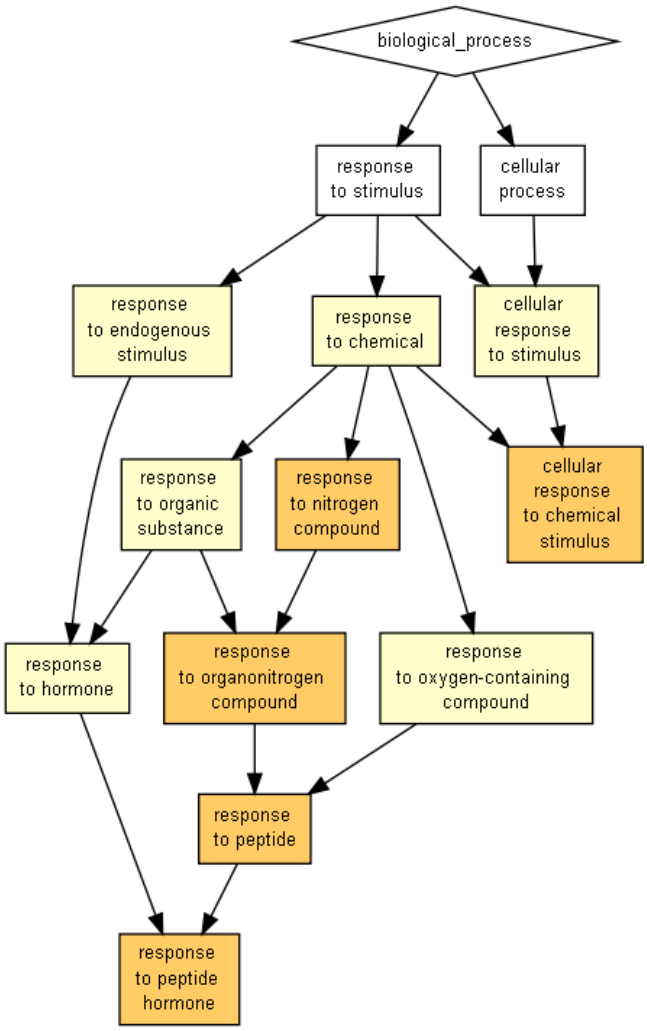


***Figure 7.6**: Pathways related to progression in the LUMC cohort. Diagrammatic illustration using GOrilla(Eden et al., 2009) of the GO terms enriched in the top vs bottom of the list of transcripts in which transcripts are ranked according to differential expression with rate of HD progression in 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 (Mastrokolias 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 if likely to be much less robust and reproducible, such that much larger samples sizes would be needed to use it 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 ($1.58x10^{-8}$) (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**).

Repair pathway	NHEJ	HR	alt-NHEJ/ MMEJ	SSA	ICL repair	SSB repair	BER	TLS	NER	MMR	Expanded TNRs
Source of DNA damage	IR, radiomimetics, Topo II inhibitors	X-linking agents, replication inhibitors, antimetabolites, Topo I inhibitors			X-linking agents	IR, ROS, radiomimetics Topo I inhibitors H_2O_2, alkylating agents	Alkylating agents	UV, alkylating agents	Alkylating agents, X-linkers	DNA Pol proofreading errors	Repeated sequence
Damage sensors	Ku70/Ku80	MRN	PARP	MRN	FA core complex (FANCA, B, C, E, F, G, L and M)	PARP	DNA glycosylases, APE1	PCNA	XPC DDB2 CSA	MSH2, MSH3, MSH6, MLH1, PMS2	MSH3, MLH1, PMS1, PMS2, FAN1, RRM2B
Signaling/ mediator proteins	DNAPK	ATM, ATR, MK2, CtIP, BRCA1/BARD1, BRCA2, PALB2, RPA		CtIP	FANCD1 [BRCA2] D2, I J [BRIP1] N [PALB2] O [RAD51C] P [SLX4]			RAD6 RAD18	XPA, XPF RPA		
Effector proteins	XRCC4 XLF LIG4 APLF Artemis PAXX WRN	RAD51 MUS81/EME1 SLX1/SLX4 RTEL1 BLM TOPOIII POLQ PARI RECQL5 FANCJ, BLM	XRCC1 LIG3, LIG1 CtIP POLQ	RAD52, others?	Shared with HR, TLS, and NER	XRCC1 PNKP POLβ FEN1, TDP1 Aprataxin, LIG1, LIG3A	As for SSB repair	REV1, POLH POLI, POLK	XPG ERCC1 POLE POLD1 LIG1, LIG3	EXO1 POLD LIG1	LIG1, POLD2

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 ($p = 1{\cdot}39 \times 10^{-5}$) was much lower than in TRACK-HD ($p = 5{\cdot}8 \times 10^{-8}$) 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 pathogeneis of many neurodegerative 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).

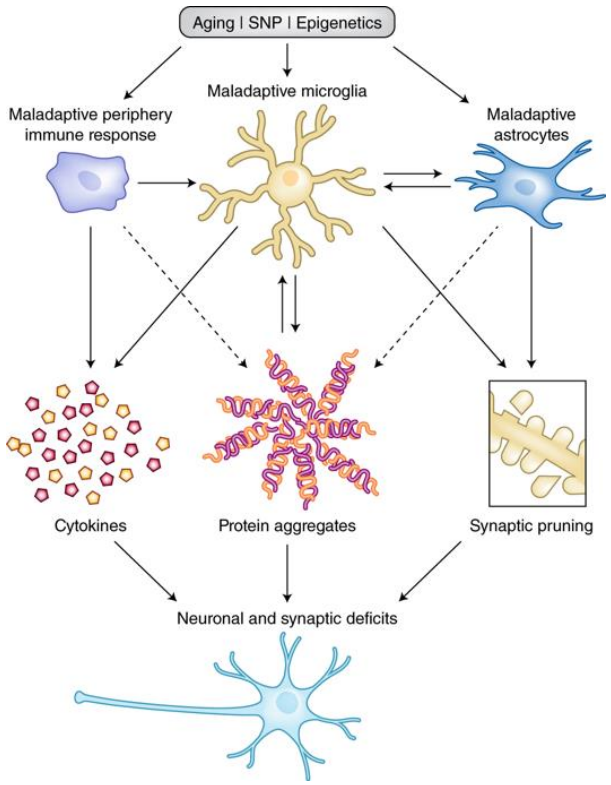


***Figure 8.2**: Innate immune pathways in neurodegenerative diseases. A maladaptive innate immune response has emerged as a critical driving force in the pathogenesis of many neurodegenerative diseases. SNPs on many disease-associated genes induce maladaptive innate immune responses that are also associated 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

1000 GENOMES PROJECT CONSORTIUM 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

ABECASIS, G. 2017. Genotype Imputation Tutorial.

ADVANCED RESEARCH COMPUTING @ CARDIFF (ARCCA). *Introduction to RAVEN* [Online]. accessed: 29/03/2016. Available: http://www.cardiff.ac.uk/arcca/services/equipment/ravenintroduction.html [Accessed 29/03/2016].

ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods,* 7**,** 248-9.

AFFYMETRIX. 2016. *www.affymetrix.com* [Online]. Available: http://www.affymetrix.com/estore/index.jsp [Accessed].

AKIMOTO, C., FORSGREN, L., LINDER, J., BIRVE, A., BACKLUND, I., ANDERSSON, J., NILSSON, A. C., ALSTERMARK, H. & ANDERSEN, P. M. 2012. No GGGGCC-hexanucleotide repeat expansion in C9ORF72 in parkinsonism patients in Sweden. *Amyotroph Lateral Scler*.

ALBERCH, J., LOPEZ, M., BADENAS, C., CARRASCO, J. L., MILA, M., MUNOZ, E. & CANALS, J. M. 2005. Association between BDNF Val66Met polymorphism and age at onset in Huntington disease. *Neurology,* 65**,** 964-5.

ALEXANDER, S. K., RITTMAN, T., XUEREB, J. H., BAK, T. H., HODGES, J. R. & ROWE, J. B. 2014. Validation of the new consensus criteria for the diagnosis of corticobasal degeneration. *Journal of Neurology, Neurosurgery & Psychiatry,* 85**,** 925-929.

ANDERSON, C. A., PETTERSSON, F. H., CLARKE, G. M., CARDON, L. R., MORRIS, A. P. & ZONDERVAN, K. T. 2010. Data quality control in genetic case-control association studies. *Nat. Protocols,* 5**,** 1564-1573.

ANDREW S. E., G. Y. P., KREMER B., SQUITIERI F., THEILMANN J., ZEISLER J., TELENIUS H., ADAM S., ALMQUIST E., ANVRET M., LUCOTTE G., JON STOESSL A., CAMPANELLA G., HAYDEN M. R. 1994. Huntington Disease without CAG Expansion: Phenocopies or Errors in Assignment? *Am J Hum Genet,* 54**,** 852-863.

ANDREW, S. E., GOLDBERG, Y. P., KREMER, B., TELENIUS, H., THEILMANN, J., ADAM, S., STARR, E., SQUITIERI, F., LIN, B., KALCHMAN, M. A. & ET AL. 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nat Genet,* 4**,** 398-403.

ARNOLD, M., RAFFLER, J., PFEUFER, A., SUHRE, K. & KASTENMÜLLER, G. 2015. SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics,* 31**,** 1334-1336.

ASH, P. E., BIENIEK, K. F., GENDRON, T. F., CAULFIELD, T., LIN, W. L., DEJESUS-HERNANDEZ, M., VAN BLITTERSWIJK, M. M., JANSEN-WEST, K., PAUL, J. W., 3RD, RADEMAKERS, R., BOYLAN, K. B., DICKSON, D. W. & PETRUCELLI, L. 2013. Unconventional Translation of C9ORF72 GGGGCC Expansion Generates Insoluble Polypeptides Specific to c9FTD/ALS. *Neuron,* 77**,** 639-646.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J.

E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ATSUTA, N., WATANABE, H., ITO, M., BANNO, H., SUZUKI, K., KATSUNO, M., TANAKA, F., TAMAKOSHI, A. & SOBUE, G. 2006. Natural history of spinal and bulbar muscular atrophy (SBMA): a study of 223 Japanese patients. *Brain,* 129**,** 1446-55.

AZIZ, N. A., VAN DER BURG, J. M. M., TABRIZI, S. J. & LANDWEHRMEYER, G. B. 2018. Overlap between age-at-onset and disease-progression determinants in Huntington disease. *Neurology,* 90**,** 10.1212/WNL.0000000000005690.

BATES, G., TABRIZI, S. & JONES, L. 2014. *Huntington's disease*, Oxford University Press.

BATES, G. P., DORSEY, R., GUSELLA, J. F., HAYDEN, M. R., KAY, C., LEAVITT, B. R., NANCE, M., ROSS, C. A., SCAHILL, R. I., WETZEL, R., WILD, E. J. & TABRIZI, S. J. 2015. Huntington disease. *Nature Reviews Disease Primers***,** 15005.

BATES, G. P., HARPER P. S., JONES, L. (EDITORS) 2002. *Huntington's disease,* Oxford, Oxford University Press.

BECANOVIC, K., NORREMOLLE, A., NEAL, S. J., KAY, C., COLLINS, J. A., ARENILLAS, D., LILJA, T., GAUDENZI, G., MANOHARAN, S., DOTY, C. N., BECK, J., LAHIRI, N., PORTALES-CASAMAR, E., WARBY, S. C., CONNOLLY, C., DE SOUZA, R. A., NETWORK, R. I. O. T. E. H. S. D., TABRIZI, S. J., HERMANSON, O., LANGBEHN, D. R., HAYDEN, M. R., WASSERMAN, W. W. & LEAVITT, B. R. 2015. A SNP in the HTT promoter alters NF-kappaB binding and is a bidirectional genetic modifier of Huntington disease. *Nat Neurosci,* 18**,** 807-16.

BECK, J., POULTER, M., HENSMAN, D., ROHRER, J. D., MAHONEY, C. J., ADAMSON, G., CAMPBELL, T., UPHILL, J., BORG, A., FRATTA, P., ORRELL, R. W., MALASPINA, A., ROWE, J., BROWN, J., HODGES, J., SIDLE, K., POLKE, J. M., HOULDEN, H., SCHOTT, J. M., FOX, N. C., ROSSOR, M. N., TABRIZI, S. J., ISAACS, A. M., HARDY, J., WARREN, J. D., COLLINGE, J. & MEAD, S. 2013. Large C9orf72 Hexanucleotide Repeat Expansions Are Seen in Multiple Neurodegenerative Syndromes and Are More Frequent Than Expected in the UK Population. *Am J Hum Genet,* 92**,** 345-353.

BECK, J., ROHRER, J. D., CAMPBELL, T., ISAACS, A., MORRISON, K. E., GOODALL, E. F., WARRINGTON, E. K., STEVENS, J., REVESZ, T., HOLTON, J., AL-SARRAJ, S., KING, A., SCAHILL, R., WARREN, J. D., FOX, N. C., ROSSOR, M. N., COLLINGE, J. & MEAD, S. 2008. A distinct clinical, neuropsychological and radiological phenotype is associated with progranulin gene mutations in a large UK series. *Brain,* 131**,** 706-20.

BERISA, T. & PICKRELL, J. K. 2016. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics,* 32**,** 283-285.

BETTENCOURT, C., HENSMAN-MOSS, D., FLOWER, M., WIETHOFF, S., BRICE, A., GOIZET, C., STEVANIN, G., KOUTSIS, G., KARADIMA, G., PANAS, M., YESCAS-GOMEZ, P., GARCIA-VELAZQUEZ, L. E., ALONSO-VILATELA, M. E., LIMA, M., RAPOSO, M., TRAYNOR, B., SWEENEY, M., WOOD, N., GIUNTI, P., NETWORK, S., DURR, A., HOLMANS, P., HOULDEN, H., TABRIZI, S. J. & JONES, L. 2016. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann Neurol,* 79**,** 983-90.

BETTENCOURT, C. & LIMA, M. 2011. Machado-Joseph Disease: from first descriptions to new perspectives. *Orphanet J Rare Dis,* 6**,** 35.

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

L. B., SEQUEIROS, J. & LIMA, M. 2011. The APOE epsilon2 allele increases the risk of earlier age at onset in Machado-Joseph disease. *Arch Neurol,* 68**,** 1580-3.

BICHARA, M., WAGNER, J. & LAMBERT, I. B. 2006. Mechanisms of tandem repeat instability in bacteria. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 598**,** 144-163.

BIOCARTA 2016. www.biocarta.com.

BJORKQVIST, M., WILD, E. J., THIELE, J., SILVESTRONI, A., ANDRE, R., LAHIRI, N., RAIBON, E., LEE, R. V., BENN, C. L., SOULET, D., MAGNUSSON, A., WOODMAN, B., LANDLES, C., POULADI, M. A., HAYDEN, M. R., KHALILI-SHIRAZI, A., LOWDELL, M. W., BRUNDIN, P., BATES, G. P., LEAVITT, B. R., MOLLER, T. & TABRIZI, S. J. 2008. A novel pathogenic pathway of immune activation detectable before clinical onset in Huntington's disease. *J Exp Med,* 205**,** 1869-77.

BOEVE, B. F., BOYLAN, K. B., GRAFF-RADFORD, N. R., DEJESUS-HERNANDEZ, M., KNOPMAN, D. S., PEDRAZA, O., VEMURI, P., JONES, D., LOWE, V., MURRAY, M. E., DICKSON, D. W., JOSEPHS, K. A., RUSH, B. K., MACHULDA, M. M., FIELDS, J. A., FERMAN, T. J., BAKER, M., RUTHERFORD, N. J., ADAMSON, J., WSZOLEK, Z. K., ADELI, A., SAVICA, R., BOOT, B., KUNTZ, K. M., GAVRILOVA, R., REEVES, A., WHITWELL, J., KANTARCI, K., JACK, C. R., JR., PARISI, J. E., LUCAS, J. A., PETERSEN, R. C. & RADEMAKERS, R. 2012. Characterization of frontotemporal dementia and/or amyotrophic lateral sclerosis associated with the GGGGCC repeat expansion in C9ORF72. *Brain,* 135**,** 765-83.

BOROVECKI, F., LOVRECIC, L., ZHOU, J., JEONG, H., THEN, F., ROSAS, H. D., HERSCH, S. M., HOGARTH, P., BOUZOU, B., JENSEN, R. V. & KRAINC, D. 2005. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc Natl Acad Sci U S A,* 102**,** 11023-8.

BOTSTEIN, D. & RISCH, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet,* 33 Suppl**,** 228-37.

BOUCHARD, J., TRUONG, J., BOUCHARD, K., DUNKELBERGER, D., DESRAYAUD, S., MOUSSAOUI, S., TABRIZI, S. J., STELLA, N. & MUCHOWSKI, P. J. 2012. Cannabinoid receptor 2 signaling in peripheral immune cells modulates disease onset and severity in mouse models of Huntington's disease. *J Neurosci,* 32**,** 18259-68.

BOYLE, E. A., LI, Y. I. & PRITCHARD, J. K. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell,* 169**,** 1177-1186.

BRAINEAC. 2016. *Braineac - The Brain eQTL Almanac* [Online]. Available: http://www.braineac.org/ [Accessed 21/01/2016].

BRANDT, J. & BUTTERS, N. 1986. The neuropsychological characteristics of Huntington's disease. *Trends in Neurosciences,* 9.

BRAS, J., ALONSO, I., BARBOT, C., COSTA, M. M., DARWENT, L., ORME, T., SEQUEIROS, J., HARDY, J., COUTINHO, P. & GUERREIRO, R. 2015. Mutations in PNKP cause recessive ataxia with oculomotor apraxia type 4. *Am J Hum Genet,* 96**,** 474-9.

BRINKMAN, R. R., MEZEI, M. M., THEILMANN, J., ALMQVIST, E. & HAYDEN, M. R. 1997. The Likelihood of Being Affected with Huntington Disease by a Particular Age, for a Specific CAG Size. *Am J Hum Genet,* 60.

BROWN, A. H. 1975a. Sample sizes required to detect linkage disequilibrium between two or three loci. *Theor Popul Biol,* 8**,** 184-201.

BROWN, M. B. 1975b. A method for combining non-independent, one-sided tests of significance. *Biometrics*, 987-992.

BROWN, M. W., KIM, Y., WILLIAMS, G. M., HUCK, J. D., SURTEES, J. A. & FINKELSTEIN, I. J. 2016. Dynamic DNA binding licenses a repair factor to bypass roadblocks in search of DNA lesions. *Nature Communications,* 7, 10607.

BUDWORTH, H., HARRIS, F. R., WILLIAMS, P., LEE DO, Y., HOLT, A., PAHNKE, J., SZCZESNY, B., ACEVEDO-TORRES, K., AYALA-PENA, S. & MCMURRAY, C. T. 2015. Suppression of Somatic Expansion Delays the Onset of Pathophysiology in a Mouse Model of Huntington's Disease. *PLoS Genet,* 11, e1005267.

BURK, K., GLOBAS, C., BOSCH, S., GRABER, S., ABELE, M., BRICE, A., DICHGANS, J., DAUM, I., KLOCKGETHER T. 1999. Cognitive deficits in spinocerebellar ataxia 2. *Brain,* 122, 769-777.

BUSSE, M. E., HUGHES, G., WILES, C. M. & ROSSER, A. E. 2008. Use of hand-held dynamometry in the evaluation of lower limb muscle strength in people with Huntington's disease. *Journal of Neurology,* 255, 1534-1540.

CAI, C., LANGFELDER, P., FULLER, T. F., OLDHAM, M. C., LUO, R., VAN DEN BERG, L. H., OPHOFF, R. A. & HORVATH, S. 2010. Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics,* 11, 589.

CALABRESI, V., GUIDA, S., SERVADIO, A. & JODICE, C. 2001. Phenotypic effects of expanded ataxin-1 polyglutamines with interruptions in vitro. *Brain Res Bull,* 56, 337-42.

CANNAVO, E., GERRITS, B., MARRA, G., SCHLAPBACH, R. & JIRICNY, J. 2007. Characterization of the interactome of the human MutL homologues MLH1, PMS1, and PMS2. *J Biol Chem,* 282, 2976-86.

CARROLL, J. B., BATES, G. P., STEFFAN, J., SAFT, C. & TABRIZI, S. J. 2015. Treating the whole body in Huntington's disease. *Lancet Neurol,* 14, 1135-42.

CARVALHO, B. S. & IRIZARRY, R. A. 2010. A framework for oligonucleotide microarray preprocessing. *Bioinformatics,* 26, 2363-7.

CHANG, C. C., CHOW, C. C., TELLIER, L., VATTIKUTI, S., PURCELL, S. M. & LEE, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience,* 4.

CHATURVEDI, R. K., CALINGASAN, N. Y., YANG, L., HENNESSEY, T., JOHRI, A. & BEAL, M. F. 2010. Impairment of PGC-1alpha expression, neuropathology and hepatic steatosis in a transgenic mouse model of Huntington's disease following chronic energy deprivation. *Human Molecular Genetics,* 19, 3190-3205.

CHE, H. V., METZGER, S., PORTAL, E., DEYLE, C., RIESS, O. & NGUYEN, H. P. 2011. Localization of sequence variations in PGC-1alpha influence their modifying effect in Huntington disease. *Mol Neurodegener,* 6, 1.

CHOUDHRY, S., MUKERJI, M., SRIVASTAVA, A. K., JAIN, S. & BRAHMACHARI, S. K. 2001. CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. *Hum Mol Gen,* 10, 2437-2446.

CLARK, A. B., VALLE, F., DROTSCHMANN, K., GARY, R. K. & KUNKEL, T. A. 2000. Functional interaction of proliferating cell nuclear antigen with MSH2-MSH6 and MSH2-MSH3 complexes. *J Biol Chem,* 275, 36498-501.

COHEN, J. C., KISS, R. S., PERTSEMLIDIS, A., MARCEL, Y. L., MCPHERSON, R. & HOBBS, H. H. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science,* 305, 869-72.

CONSORTIUM, G. M. O. H. S. D. G.-H. 2015a. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell,* 162, 516-26.

CONSORTIUM, G. O. 2016. *Gene Ontology Consortium* [Online]. Available: http://geneontology.org/ [Accessed 21/01/2016].
CONSORTIUM, G. T. 2015b. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science,* 348**,** 648-60.
CONSORTIUM, H. S. D. G.-H. 2015c. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell,* 162**,** 516-26.
CONSORTIUM, T. G. O. 2015d. Gene Ontology Consortium: going forward. *Nucleic Acids Research,* 43**,** D1049-D1056.
COOPER, G. M. & SHENDURE, J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet,* 12**,** 628-40.
CORVOL, H., BLACKMAN, S. M., BOELLE, P. Y., GALLINS, P. J., PACE, R. G., STONEBRAKER, J. R., ACCURSO, F. J., CLEMENT, A., COLLACO, J. M., DANG, H., DANG, A. T., FRANCA, A., GONG, J., GUILLOT, L., KEENAN, K., LI, W., LIN, F., PATRONE, M. V., RARAIGH, K. S., SUN, L., ZHOU, Y. H., O'NEAL, W. K., SONTAG, M. K., LEVY, H., DURIE, P. R., ROMMENS, J. M., DRUMM, M. L., WRIGHT, F. A., STRUG, L. J., CUTTING, G. R. & KNOWLES, M. R. 2015. Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nat Commun,* 6**,** 8382.
COSTA, V., ANGELINI, C., DE FEIS, I. & CICCODICOLA, A. 2010. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol,* 2010**,** 853916.
CRAUFURD, D. & SNOWDEN, J. 2002. Neuropsychological and neuropsychiatric aspects of Huntington's disease. *In:* BATES, G. P., HARPER, P. & JONES, L. (eds.) *Huntington's Disease.* Oxford: Oxford University Press.
CUI, L., JEONG, H., BOROVECKI, F., PARKHURST, C. N., TANESE, N. & KRAINC, D. 2006. Transcriptional repression of PGC-1alpha by mutant huntingtin leads to mitochondrial dysfunction and neurodegeneration. *Cell,* 127**,** 59-69.
DAUSSET, J., CANN, H., COHEN, D., LATHROP, M., LALOUEL, J. M. & WHITE, R. 1990. Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics,* 6**,** 575-7.
DE KONING, A. P., GU, W., CASTOE, T. A., BATZER, M. A. & POLLOCK, D. D. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet,* 7**,** e1002384.
DE LEEUW, C. A., MOOIJ, J. M., HESKES, T. & POSTHUMA, D. 2015. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol,* 11**,** e1004219.
DEJESUS-HERNANDEZ, M., MACKENZIE, I. R., BOEVE, B. F., BOXER, A. L., BAKER, M., RUTHERFORD, N. J., NICHOLSON, A. M., FINCH, N. A., FLYNN, H., ADAMSON, J., KOURI, N., WOJTAS, A., SENGDY, P., HSIUNG, G. Y., KARYDAS, A., SEELEY, W. W., JOSEPHS, K. A., COPPOLA, G., GESCHWIND, D. H., WSZOLEK, Z. K., FELDMAN, H., KNOPMAN, D. S., PETERSEN, R. C., MILLER, B. L., DICKSON, D. W., BOYLAN, K. B., GRAFF-RADFORD, N. R. & RADEMAKERS, R. 2011. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron,* 72**,** 245-56.
DEJESUS-HERNANDEZ, M., RAYAPROLU, S., SOTO-ORTOLAZA, A. I., RUTHERFORD, N. J., HECKMAN, M. G., TRAYNOR, S., STRONGOSKY, A., GRAFF-RADFORD, N., VAN GERPEN, J., UITTI, R. J., SHIH, J. J., LIN, S. C., WSZOLEK, Z. K., RADEMAKERS, R. & ROSS, O. A. 2012. Analysis of the C9orf72 repeat in Parkinson's disease, essential tremor and restless legs syndrome. *Parkinsonism Relat Disord*.

DELANEAU, O., ZAGURY, J.-F. & MARCHINI, J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods,* 10**,** 5-6.
DELUCA, D. S., LEVIN, J. Z., SIVACHENKO, A., FENNELL, T., NAZAIRE, M. D., WILLIAMS, C., REICH, M., WINCKLER, W. & GETZ, G. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics,* 28**,** 1530-2.
DESAI, A. & GERSON, S. 2014. Exo1 independent DNA mismatch repair involves multiple compensatory nucleases. *DNA Repair (Amst),* 21**,** 55-64.
DEVLIN, B. & ROEDER, K. 1999. Genomic control for association studies. *Biometrics,* 55**,** 997-1004.
DHAENENS, C. M., BURNOUF, S., SIMONIN, C., VAN BRUSSEL, E., DUHAMEL, A., DEFEBVRE, L., DURU, C., VUILLAUME, I., CAZENEUVE, C., CHARLES, P., MAISON, P., DEBRUXELLES, S., VERNY, C., GERVAIS, H., AZULAY, J. P., TRANCHANT, C., BACHOUD-LEVI, A. C., DURR, A., BUEE, L., KRYSTKOWIAK, P., SABLONNIERE, B. & BLUM, D. 2009. A genetic variation in the ADORA2A gene modifies age at onset in Huntington's disease. *Neurobiol Dis,* 35**,** 474-6.
DJOUSSE, L., KNOWLTON, B., HAYDEN, M., ALMQVIST, E. W., BRINKMAN, R., ROSS, C., MARGOLIS, R., ROSENBLATT, A., DURR, A., DODE, C., MORRISON, P. J., NOVELLETTO, A., FRONTALI, M., TRENT, R. J., MCCUSKER, E., GOMEZ-TORTOSA, E., MAYO, D., JONES, R., ZANKO, A., NANCE, M., ABRAMSON, R., SUCHOWERSKY, O., PAULSEN, J., HARRISON, M., YANG, Q., CUPPLES, L. A., GUSELLA, J. F., MACDONALD, M. E. & MYERS, R. H. 2003. Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *Am J Med Genet A,* 119A**,** 279-82.
DJOUSSE, L., KNOWLTON, B., HAYDEN, M. R., ALMQVIST, E. W., BRINKMAN, R. R., ROSS, C. A., MARGOLIS, R. L., ROSENBLATT, A., DURR, A., DODE, C., MORRISON, P. J., NOVELLETTO, A., FRONTALI, M., TRENT, R. J., MCCUSKER, E., GOMEZ-TORTOSA, E., MAYO CABRERO, D., JONES, R., ZANKO, A., NANCE, M., ABRAMSON, R. K., SUCHOWERSKY, O., PAULSEN, J. S., HARRISON, M. B., YANG, Q., CUPPLES, L. A., MYSORE, J., GUSELLA, J. F., MACDONALD, M. E. & MYERS, R. H. 2004. Evidence for a modifier of onset age in Huntington disease linked to the HD gene in 4p16. *Neurogenetics,* 5**,** 109-14.
DRAGILEVA, E., HENDRICKS, A., TEED, A., GILLIS, T., LOPEZ, E. T., FRIEDBERG, E. C., KUCHERLAPATI, R., EDELMANN, W., LUNETTA, K. L., MACDONALD, M. E. & WHEELER, V. C. 2009. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiol Dis,* 33**,** 37-47.
DRUMM, M. L., KONSTAN, M. W., SCHLUCHTER, M. D., HANDLER, A., PACE, R., ZOU, F., ZARIWALA, M., FARGO, D., XU, A., DUNN, J. M., DARRAH, R. J., DORFMAN, R., SANDFORD, A. J., COREY, M., ZIELENSKI, J., DURIE, P., GODDARD, K., YANKASKAS, J. R., WRIGHT, F. A. & KNOWLES, M. R. 2005. Genetic modifiers of lung disease in cystic fibrosis. *N Engl J Med,* 353**,** 1443-53.
DU, J. T., CAMPAU, E., SORAGNI, E., JESPERSEN, C. & GOTTESFELD, J. M. 2013. Length-dependent CTG.CAG triplet-repeat expansion in myotonic dystrophy patient-derived induced pluripotent stem cells. *Human Molecular Genetics,* 22**,** 5276-5287.
DUFF, K., PAULSEN, J. S., BEGLINGER, L. J., LANGBEHN, D. R., WANG, C., STOUT, J. C., ROSS, C. A., AYLWARD, E., CARLOZZI, N. E. & QUELLER, S. 2010. "Frontal" behaviors before the diagnosis of Huntington's disease and their relationship

to markers of disease progression: evidence of early lack of awareness. *J Neuropsychiatry Clin Neurosci,* 22**,** 196-207.
DURR, A. 2010. Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. *Lancet Neurol,* 9**,** 885-94.
DURR, A., STEVANIN, G., CANCEL, G., DUYCKAERTS, C., ABBAS, N., DIDIERJEAN, O., CHNEIWEISS, H., BENOMAR, A., LYON-CAEN, O., JULIEN, J., SERDARU, M., PENET, C., AGID, Y. & BRICE, A. 1996. Spinocerebellar ataxia 3 and Machado-Joseph disease: clinical, molecular, and neuropathological features. *Ann Neurol,* 39**,** 490-9.
DUYAO, M., AMBROSE, C., MYERS, R., NOVELLETTO, A., PERSICHETTI, F., FRONTALI, M., FOLSTEIN, S., ROSS, C., FRANZ, M., ABBOTT, M. & ET AL. 1993. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat Genet,* 4**,** 387-92.
EDELMANN, W., UMAR, A., YANG, K., HEYER, J., KUCHERLAPATI, M., LIA, M., KNEITZ, B., AVDIEVICH, E., FAN, K., WONG, E., CROUSE, G., KUNKEL, T., LIPKIN, M., KOLODNER, R. D. & KUCHERLAPATI, R. 2000. The DNA mismatch repair genes Msh3 and Msh6 cooperate in intestinal tumor suppression. *Cancer Res,* 60**,** 803-7.
EDEN, E., NAVON, R., STEINFELD, I., LIPSON, D. & YAKHINI, Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics,* 10**,** 48.
EHDN. 2018. *About EHDN* [Online]. Available: http://www.ehdn.org/about-ehdn/ [Accessed 15/03/2018 2018].
EPPIG, J. T., BLAKE, J. A., BULT, C. J., KADIN, J. A., RICHARDSON, J. E. & GROUP, T. M. G. D. 2015. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Research,* 43**,** D726-D736.
EVANS, E. & ALANI, E. 2000. Roles for Mismatch Repair Factors in Regulating Genetic Recombination. *Mol Cell Biol,* 20**,** 7839–7844.
FABREGAT, A., SIDIROPOULOS, K., GARAPATI, P., GILLESPIE, M., HAUSMANN, K., HAW, R., JASSAL, B., JUPE, S., KORNINGER, F., MCKAY, S., MATTHEWS, L., MAY, B., MILACIC, M., ROTHFELS, K., SHAMOVSKY, V., WEBBER, M., WEISER, J., WILLIAMS, M., WU, G., STEIN, L., HERMJAKOB, H. & D'EUSTACHIO, P. 2016. The Reactome pathway Knowledgebase. *Nucleic Acids Research,* 44**,** D481-D487.
FEIL, R. & FRAGA, M. F. 2011. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet,* 13**,** 97-109.
FINN, R. D., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., MISTRY, J., MITCHELL, A. L., POTTER, S. C., PUNTA, M., QURESHI, M., SANGRADOR-VEGAS, A., SALAZAR, G. A., TATE, J. & BATEMAN, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res,* 44**,** D279-85.
FISHER, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb. ,* 52**,** 399–433.
FISHER, R. A. 1925. *Statistical Methods for Research Workers*.
FITZPATRICK, A. W. P., FALCON, B., HE, S., MURZIN, A. G., MURSHUDOV, G., GARRINGER, H. J., CROWTHER, R. A., GHETTI, B., GOEDERT, M. & SCHERES, S. H. W. 2017. Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature,* 547**,** 185-190.
FLORES-ROZAS, H., CLARK, D. & KOLODNER, R. D. 2000. Proliferating cell nuclear antigen and Msh2p-Msh6p interact to form an active mispair recognition complex. *Nat Genet,* 26**,** 375-8.

FLOWER, M., LOMEIKAITE, V., CIOSI, M., CUMMING, S., MORALES, F., LO, K. K., HENSMAN MOSS, D. J., JONES, L., HOLMANS, P., INVESTIGATORS, T. T.-H., THE OPTIMISTIC CONSORTIUM, MONCKTON, D. G. & TABRIZI, S. J. 2019. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. *Brain,* 142**,** 1876-1886.

FOIRY, L., DONG, L., SAVOURET, C., HUBERT, L., TE RIELE, H., JUNIEN, C. & GOURDON, G. 2006. Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Hum Genet,* 119**,** 520-6.

FRATTA, P., MIZIELINSKA, S., NICOLL, A. J., ZLOH, M., FISHER, E. M., PARKINSON, G. & ISAACS, A. M. 2012. C9orf72 hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes. *Sci Rep,* 2**,** 1016.

FRATTA, P., POULTER, M., LASHLEY, T., ROHRER, J. D., POLKE, J. M., BECK, J., RYAN, N., HENSMAN, D., MIZIELINSKA, S., WAITE, A. J., LAI, M. C., GENDRON, T. F., PETRUCELLI, L., FISHER, E. M., REVESZ, T., WARREN, J. D., COLLINGE, J., ISAACS, A. M. & MEAD, S. 2013. Homozygosity for the C9orf72 GGGGCC repeat expansion in frontotemporal dementia. *Acta Neuropathol,* 126**,** 401-9.

FRIEDLAND, R. P., SHAH, J. J., FARRER, L. A., VARDARAJAN, B., REBOLLEDO-MENDEZ, J. D., MOK, K. & HARDY, J. 2012. Behavioral variant frontotemporal lobar degeneration with amyotrophic lateral sclerosis with a chromosome 9p21 hexanucleotide repeat. *Front Neurol,* 3**,** 136.

FU, H., HARDY, J. & DUFF, K. E. 2018. Selective vulnerability in neurodegenerative diseases. *Nat Neurosci,* 21**,** 1350-1358.

GACY, A. M., GOELLNER, G., JURANIC, N., MACURA, S. & MCMURRAY, C. T. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell,* 81**,** 533-40.

GAN, L., COOKSON, M. R., PETRUCELLI, L. & LA SPADA, A. R. 2018. Converging pathways in neurodegeneration, from genetics to mechanisms. *Nat Neurosci,* 21**,** 1300-1309.

GANDHI, S. & WOOD, N. W. 2010. Genome-wide association studies: the key to unlocking neurodegeneration? *Nat Neurosci,* 13**,** 789-94.

GATCHEL, J. R. & ZOGHBI, H. Y. 2005. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet,* 6**,** 743-55.

GEM-HD-CONSORTIUM 2015. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell,* 162**,** 516-26.

GENECARDS. *GeneCards Human Gene Database* [Online]. Available: http://www.genecards.org/ [Accessed].

GENTLEMAN RC, C. V., BATES DM, BOLSTAD B, DETTLING M, DUDOIT S, ELLIS B, GAUTIER L, GE Y, GENTRY J, HORNIK K, HOTHORN T, HUBER W, IACUS S, IRIZARRY R, LEISCH F, LI C, MAECHLER M, ROSSINI AJ, SAWITZKI G, SMITH C, SMYTH G, TIERNEY L, YANG JYH, ZHANG J 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*.

GESCHWIND, D. H., PERLMAN, S., FIGUEROA, C. P., TREIMAN, L. J. & PULST, S. M. 1997. The prevalence and wide clinical spectrum of the spinocerebellar ataxia type 2 trinucleotide repeat in patients with autosomal dominant cerebellar ataxia. *Am J Hum Genet,* 60**,** 842-50.

GIAMBARTOLOMEI, C., VUKCEVIC, D., SCHADT, E. E., FRANKE, L., HINGORANI, A. D., WALLACE, C. & PLAGNOL, V. 2014. Bayesian Test for Colocalisation between

Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet,* 10**,** e1004383.

GIBBS, D. L., BARATT, A., BARIC, R. S., KAWAOKA, Y., SMITH, R. D., ORWOLL, E. S., KATZE, M. G. & MCWEENEY, S. K. 2013. Protein co-expression network analysis (ProCoNA). *J Clin Bioinforma,* 3**,** 11.

GIBBS, J. R., VAN DER BRUG, M. P., HERNANDEZ, D. G., TRAYNOR, B. J., NALLS, M. A., LAI, S. L., AREPALLI, S., DILLMAN, A., RAFFERTY, I. P., TRONCOSO, J., JOHNSON, R., ZIELKE, H. R., FERRUCCI, L., LONGO, D. L., COOKSON, M. R. & SINGLETON, A. B. 2010. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet,* 6**,** e1000952.

GIRDEA, M., DUMITRIU, S., FIUME, M., BOWDIN, S., BOYCOTT, K. M., CHÉNIER, S., CHITAYAT, D., FAGHFOURY, H., MEYN, M. S., RAY, P. N., SO, J., STAVROPOULOS, D. J. & BRUDNO, M. 2013. PhenoTips: Patient Phenotyping Software for Clinical and Research Use. *Human Mutation,* 34**,** 1057-1065.

GIUNTI, P., SABBADINI, G., SWEENEY, M. G., DAVIS, M. B., VENEZIANO, L., MANTUANO, E., FEDERICO, A., PLASMATI, R., FRONTALI, M. & WOOD, N. W. 1998. The role of the SCA2 trinucleotide repeat expansion in 89 autosomal dominant cerebellar ataxia families. Frequency, clinical and genetic correlates. *Brain,* 121 (Pt 3)**,** 459-67.

GLOBAS, C., DU MONTCEL, S. T., BALIKO, L., BOESCH, S., DEPONDT, C., DIDONATO, S., DURR, A., FILLA, A., KLOCKGETHER, T., MARIOTTI, C., MELEGH, B., RAKOWICZ, M., RIBAI, P., ROLA, R., SCHMITZ-HUBSCH, T., SZYMANSKI, S., TIMMANN, D., VAN DE WARRENBURG, B. P., BAUER, P. & SCHOLS, L. 2008. Early symptoms in spinocerebellar ataxia type 1, 2, 3, and 6. *Mov Disord,* 23**,** 2232-8.

GODIN, J. D., COLOMBO, K., MOLINA-CALAVITA, M., KERYER, G., ZALA, D., CHARRIN, B. C., DIETRICH, P., VOLVERT, M. L., GUILLEMOT, F., DRAGATSIS, I., BELLAICHE, Y., SAUDOU, F., NGUYEN, L. & HUMBERT, S. 2010. Huntingtin is required for mitotic spindle orientation and mammalian neurogenesis. *Neuron,* 67**,** 392-406.

GOELLNER, E. M., PUTNAM, C. D. & KOLODNER, R. D. 2015. Exonuclease 1-dependent and independent mismatch repair. *DNA Repair (Amst),* 32**,** 24-32.

GOELLNER, G. M., TESTER, D., THIBODEAU, S., ALMQVIST, E., GOLDBERG, Y. P., HAYDEN, M. R. & MCMURRAY, C. T. 1997. Different Mechanisms Underlie DNA Instability in Huntington Disease and Colorectal Cancer. *Am. J. Hum. Genet.,* 60**,** 879-890.

GOLDMAN, J. S., FARMER, J. M., WOOD, E. M., JOHNSON, J. K., BOXER, A., NEUHAUS, J., LOMEN-HOERTH, C., WILHELMSEN, K. C., LEE, V. M., GROSSMAN, M. & MILLER, B. L. 2005. Comparison of family histories in FTLD subtypes and related tauopathies. *Neurology,* 65**,** 1817-9.

GOMES-PEREIRA, M. & MONCKTON, D. G. 2006. Chemical modifiers of unstable expanded simple sequence repeats: what goes up, could come down. *Mutat Res,* 598**,** 15-34.

GOMEZ-NICOLA, D., FRANSEN, N. L., SUZZI, S. & PERRY, V. H. 2013. Regulation of microglial proliferation during chronic neurodegeneration. *J Neurosci,* 33**,** 2481-93.

GONITEL, R., MOFFITT, H., SATHASIVAM, K., WOODMAN, B., DETLOFF, P. J., FAULL, R. L. & BATES, G. P. 2008. DNA instability in postmitotic neurons. *Proc Natl Acad Sci U S A,* 105**,** 3467-72.

GOOLD, R., FLOWER, M., MOSS, D. H., MEDWAY, C., WOOD-KACZMAR, A., ANDRE, R., FARSHIM, P., BATES, G. P., HOLMANS, P., JONES, L. & TABRIZI, S. J. 2019. FAN1 modifies Huntington's disease progression by stabilising the expanded HTT CAG repeat. *Hum Mol Genet,* 28**,** 650-661.

GOULA, A. V., BERQUIST, B. R., WILSON, D. M., 3RD, WHEELER, V. C., TROTTIER, Y. & MERIENNE, K. 2009. Stoichiometry of base excision repair proteins correlates with increased somatic CAG instability in striatum over cerebellum in Huntington's disease transgenic mice. *PLoS Genet,* 5**,** e1000749.

GROUP, H. S. 1996a. Unified Huntington's Disease Rating Scale: reliability and consistency. Huntington Study Group. *Mov Disord,* 11**,** 136-42.

GROUP, H. S. 1996b. Unified Huntington's disease ratinug scale: reliability and consistency. *Movement Disorders,* 11**,** 136-142.

GUERREIRO, R., WOJTAS, A., BRAS, J., CARRASQUILLO, M., ROGAEVA, E., MAJOUNIE, E., CRUCHAGA, C., SASSI, C., KAUWE, J. S., YOUNKIN, S., HAZRATI, L., COLLINGE, J., POCOCK, J., LASHLEY, T., WILLIAMS, J., LAMBERT, J. C., AMOUYEL, P., GOATE, A., RADEMAKERS, R., MORGAN, K., POWELL, J., ST GEORGE-HYSLOP, P., SINGLETON, A. & HARDY, J. 2013. TREM2 variants in Alzheimer's disease. *N Engl J Med,* 368**,** 117-27.

GUERRETTE, S., WILSON, T., GRADIA, S. & FISHEL, R. 1998. Interactions of human hMSH2 with hMSH3 and hMSH2 with hMSH6: examination of mutations found in hereditary nonpolyposis colorectal cancer. *Mol Cell Biol,* 18**,** 6616-23.

GUESELLA, J. F., MACDONALD, M. E. & LEE, J. M. 2014. Genetic modifiers of Huntington's disease. *Mov Disord,* 29**,** 1359-65.

GUNDERSON, K. L., KRUGLYAK, S., GRAIGE, M. S., GARCIA, F., KERMANI, B. G., ZHAO, C., CHE, D., DICKINSON, T., WICKHAM, E., BIERLE, J., DOUCET, D., MILEWSKI, M., YANG, R., SIEGMUND, C., HAAS, J., ZHOU, L., OLIPHANT, A., FAN, J.-B., BARNARD, S. & CHEE, M. S. 2004. Decoding Randomly Ordered DNA Arrays. *Genome Res,* 14**,** 870-877.

GUPTA, S., GELLERT, M. & YANG, W. 2012. Mechanism of mismatch recognition revealed by human MutSbeta bound to unpaired DNA loops. *Nat Struct Mol Biol,* 19**,** 72-8.

GUSELLA, J. F. 1984. Genetic linkage of the Huntington's disease gene to a DNA marker. *Can J Neurol Sci,* 11**,** 421-5.

GUSELLA, J. F., MACDONALD, M. E. & LEE, J. M. 2014. Genetic modifiers of Huntington's disease. *Mov Disord,* 29**,** 1359-65.

HABRAKEN, Y., SUNG, P., PRAKASH, L. & PRAKASH, S. 1997. Enhancement of MSH2-MSH3-mediated mismatch recognition by the yeast MLH1-PMS1 complex. *Current Biology,* 7**,** 790-793.

HALLIDAY, G. M., MCRITCHIE, D. A., MACDONALD, V., DOUBLE, K. L., TRENT, R. J. & MCCUSKER, E. 1998. Regional specificity of brain atrophy in Huntington's disease. *Exp Neurol,* 154**,** 663-72.

HARDING, A. E. 1984. Autosomal dominant cerebellar ataxias. *The Hereditary Ataxias and Related Disorders.* Churchill Livingstone.

HARDY, J. & SINGLETON, A. 2009. Genomewide Association Studies and Human Disease. *N Engl J Med,* 360**,** 1759-68.

HARRELL, F. E. 2001. *Regression Modeling Strategies,* New York, Wiley.

HASHIDA, H., GOTO, J., KURISAKI, H., MIZUSAWA, H. & KANAZAWA, I. 1997. Brain regional differences in the expansion of a CAG repeat in the spinocerebellar

ataxias: dentatorubral-pallidoluysian atrophy, Machado-Joseph disease, and spinocerebellar ataxia type 1. *Ann Neurol,* 41**,** 505-11.

HATCHI, E., SKOURTI-STATHAKI, K., VENTZ, S., PINELLO, L., YEN, A., KAMIENIARZ-GDULA, K., DIMITROV, S., PATHANIA, S., MCKINNEY, K. M., EATON, M. L., KELLIS, M., HILL, S. J., PARMIGIANI, G., PROUDFOOT, N. J. & LIVINGSTON, D. M. 2015. BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. *Mol Cell,* 57**,** 636-47.

HAUGEN, A. C., GOEL, A., YAMADA, K., MARRA, G., NGUYEN, T. P., NAGASAKA, T., KANAZAWA, S., KOIKE, J., KIKUCHI, Y., ZHONG, X., ARITA, M., SHIBUYA, K., OSHIMURA, M., HEMMI, H., BOLAND, C. R. & KOI, M. 2008. Genetic instability caused by loss of MutS homologue 3 in human colorectal cancer. *Cancer Res,* 68**,** 8465-72.

HAYES, S., TURECKI, G., BRISEBOIS, K., LOPES-CENDES, I., GASPAR, C., RIESS, O., RANUM, L. P., PULST, S. M. & ROULEAU, G. A. 2000. CAG repeat length in RAI1 is associated with age at onset variability in spinocerebellar ataxia type 2 (SCA2). *Hum Mol Genet,* 9**,** 1753-8.

HENSMAN MOSS, D. J., FLOWER, M. D., LO, K. K., MILLER, J. R., VAN OMMEN, G. B., T HOEN, P. A., STONE, T. C., GUINEE, A., LANGBEHN, D. R., JONES, L., PLAGNOL, V., VAN ROON-MOM, W. M., HOLMANS, P. & TABRIZI, S. J. 2017a. Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. *Sci Rep,* 7**,** 44849.

HENSMAN MOSS, D. J., PARDIÑAS, A. F., LANGBEHN, D., LO, K. K., LEAVITT, B. R., RA, R., DURR, A., MEAD, S., INVESTIGATORS, T.-H., INVESTIGATORS, R., HOLMANS, P., JONES, L. & TABRIZI, S. J. 2017b. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *The Lancet Neurology,* 16**,** 701-711.

HENSMAN MOSS, D. J., POULTER, M., BECK, J., HEHIR, J., POLKE, J. M., CAMPBELL, T., ADAMSON, G., MUDANOHWO, E., MCCOLGAN, P., HAWORTH, A., WILD, E. J., SWEENEY, M. G., HOULDEN, H., MEAD, S. & TABRIZI, S. J. 2014. C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. *Neurology,* 82**,** 292-9.

HODGES, A. 2006. Regional and cellular gene expression changes in human Huntington's disease brain. *Human Molecular Genetics,* 15**,** 965-977.

HODGES, A., STRAND, A. D., ARAGAKI, A. K., KUHN, A., SENGSTAG, T., HUGHES, G., ELLISTON, L. A., HARTOG, C., GOLDSTEIN, D. R., THU, D., HOLLINGSWORTH, Z. R., COLLIN, F., SYNEK, B., HOLMANS, P. A., YOUNG, A. B., WEXLER, N. S., DELORENZI, M., KOOPERBERG, C., AUGOOD, S. J., FAULL, R. L., OLSON, J. M., JONES, L. & LUTHI-CARTER, R. 2006. Regional and cellular gene expression changes in human Huntington's disease brain. *Hum Mol Genet,* 15**,** 965-77.

HOGARTH, P., KAYSON, E., KIEBURTZ, K., MARDER, K., OAKES, D., ROSAS, D., SHOULSON, I., WEXLER, N. S., YOUNG, A. B. & ZHAO, H. 2005. Interrater agreement in the assessment of motor manifestations of Huntington's disease. *Mov Disord,* 20**,** 293-7.

HOLBERT, S., DENGHIEN, I., KIECHLE, T., ROSENBLATT, A., WELLINGTON, C., HAYDEN, M. R., MARGOLIS, R. L., ROSS, C. A., DAUSSET, J., FERRANTE, R. J. & NERI, C. 2001. The Gln-Ala repeat transcriptional activator CA150 interacts with huntingtin: neuropathologic and genetic evidence for a role in Huntington's disease pathogenesis. *Proc Natl Acad Sci U S A,* 98**,** 1811-6.

HOLMANS, P. 2010. Statistical Methods for Pathway Analysis of Genome-Wide Data for Association with Complex Genetic Traits. *In:* DUNLAP, J. C. M., JASON H (ed.) *Advances in Genetics- Computational Methods for Genetics of Complex Traits.* Burlington: Academic Press.

HOLMANS, P., GREEN, E. K., PAHWA, J. S., FERREIRA, M. A., PURCELL, S. M., SKLAR, P., OWEN, M. J., O'DONOVAN, M. C. & CRADDOCK, N. 2009. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet,* 85**,** 13-24.

HONG, S., BEJA-GLASSER, V. F., NFONOYIM, B. M., FROUIN, A., LI, S., RAMAKRISHNAN, S., MERRY, K. M., SHI, Q., ROSENTHAL, A., BARRES, B. A., LEMERE, C. A., SELKOE, D. J. & STEVENS, B. 2016a. Complement and microglia mediate early synapse loss in Alzheimer mouse models. *Science*.

HONG, S., DISSING-OLESEN, L. & STEVENS, B. 2016b. New insights on the role of microglia in synaptic pruning in health and disease. *Curr Opin Neurobiol,* 36**,** 128-34.

HORVATH, S., LANGFELDER, P., KWAK, S., AARONSON, J., ROSINSKI, J., VOGT, T. F., ESZES, M., FAULL, R. L. M., CURTIS, M. A., WALDVOGEL, H. J., CHOI, O.-W., TUNG, S., VINTERS, H. V., COPPOLA, G. & YANG, X. W. 2016. Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. *Aging,* 8**,** 1485-1512.

HORVATH, S., ZHANG, Y., LANGFELDER, P., KAHN, R. S., BOKS, M. P., VAN EIJK, K., VAN DEN BERG, L. H. & OPHOFF, R. A. 2012. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol,* 13**,** R97.

HOWIE, B., FUCHSBERGER, C., STEPHENS, M., MARCHINI, J. & ABECASIS, G. R. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics,* 44**,** 955-959.

HUANG, M. & D'ANDREA, A. D. 2010. A new nuclease member of the FAN club. *Nat Struct Mol Biol,* 17**,** 926-8.

HUGHES, A. & JONES, L. 2014. Pathogenic Mechanisms. *In:* BATES, G. P., TABRIZI, S. J. & JONES, L. (eds.) *Huntington's Disease.* 4th Edition ed. Oxford: OUP.

HUNTINGTON'S, DISEASE, COLLABORATIVE, RESEARCH & GROUP 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell,* 72**,** 971-83.

ILLUMINA. 2014. *TruSeq(R) RNA Sample Preparation v2 Guide* [Online]. Available: http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqrna/truseq-rna-sample-prep-v2-guide-15026495-f.pdf [Accessed 12/01/2016].

ILLUMINA. 2015. *An introduction to Next-Generation Sequencing Technology* [Online]. Available: https://emea.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf [Accessed].

ILLUMINA. 2017. *Omni2.5-8 Kit product information* [Online]. Available: https://emea.illumina.com/products/by-type/microarray-kits/infinium-omni25-8.html [Accessed 26/10/2017 2017].

ILLUMINA. 2018a. *HiSeq 2000* [Online]. Available: https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf [Accessed 28/03/2018 2018].

ILLUMINA 2018b. Nextera DNA Exome. *In:* ILLUMINA (ed.).
INC., S. I. 2013. SAS/STAT 13.1 User's Guide: High-Performance Procedures. . *In:* INC., S. I. (ed.). NC.
INSTITUTE, N. C. last updated: 18 September 2012. *Home : Pathway Interaction Database* [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/ [Accessed].
INTERNATIONAL GENOMICS OF ALZHEIMER'S DISEASE, C. 2015. Convergent genetic and expression data implicate immunity in Alzheimer's disease. *Alzheimers Dement,* 11**,** 658-71.
IYER, R. R., PLUCIENNIK, A., NAPIERALA, M. & WELLS, R. D. 2015. DNA Triplet Repeat Expansion and Mismatch Repair. *Annu Rev Biochem,* 84**,** 199-226.
JAYADEV, S. & BIRD, T. D. 2013. Hereditary ataxias: overview. *Genet Med,* 15**,** 673-83.
JIN, H. & CHO, Y. 2017. Structural and functional relationships of FAN1. *DNA Repair (Amst)*.
JIRICNY, J. 2006. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol,* 7**,** 335-46.
JOLLIFFE, I. T. & CADIMA, J. 2016. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci,* 374**,** 20150202.
JONES, L., HOULDEN, H. & TABRIZI, S. J. 2017. DNA repair in the trinucleotide repeat disorders. *The Lancet Neurology,* 16**,** 88-96.
JONSON, I., OUGLAND, R. & LARSEN, E. 2013. DNA repair mechanisms in Huntington's disease. *Mol Neurobiol,* 47**,** 1093-102.
KALLBERG, M., WANG, H., WANG, S., PENG, J., WANG, Z., LU, H. & XU, J. 2012. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc,* 7**,** 1511-22.
KÄLLBERG, M., WANG, H., WANG, S., PENG, J., WANG, Z., LU, H. & XU, J. 2012. Template-based protein structure modeling using the RaptorX web server. *Nat. Protocols,* 7**,** 1511-1522.
KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.
KANEHISA, M., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research,* 44**,** D457-D462.
KANTARTZIS, A., WILLIAMS, G. M., BALAKRISHNAN, L., ROBERTS, R. L., SURTEES, J. A. & BAMBARA, R. A. 2012. Msh2-Msh3 interferes with Okazaki fragment processing to promote trinucleotide repeat expansions. *Cell Rep,* 2**,** 216-22.
KARCH, C. M., WEN, N., FAN, C. C., YOKOYAMA, J. S., KOURI, N., ROSS, O. A., HOGLINGER, G., MULLER, U., FERRARI, R., HARDY, J., SCHELLENBERG, G. D., SLEIMAN, P. M., MOMENI, P., HESS, C. P., MILLER, B. L., SHARMA, M., VAN DEERLIN, V., SMELAND, O. B., ANDREASSEN, O. A., DALE, A. M. & DESIKAN, R. S. 2018. Selective Genetic Overlap Between Amyotrophic Lateral Sclerosis and Diseases of the Frontotemporal Dementia Spectrum. *JAMA Neurol*.
KATHIRESAN, S., WILLER, C. J., PELOSO, G. M., DEMISSIE, S., MUSUNURU, K., SCHADT, E. E., KAPLAN, L., BENNETT, D., LI, Y., TANAKA, T., VOIGHT, B. F., BONNYCASTLE, L. L., JACKSON, A. U., CRAWFORD, G., SURTI, A., GUIDUCCI, C., BURTT, N. P., PARISH, S., CLARKE, R., ZELENIKA, D., KUBALANZA, K. A., MORKEN, M. A., SCOTT, L. J., STRINGHAM, H. M., GALAN, P., SWIFT, A. J., KUUSISTO, J., BERGMAN, R. N., SUNDVALL, J., LAAKSO, M., FERRUCCI, L., SCHEET, P., SANNA, S., UDA, M., YANG, Q., LUNETTA, K. L., DUPUIS, J., DE

BAKKER, P. I., O'DONNELL, C. J., CHAMBERS, J. C., KOONER, J. S., HERCBERG, S., MENETON, P., LAKATTA, E. G., SCUTERI, A., SCHLESSINGER, D., TUOMILEHTO, J., COLLINS, F. S., GROOP, L., ALTSHULER, D., COLLINS, R., LATHROP, G. M., MELANDER, O., SALOMAA, V., PELTONEN, L., ORHO-MELANDER, M., ORDOVAS, J. M., BOEHNKE, M., ABECASIS, G. R., MOHLKE, K. L. & CUPPLES, L. A. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet,* 41**,** 56-65.

KAWAI, Y., SUENAGA, M., WATANABE, H. & SOBUE, G. 2009. Cognitive impairment in spinocerebellar degeneration. *Eur Neurol,* 61**,** 257-68.

KAY, C., COLLINS, J. A., MIEDZYBRODZKA, Z., MADORE, S. J., GORDON, E. S., GERRY, N., DAVIDSON, M., SLAMA, R. A. & HAYDEN, M. R. 2016a. Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology*.

KAY, C., TIRADO-HURTADO, I., CORNEJO-OLIVAS, M., COLLINS, J. A., WRIGHT, G., INCA-MARTINEZ, M., VELIZ-OTANI, D., KETELAAR, M. E., SLAMA, R. A., ROSS, C. J., MAZZETTI, P. & HAYDEN, M. R. 2016b. The targetable A1 Huntington disease haplotype has distinct Amerindian and European origins in Latin America. *Eur J Hum Genet*.

KEGG. 2016. *KEGG: Kyoto Encyclopedia of Genes and Genomes* [Online]. Available: http://www.kegg.jp/ [Accessed].

KENNEDY, L., EVANS, E., CHEN, C. M., CRAVEN, L., DETLOFF, P. J., ENNIS, M. & SHELBOURNE, P. F. 2003. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum Mol Genet,* 12**,** 3359-67.

KENNEDY, L. & SHELBOURNE, P. F. 2000. Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? *Hum Mol Gen,* 9**,** 2539-2344.

KENNEDY, W. R., ALTER, M. & SUNG, J. H. 1968. Progressive proximal spinal and bulbar muscular atrophy of late onset. A sex-linked recessive trait. *Neurology,* 18**,** 671-80.

KEOGH, R., FROST, C., OWEN, G., DANIEL, R. M., LANGBEHN, D. R., LEAVITT, B., DURR, A., ROOS, R. A., LANDWEHRMEYER, G. B., REILMANN, R., BOROWSKY, B., STOUT, J., CRAUFURD, D. & TABRIZI, S. J. 2016. Medication Use in Early-HD Participants in Track-HD: an Investigation of its Effects on Clinical Performance. *PLoS Curr,* 8.

KEUM, J. W., SHIN, A., GILLIS, T., MYSORE, J. S., ABU ELNEEL, K., LUCENTE, D., HADZI, T., HOLMANS, P., JONES, L., ORTH, M., KWAK, S., MACDONALD, M. E., GUSELLA, J. F. & LEE, J. M. 2016. The HTT CAG-Expansion Mutation Determines Age at Death but Not Disease Duration in Huntington Disease. *Am J Hum Genet,* 98**,** 287-98.

KHAN, N., KOLIMI, N. & RATHINAVELAN, T. 2015. Twisting Right to Left: A...A Mismatch in a CAG Trinucleotide Repeat Overexpansion Provokes Left-Handed Z-DNA Conformation. *PLoS Comput Biol,* 11**,** e1004162.

KIERNAN, M. C., VUCIC, S., CHEAH, B. C., TURNER, M. R., EISEN, A., HARDIMAN, O., BURRELL, J. R. & ZOING, M. C. 2011. Amyotrophic lateral sclerosis. *Lancet,* 377**,** 942-55.

KIEZUN, A., GARIMELLA, K., DO, R., STITZIEL, N. O., NEALE, B. M., MCLAREN, P. J., GUPTA, N., SKLAR, P., SULLIVAN, P. F., MORAN, J. L., HULTMAN, C. M., LICHTENSTEIN, P., MAGNUSSON, P., LEHNER, T., SHUGART, Y. Y., PRICE, A. L.,

DE BAKKER, P. I., PURCELL, S. M. & SUNYAEV, S. R. 2012. Exome sequencing and the genetic basis of complex traits. *Nat Genet,* 44**,** 623-30.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol,* 14**,** R36.

KIM, J. I., LONG, J. D., MILLS, J. A., MCCUSKER, E. & PAULSEN, J. S. 2015. Multivariate Clustering of Progression Profiles Reveals Different Depression Patterns in Prodromal Huntington Disease. *Neuropsychology*.

KLECZKOWSKA, H. E., MARRA, G., LETTIERI, T. & JIRICNY, J. 2001. hMSH3 and hMSH6 interact with PCNA and colocalize with it to replication foci. *Genes Dev,* 15**,** 724-36.

KLEIN, R. J., ZEISS, C., CHEW, E. Y., TSAI, J.-Y., SACKLER, R. S., HAYNES, C., HENNING, A. K., SANGIOVANNI, J. P., MANE, S. M., MAYNE, S. T., BRACKEN, M. B., FERRIS, F. L., OTT, J., BARNSTABLE, C. & HOH, J. 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* 308**,** 385-388.

KÖHLER, S., VASILEVSKY, N., ENGELSTAD, M., FOSTER, E. & MCMURRAY 2017. The Human Phenotype Ontology in 2017. *Nucl. Acids Res.*

KOVALENKO, M., DRAGILEVA, E., ST CLAIRE, J., GILLIS, T., GUIDE, J. R., NEW, J., DONG, H., KUCHERLAPATI, R., KUCHERLAPATI, M. H., EHRLICH, M. E., LEE, J. M. & WHEELER, V. C. 2012. Msh2 acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. *PLoS One,* 7**,** e44273.

KRATZ, K., SCHOPF, B., KADEN, S., SENDOEL, A., EBERHARD, R., LADEMANN, C., CANNAVO, E., SARTORI, A. A., HENGARTNER, M. O. & JIRICNY, J. 2010. Deficiency of FANCD2-associated nuclease KIAA1018/FAN1 sensitizes cells to interstrand crosslinking agents. *Cell,* 142**,** 77-88.

KUMAR, P., HENIKOFF, S. & NG, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc,* 4**,** 1073-81.

KWAN, W., MAGNUSSON, A., CHOU, A., ADAME, A., CARSON, M. J., KOHSAKA, S., MASLIAH, E., MÖLLER, T., RANSOHOFF, R., TABRIZI, S. J., BJÖRKQVIST, M. & MUCHOWSKI, P. J. 2012a. Bone Marrow Transplantation Confers Modest Benefits in Mouse Models of Huntington's Disease. *The Journal of Neuroscience,* 32**,** 133-142.

KWAN, W., TRAGER, U., DAVALOS, D., CHOU, A., BOUCHARD, J., ANDRE, R., MILLER, A., WEISS, A., GIORGINI, F., CHEAH, C., MOLLER, T., STELLA, N., AKASSOGLOU, K., TABRIZI, S. J. & MUCHOWSKI, P. J. 2012b. Mutant huntingtin impairs immune cell migration in Huntington disease. *J Clin Invest,* 122**,** 4737-47.

KWAN, W., TRÄGER, U., DAVALOS, D., CHOU, A., BOUCHARD, J., ANDRE, R., MILLER, A., WEISS, A., GIORGINI, F., CHEAH, C., MÖLLER, T., STELLA, N., AKASSOGLOU, K., TABRIZI, S. J. & MUCHOWSKI, P. J. 2012c. Mutant huntingtin impairs immune cell migration in Huntington disease. *Journal of Clinical Investigation,* 122**,** 4737-4747.

LABADORF, A., HOSS, A. G., LAGOMARSINO, V., LATOURELLE, J. C., HADZI, T. C., BREGU, J., MACDONALD, M. E., GUSELLA, J. F., CHEN, J.-F., AKBARIAN, S., WENG, Z. & MYERS, R. H. 2015a. RNA Sequence Analysis of Human Huntington Disease Brain Reveals an Extensive Increase in Inflammatory and Developmental Gene Expression. *PLOS ONE,* 10**,** e0143563.

LABADORF, A., HOSS, A. G., LAGOMARSINO, V., LATOURELLE, J. C., HADZI, T. C., BREGU, J., MACDONALD, M. E., GUSELLA, J. F., CHEN, J. F., AKBARIAN, S., WENG, Z. & MYERS, R. H. 2015b. RNA Sequence Analysis of Human Huntington Disease Brain Reveals an Extensive Increase in Inflammatory and Developmental Gene Expression. *PLoS One,* 10**,** e0143563.

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

LANDWEHRMEYER, G. B., FITZER-ATTAS, C. J., GIULIANO, J. D., GONÇALVES, N., ANDERSON, K. E., CARDOSO, F., FERREIRA, J. J., MESTRE, T. A., STOUT, J. C. & SAMPAIO, C. 2016. Data Analytics from Enroll-HD, a Global Clinical Research Platform for Huntington's Disease. *Movement Disorders Clinical Practice*.

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

LANGBEHN, D. R., BRINKMAN, R. R., FALUSH, D., PAULSEN, J. S. & HAYDEN, M. R. 2004. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin Genet,* 65**,** 267-77.

LANGFELDER, P., GAO, F., WANG, N., HOWLAND, D., KWAK, S., VOGT, T. F., AARONSON, J. S., ROSINSKI, J., COPPOLA, G., HORVATH, S. & YANG, X. W. 2018. MicroRNA signatures of endogenous Huntingtin CAG repeat expansion in mice. *PLoS One,* 13**,** e0190550.

LANGFELDER, P. & HORVATH, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* 9**,** 559.

LANSKA, D. J., LAVINE, L., LANSKA, M. J. & SCHOENBERG, B. S. 1988. Huntington's disease mortality in the United States. *Neurology,* 38**,** 769-72.

LARREA, A. A., LUJAN, S. A. & KUNKEL, T. A. 2010. SnapShot: DNA mismatch repair. *Cell,* 141**,** 730 e1.

LASHLEY, T., HARDY, J. & ISAACS, A. M. 2013. RANTing about C9orf72. *Neuron,* 77**,** 597-8.

LEE, J., HWANG, Y. J., KIM, K. Y., KOWALL, N. W. & RYU, H. 2013. Epigenetic Mechanisms of Neurodegeneration in Huntington's Disease. *Neurotherapeutics*.

LEE, J. H., LEE, J. M., RAMOS, E. M., GILLIS, T., MYSORE, J. S., KISHIKAWA, S., HADZI, T., HENDRICKS, A. E., HAYDEN, M. R., MORRISON, P. J., NANCE, M., ROSS, C. A., MARGOLIS, R. L., SQUITIERI, F., GELLERA, C., GOMEZ-TORTOSA, E., AYUSO, C., SUCHOWERSKY, O., TRENT, R. J., MCCUSKER, E., NOVELLETTO, A., FRONTALI, 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., MARDER, K., SEQUEIROS, J., LANDWEHRMEYER, G. B., SHOULSON, I., MYERS, R. H., MACDONALD, M. E. & GUSELLA, J. F. 2012a. TAA repeat variation in the GRIK2 gene does not influence age at onset in Huntington's disease. *Biochem Biophys Res Commun,* 424**,** 404-8.

LEE, J. M., CHAO, M. J., HAROLD, D., ABU ELNEEL, K., GILLIS, T., HOLMANS, P., JONES, L., ORTH, M., MYERS, R. H., KWAK, S., WHEELER, V. C., MACDONALD, M. E. & GUSELLA, J. F. 2017. A modifier of Huntington's disease onset at the MLH1 locus. *Hum Mol Genet,* 26**,** 3859-3867.

LEE, J. M., GILLIS, T., MYSORE, J. S., RAMOS, E. M., MYERS, R. H., HAYDEN, M. R., MORRISON, P. J., NANCE, M., ROSS, C. A., MARGOLIS, R. L., SQUITIERI, F., GRIGUOLI, A., DI DONATO, S., GOMEZ-TORTOSA, E., AYUSO, C., SUCHOWERSKY, O., TRENT, R. J., MCCUSKER, E., NOVELLETTO, A., FRONTALI,

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., MARDER, K., SEQUEIROS, J., MACDONALD, M. E. & GUSELLA, J. F. 2012b. Common SNP-based haplotype analysis of the 4p16.3 Huntington disease gene region. *Am J Hum Genet,* 90**,** 434-44.

LEEK, J. T. 2014. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res,* 42.

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. & EXOME AGGREGATION, C. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-91.

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 of parkinsonism. *Brain,* 136**,** 385-91.

LEVINE, T. P., DANIELS, R. D., GATTA, A. T., WONG, L. H. & HAYES, M. J. 2013. The product of C9orf72, a gene strongly implicated in neurodegeneration, is structurally related to DENN Rab-GEFs. *Bioinformatics,* 29**,** 499-503.

LGC. 2018. *How does KASP work* [Online]. Available: https://www.lgcgroup.com/kasp/#.WrPzknnLhEZ [Accessed 28/03/2018 2018].

LI, Y., WILLER, C., SANNA, S. & ABECASIS, G. 2009. Genotype imputation. *Annu Rev Genomics Hum Genet,* 10**,** 387-406.

LI, Y. I., VAN DE GEIJN, B., RAJ, A., KNOWLES, D. A., PETTI, A. A., GOLAN, D., GILAD, Y. & PRITCHARD, J. K. 2016. RNA splicing is a primary link between genetic variation and disease. *Science,* 352**,** 600-4.

LING, S. C., POLYMENIDOU, M. & CLEVELAND, D. W. 2013. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron,* 79**,** 416-38.

LIU, T., GHOSAL, G., YUAN, J., CHEN, J. & HUANG, J. 2010. FAN1 acts with FANCI-FANCD2 to promote DNA interstrand cross-link repair. *Science,* 329**,** 693-6.

LIU, Y. & WILSON, S. H. 2012. DNA base excision repair: a mechanism of trinucleotide repeat expansion. *Trends Biochem Sci,* 37**,** 162-72.

LOCKE, D. P., HILLIER, L. W., WARREN, W. C., WORLEY, K. C., NAZARETH, L. V., MUZNY, 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.,

LUNTER, G., MEADER, S., HEGER, A., PONTING, C. P., MARQUES-BONET, T., ALKAN, C., CHEN, L., CHENG, Z., KIDD, J. M., EICHLER, E. E., WHITE, S., SEARLE, S., VILELLA, A. J., CHEN, Y., FLICEK, P., MA, J., RANEY, B., SUH, B., BURHANS, R., HERRERO, J., HAUSSLER, D., FARIA, R., FERNANDO, O., DARRE, F., FARRE, D., GAZAVE, E., OLIVA, M., NAVARRO, A., ROBERTO, R., CAPOZZI, O., ARCHIDIACONO, N., DELLA VALLE, G., PURGATO, S., ROCCHI, M., KONKEL, M. K., WALKER, J. A., ULLMER, B., BATZER, M. A., SMIT, A. F., HUBLEY, R., CASOLA, C., SCHRIDER, D. R., HAHN, M. W., QUESADA, V., PUENTE, X. S., ORDONEZ, G. R., LOPEZ-OTIN, C., VINAR, T., BREJOVA, B., RATAN, A., HARRIS, R. S., MILLER, W., KOSIOL, C., LAWSON, H. A., TALIWAL, V., MARTINS, A. L., SIEPEL, A., ROYCHOUDHURY, A., MA, X., DEGENHARDT, J., BUSTAMANTE, C. D., GUTENKUNST, R. N., MAILUND, T., DUTHEIL, J. Y., HOBOLTH, A., SCHIERUP, M. H., RYDER, O. A., YOSHINAGA, Y., DE JONG, P. J., WEINSTOCK, G. M., ROGERS, J., MARDIS, E. R., GIBBS, R. A., et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature,* 469**,** 529-33.

LONG, J. D., LANGBEHN, D. R., TABRIZI, S. J., LANDWEHRMEYER, B. G., PAULSEN, J. S., WARNER, J. & SAMPAIO, C. 2017. Validation of a prognostic index for Huntington's disease. *Mov Disord,* 32**,** 256-263.

LONG, J. D., PAULSEN, J. S., MARDER, K., ZHANG, Y., KIM, J. I. & MILLS, J. A. 2013. Tracking motor impairments in the progression of Huntington's disease. *Mov Disord,* 29**,** 311-319.

LOPEZ CASTEL, A., CLEARY, J. D. & PEARSON, C. E. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol,* 11**,** 165-70.

LOPEZ CASTEL, A., TOMKINSON, A. E. & PEARSON, C. E. 2009. CTG/CAG repeat instability is modulated by the levels of human DNA ligase I and its interaction with proliferating cell nuclear antigen: a distinction between replication and slipped-DNA repair. *J Biol Chem,* 284**,** 26631-45.

LORENZETTI, D., BOHLEGA, S. & ZOGHBI, H. Y. 1997. The expansion of the CAG repeat in ataxin-2 is a frequent cause of autosomal dominant spinocerebellar ataxia. *Neurology,* 49**,** 1009-13.

LOUIS, E. D., LEE, P., QUINN, L. & MARDER, K. 1999. Dystonia in Huntington's disease: prevalence and clinical characteristics. *Mov Disord,* 14**,** 95-101.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15**,** 550.

LOVESTONE, S., HODGSON, S., SHAM, P., DIFFER, A. M. & LEVY, R. 1996. Familial psychiatric presentation of Huntington's disease. *J Med Genet,* 33**,** 128-31.

LOVRECIC, L., KASTRIN, A., KOBAL, J., PIRTOSEK, Z., KRAINC, D. & PETERLIN, B. 2009. Gene expression changes in blood as a putative biomarker for Huntington's disease. *Mov Disord,* 24**,** 2277-81.

LU, M., BOSCHETTI, C. & TUNNACLIFFE, A. 2015. Long Term Aggresome Accumulation Leads to DNA Damage, p53-dependent Cell Cycle Arrest, and Steric Interference in Mitosis. *J Biol Chem,* 290**,** 27986-8000.

LYNDAKER, A. M. & ALANI, E. 2009. A tale of tails: insights into the coordination of 3′ end processing during homologous recombination. *BioEssays,* 31**,** 315-321.

MACKAY, C., DECLAIS, A. C., LUNDIN, C., AGOSTINHO, A., DEANS, A. J., MACARTNEY, T. J., HOFMANN, K., GARTNER, A., WEST, S. C., HELLEDAY, T., LILLEY, D. M. & ROUSE, J. 2010. Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. *Cell,* 142**,** 65-76.

MAGAZI, D. S., KRAUSE, A., BONEV, V., MOAGI, M., IQBAL, Z., DLUDLA, M. & VAN DER MEYDEN, C. H. 2008. Huntington's disease: genetic heterogeneity in black African patients. *S Afr Med J,* 98**,** 200-3.

MAHER 2008. Personal genomes: The case of the missing heritability. *Nature,* 456**,** 18-21.

MAHONEY, C. J., BECK, J., ROHRER, J. D., LASHLEY, T., MOK, K., SHAKESPEARE, T., YEATMAN, T., WARRINGTON, E. K., SCHOTT, J. M., FOX, N. C., ROSSOR, M. N., HARDY, J., COLLINGE, J., REVESZ, T., MEAD, S. & WARREN, J. D. 2012. Frontotemporal dementia with the C9ORF72 hexanucleotide repeat expansion: clinical, neuroanatomical and neuropathological features. *Brain,* 135**,** 736-50.

MAJEWSKI, J. & PASTINEN, T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet,* 27**,** 72-9.

MAJEWSKI, J., SCHWARTZENTRUBER, J., LALONDE, E., MONTPETIT, A. & JABADO, N. 2011. What can exome sequencing do for you? *J Med Genet,* 48**,** 580-9.

MAJOUNIE, E., ABRAMZON, Y., RENTON, A. E., KELLER, M. F., TRAYNOR, B. J. & SINGLETON, A. B. 2012. Large C9orf72 repeat expansions are not a common cause of Parkinson's disease. *Neurobiol Aging,* 33**,** 2527 e1-2.

MALKKI, M. & PETERSDORF, E. W. 2012. Genotyping of single nucleotide polymorphisms by 5' nuclease allelic discrimination. *Methods in molecular biology (Clifton, N.J.),* 882**,** 173-182.

MANLEY, K., SHIRLEY, T. L., FLAHERTY, L. & MESSER, A. 1999. *Msh2* deficiency prevents *in vivo* somatic instability of the CAG repeat in Huntington disease transgenei mice. *Nat. Gen.,* 23**,** 471-3.

MANOLIO, T. A. 2017. A decade of shared genomic associations. *Nature,* 546**,** 360-361.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A. & VISSCHER, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

MARTIN, S. A., LORD, C. J. & ASHWORTH, A. 2010. Therapeutic targeting of the DNA mismatch repair pathway. *Clin Cancer Res,* 16**,** 5107-13.

MASON, A. G., TOME, S., SIMARD, J. P., LIBBY, R. T., BAMMLER, T. K., BEYER, R. P., MORTON, A. J., PEARSON, C. E. & LA SPADA, A. R. 2014. Expression levels of DNA replication and repair genes predict regional somatic repeat instability in the brain but are not altered by polyglutamine disease protein expression or age. *Hum Mol Genet,* 23**,** 1606-18.

MASSEY, T. & JONES, L. 2018. The central role of DNA damage and repair in CAG repeat diseases. *Disease Models & Mechanisms,* 11.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. 2006. Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics,* 7**,** 29-59.

MASTROKOLIAS, A., ARIYUREK, Y., GOEMAN, J. J., VAN DUIJN, E., ROOS, R. A., VAN DER MAST, R. C., VAN OMMEN, G. B., DEN DUNNEN, J. T., T HOEN, P. A. & VAN ROON-MOM, W. M. 2015. Huntington's disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. *Eur J Hum Genet*.

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.
MATSUURA, T., SASAKI, H., YABE, I., HAMADA, K., HAMADA, T., SHITARA, M. & TASHIRO, K. 1999. Mosaicism of unstable CAG repeats in the brain of spinocerebellar ataxia type 2. *J Neurol,* 246**,** 835-9.
MCGOLDRICK, P., ZHANG, M., VAN BLITTERSWIJK, M., SATO, C., MORENO, D., XIAO, S., ZHANG, A. B., MCKEEVER, P. M., WEICHERT, A., SCHNEIDER, R., KEITH, J., PETRUCELLI, L., RADEMAKERS, R., ZINMAN, L., ROBERTSON, J. & ROGAEVA, E. 2018. Unaffected mosaic C9orf72 case: RNA foci, dipeptide proteins, but upregulated C9orf72 expression. *Neurology,* 90**,** e323-e331.
MCKINNON, P. J. 2009. DNA repair deficiency and neurological disease. *Nat Rev Neurosci,* 10**,** 100-12.
MCMURRAY, C. T. 2010. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet,* 11**,** 786-99.
MENON, R. P., NETHISINGHE, S., FAGGIANO, S., VANNOCCI, T., REZAEI, H., PEMBLE, S., SWEENEY, M. G., WOOD, N. W., DAVIS, M. B., PASTORE, A. & GIUNTI, P. 2013. The role of interruptions in polyQ in the pathology of SCA1. *PLoS Genet,* 9**,** e1003648.
METZGER, S., BAUER, P., TOMIUK, J., LACCONE, F., DIDONATO, S., GELLERA, C., SOLIVERI, P., LANGE, H. W., WEIRICH-SCHWAIGER, H., WENNING, G. K., MELEGH, B., HAVASI, V., BALIKO, L., WIECZOREK, S., ARNING, L., ZAREMBA, J., SULEK, A., HOFFMAN-ZACHARSKA, D., BASAK, A. N., ERSOY, N., ZIDOVSKA, J., KEBRDLOVA, V., PANDOLFO, M., RIBAI, P., KADASI, L., KVASNICOVA, M., WEBER, B. H., KREUZ, F., DOSE, M., STUHRMANN, M. & RIESS, O. 2006. The S18Y polymorphism in the UCHL1 gene is a genetic modifier in Huntington's disease. *Neurogenetics,* 7**,** 27-30.
METZGER, S., SAUKKO, M., VAN CHE, H., TONG, L., PUDER, Y., RIESS, O. & NGUYEN, H. P. 2010. Age at onset in Huntington's disease is modified by the autophagy pathway: implication of the V471A polymorphism in Atg7. *Hum Genet,* 128**,** 453-9.
METZGER, S., WALTER, C., RIESS, O., ROOS, R. A., NIELSEN, J. E., CRAUFURD, D., NETWORK, R. I. O. T. E. H. S. D. & NGUYEN, H. P. 2013. The V471A polymorphism in autophagy-related gene ATG7 modifies age at onset specifically in Italian Huntington disease patients. *PLoS One,* 8**,** e68951.
MGI. 2016. *MGI-Mouse Genome Informatics-The international database resource for the laboratory mouse* [Online]. Available: http://www.informatics.jax.org/ [Accessed].
MI, H., MURUGANUJAN, A., CASAGRANDE, J. T. & THOMAS, P. D. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protocols,* 8**,** 1551-1566.
MICHEGAN, U. O. Minimac Tutorial.
MIHM, M. J., AMANN, D. M., SCHANBACHER, B. L., ALTSCHULD, R. A., BAUER, J. A. & HOYT, K. R. 2007. Cardiac dysfunction in the R6/2 mouse model of Huntington's disease. *Neurobiology of Disease,* 25**,** 297-308.
MILLER, J. R., LO, K. K., ANDRE, R., HENSMAN MOSS, D. J., TRAGER, U., STONE, T. C., JONES, L., HOLMANS, P., PLAGNOL, V. & TABRIZI, S. J. 2016a. RNA-Seq of Huntington's disease patient myeloid cells reveals innate transcriptional dysregulation associated with proinflammatory pathway activation. *Hum Mol Genet*.

MILLER, J. R., LO, K. K., ANDRE, R., HENSMAN MOSS, D. J., TRAGER, U., STONE, T. C., JONES, L., HOLMANS, P., PLAGNOL, V. & TABRIZI, S. J. 2016b. RNA-Seq of Huntington's disease patient myeloid cells reveals innate transcriptional dysregulation associated with proinflammatory pathway activation. *Hum Mol Genet,* 25**,** 2893-2904.

MINA, E., VAN ROON-MOM, W. M., HETTNE, K. M., VAN ZWET, E., GOEMAN, J. J., NERI, C., MONS, B., T'HOEN, P. A. C. & ROOS, M. 2016. Common disease signatures between blood and brain in Huntington's Disease. *Orphanet Journal of Rare Diseases*.

MIRET, J. J., MILLA, M. G. & LAHUE, R. S. 1993. Characterization of a DNA mismatch-binding activity in yeast extracts. *J Biol Chem,* 268**,** 3507-13.

MIRKIN, S. M. 2007. Expandable DNA repeats and human disease. *Nature,* 447**,** 932-40.

MIRKIN, S. M. 2013. Driving past four-strand snags. *Nature,* 497**,** 449-450.

MOCHEL, F., DURANT, B., MENG, X., O'CALLAGHAN, J., YU, H., BROUILLET, E., WHEELER, V. C., HUMBERT, S., SCHIFFMANN, R. & DURR, A. 2012. Early alterations of brain cellular energy homeostasis in Huntington disease models. *J Biol Chem,* 287**,** 1361-70.

MODRICH, P. 2006. Mechanisms in eukaryotic mismatch repair. *J Biol Chem,* 281**,** 30305-9.

MONTANINI, L., FERRARI, S., CRAFA, P., GHIRARDINI, S., PONZIN, D., ORSONI, J. G. & MORA, P. 2013. Human RNA integrity after postmortem retinal tissue recovery. *Ophthalmic Genet,* 34**,** 27-31.

MOONEY, M. A. & WILMOT, B. 2015. Gene set analysis: A step-by-step guide. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics,* 168**,** 517-527.

MOORE R. C., X. F., MONAGHAN J., HAN D., ZHANG Z., EDSTROM L., ANVRET M., PRUSINER S. B. 2001. Huntington's Disease Phenocopy Is a Familiar Prion Disease. *Am J Hum Genet,* 69**,** 1385-1388.

MORALES, F., VASQUEZ, M., SANTAMARIA, C., CUENCA, P., CORRALES, E. & MONCKTON, D. G. 2016. A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair (Amst),* 40**,** 57-66.

MORI, K., WENG, S.-M., T., A., S., M., K., R., E., K., B., S., A., K. H., M., C., VAN BROECKHOVEN, C., HAASS, C. & EDBAUER, D. 2013. The *C9orf72* GGGCC Repeat Is Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS. *Science,* 339**,** 1335-8.

MORRIS, H. R., WAITE, A. J., WILLIAMS, N. M., NEAL, J. W. & BLAKE, D. J. 2012. Recent advances in the genetics of the ALS-FTLD complex. *Curr Neurol Neurosci Rep,* 12**,** 243-50.

MORTON, N. E. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet,* 7**,** 277-318.

MOSKVINA, V., O'DUSHLAINE, C., PURCELL, S., CRADDOCK, N., HOLMANS, P. & O'DONOVAN, M. C. 2011. Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. *Genet Epidemiol,* 35**,** 861-6.

MOSS, D. J. H., PARDINAS, A. F., LANGBEHN, D., LO, K., LEAVITT, B. R., ROOS, R., DURR, A., MEAD, S., INVESTIGATORS, T.-H., INVESTIGATORS, R., HOLMANS, P., JONES, L. & TABRIZI, S. J. 2017. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol*.

MURRAY, M. E., DEJESUS-HERNANDEZ, M., RUTHERFORD, N. J., BAKER, M., DUARA, R., GRAFF-RADFORD, N. R., WSZOLEK, Z. K., FERMAN, T. J., JOSEPHS, K. A., BOYLAN, K. B., RADEMAKERS, R. & DICKSON, D. W. 2011. Clinical and neuropathologic heterogeneity of c9FTD/ALS associated with hexanucleotide repeat expansion in C9ORF72. *Acta Neuropathol,* 122**,** 673-90.

NAKAJIMA, E., ORIMO, H., IKEJIMA, M. & SHIMADA, T. 1995. Nine-bp repeat polymorphism in exon 1 of the hMSH3 gene. *J Hum Genet,* 40**,** 343-345.

NAZE, P., VUILLAUME, I., DESTEE, A., PASQUIER, F. & SABLONNIERE, B. 2002. Mutation analysis and association studies of the ubiquitin carboxy-terminal hydrolase L1 gene in Huntington's disease. *Neurosci Lett,* 328**,** 1-4.

NEIL, A. J., KIM, J. C. & MIRKIN, S. M. 2017. Precarious maintenance of simple DNA repeats in eukaryotes. *Bioessays,* 39.

NELSON, M. R., TIPNEY, H., PAINTER, J. L., SHEN, J., NICOLETTI, P., SHEN, Y., FLORATOS, A., SHAM, P. C., LI, M. J., WANG, J., CARDON, L. R., WHITTAKER, J. C. & SANSEAU, P. 2015. The support of human genetic evidence for approved drug indications. *Nat Genet,* 47**,** 856-60.

NEUEDER, A. & BATES, G. P. 2014. A common gene expression signature in Huntington's disease patient brain regions. *BMC Med Genomics,* 7**,** 60.

NEW, L., LIU, K. & CROUSE, G. F. 1993. The yeast gene MSH3 defines a new class of eukaryotic MutS homologues. *Mol Gen Genet,* 239**,** 97-108.

NEYMAN, J. & PEARSON, E. S. 1933. On the problem of the most efficient tests of statistical hypotheses. . *Philos Trans R Soc Lond A,* 231**,** 289–337

NG, L., HUANG, STOCKWELL, WALENZ, LI, AXELROD, BUSAM, STRAUSBERG, VENTER 2008. Genetic Variation in an Individual Human Exome. *PLOS Genetics*.

NICA, A. C. & DERMITZAKIS, E. T. 2013. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci,* 368**,** 20120362.

NISHIMURA, D. 2001. BioCarta. *Biotech Software & Internet Report,* 2**,** 117-120.

NOVELLETTO, A., PERSICHETTI, F., SABBADINI, G., MANDICH, P., BELLONE, E., AJMAR, F., SQUITIERI, F., CAMPANELLA, G., BOZZA, A., MACDONALD, M. E. & ET AL. 1994. Polymorphism analysis of the huntingtin gene in Italian families affected with Huntington disease. *Hum Mol Genet,* 3**,** 1129-32.

O'DONNELL, L. & DUROCHER, D. 2010. DNA repair has a new FAN1 club. *Mol Cell,* 39**,** 167-9.

O'DOWD, S., CURTIN, D., WAITE, A. J., ROBERTS, K., PENDER, N., REID, V., O'CONNELL, M., WILLIAMS, N. M., MORRIS, H. R., TRAYNOR, B. J. & LYNCH, T. 2012. C9ORF72 expansion in amyotrophic lateral sclerosis/frontotemporal dementia also causes parkinsonism. *Mov Disord,* 27**,** 1072-4.

OHSU. 2017. *Illumina Bead Arrays* [Online]. Available: http://www.ohsu.edu/xd/research/research-cores/gene-profiling-shared-resource/project-design/array-technology/illumina-bead-arrays.cfm [Accessed 2017].

OLMOS-ALONSO, A., SCHETTERS, S. T., SRI, S., ASKEW, K., MANCUSO, R., VARGAS-CABALLERO, M., HOLSCHER, C., PERRY, V. H. & GOMEZ-NICOLA, D. 2016.

Pharmacological targeting of CSF1R inhibits microglial proliferation and prevents the progression of Alzheimer's-like pathology. *Brain,* 139**,** 891-907.
OMIM. 2017a. *ATAXIA-TELANGIECTASIA-LIKE DISORDER 1; ATLD1* [Online]. Available: https://www.omim.org/entry/604391 [Accessed 01/08/2017 2017].
OMIM. 2017b. *SECKEL SYNDROME 1; SCKL1* [Online]. Available: https://www.omim.org/entry/210600?search=seckel%20syndrome%201&highlight=1%20syndromic%20syndrome%20seckel [Accessed 01/08/2017 2017].
ORTH, M., COOPER, J. M., BATES, G. P. & SCHAPIRA, A. H. V. 2003. Inclusion formation in Huntington's disease R6/2 mouse muscle cultures. *Journal of Neurochemistry,* 87**,** 1-6.
ORTH, M., HANDLEY, O. J., SCHWENKE, C., DUNNETT, S. B., CRAUFURD, D., HO, A. K., WILD, E., TABRIZI, S. J. & LANDWEHRMEYER, G. B. 2010. Observing Huntington's Disease: the European Huntington's Disease Network's REGISTRY. *PLoS Curr,* 2**,** RRN1184.
OWEN, B. A., YANG, Z., LAI, M., GAJEC, M., BADGER, J. D., 2ND, HAYES, J. J., EDELMANN, W., KUCHERLAPATI, R., WILSON, T. M. & MCMURRAY, C. T. 2005. (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nat Struct Mol Biol,* 12**,** 663-70.
PAMPHLETT, R., CHEONG, P. L., TRENT, R. J. & YU, B. 2012. Transmission of C9orf72 hexanucleotide repeat expansions in sporadic amyotrophic lateral sclerosis: an Australian trio study. *Neuroreport,* 23**,** 556-9.
PANAS, M., AVRAMOPOULOS, D., KARADIMA, G., PETERSEN, M. B. & VASSILOPOULOS, D. 1999. Apolipoprotein E and presenilin-1 genotypes in Huntington's disease. *J Neurol,* 246**,** 574-7.
PANTHER. 2016. *PANTHER - Gene List Analysis* [Online]. Available: http://pantherdb.org/ [Accessed].
PATTISON, J. S., SANBE, A., MALOYAN, A., OSINSKA, H., KLEVITSKY, R. & ROBBINS, J. 2008. Cardiomyocyte Expression of a Polyglutamine Preamyloid Oligomer Causes Heart Failure. *Circulation,* 117**,** 2743-2751.
PAULL, T. T. 2015. Mechanisms of ATM Activation. *Annu Rev Biochem,* 84**,** 711-38.
PAULSEN, J. S. 2011. Cognitive impairment in Huntington disease: diagnosis and treatment. *Curr Neurol Neurosci Rep,* 11**,** 474-83.
PAULSEN, J. S., LANGBEHN, D. R., STOUT, J. C., AYLWARD, E., ROSS, C. A., NANCE, M., GUTTMAN, M., JOHNSON, S., MACDONALD, M., BEGLINGER, L. J., DUFF, K., KAYSON, E., BIGLAN, K., SHOULSON, I., OAKES, D. & HAYDEN, M. 2008. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *J Neurol Neurosurg Psychiatry,* 79**,** 874-80.
PAULSEN, J. S., NEHL, C., HOTH, K. F., KANZ, J. E., BENJAMIN, M., CONYBEARE, R., MCDOWELL, B. & TURNER, B. 2005. Depression and stages of Huntington's disease. *J Neuropsychiatry Clin Neurosci,* 17**,** 496-502.
PAULSEN, J. S., READY, R. E., HAMILTON, J. M., MEGA, M. S. & CUMMINGS, J. L. 2001a. Neuropsychiatric aspects of Huntington's disease. *J Neurol Neurosurg Psychiatry,* 71**,** 310-4.
PAULSEN, J. S., ZHAO, H., STOUT, J. C., BRINKMAN, R. R., GUTTMAN, M., ROSS, C. A., COMO, P., MANNING, C., HAYDEN, M. R. & SHOULSON, I. 2001b. Clinical markers of early disease in persons near onset of Huntington's disease. *Neurology,* 57**,** 658-662.

PEARL, L. H., SCHIERZ, A. C., WARD, S. E., AL-LAZIKANI, B. & PEARL, F. M. 2015. Therapeutic opportunities within the DNA damage response. *Nat Rev Cancer,* 15**,** 166-80.

PEARSON, C. E., NICHOL EDAMURA, K. & CLEARY, J. D. 2005. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet,* 6**,** 729-42.

PENNEY, J. B., JR., VONSATTEL, J. P., MACDONALD, M. E., GUSELLA, J. F. & MYERS, R. H. 1997. CAG repeat number governs the development rate of pathology in Huntington's disease. *Ann Neurol,* 41**,** 689-92.

PENNEY, J. B. V. J.-P. M. M. E. G. J. P. M. R. H. 1997. CAG Repeat Number Governs the Development Rate of Pathology in Hungtington's Disease. *Annals of Neurology,* 41.

PERSICHETTI, F., SRINIDHI, J., KANALEY, L., GE, P., MYERS, R. H., D'ARRIGO, K., BARNES, G. T., MACDONALD, M. E., VONSATTEL, J. P., GUSELLA, J. F. & ET AL. 1994. Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. *Neurobiol Dis,* 1**,** 159-66.

PERSICHETTI F., S. J., KANALEY L., GE P., MYERS R. H., D'ARRIGO K., BARNES G. T., MACDONALD M. E., VONSATTEL J-P., GUSELLA J. F., BIRD E. D. 1994. Huntingtton's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. *Neurobiology of Disease,* 1**,** 159-166.

PICKRELL, J. K. 2014. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet,* 94**,** 559-73.

PICKRELL, J. K., BERISA, T., LIU, J. Z., SEGUREL, L., TUNG, J. Y. & HINDS, D. A. 2016. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet,* 48**,** 709-717.

PINTO, R. M., DRAGILEVA, E., KIRBY, A., LLORET, A., LOPEZ, E., ST CLAIRE, J., PANIGRAHI, G. B., HOU, C., HOLLOWAY, K., GILLIS, T., GUIDE, J. R., COHEN, P. E., LI, G. M., PEARSON, C. E., DALY, M. J. & WHEELER, V. C. 2013. Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS Genet,* 9**,** e1003930.

PLENGE, R. M. 2017. *Omnigenic drug discovery* [Online]. Available: https://www.plengegen.com/blog/omnigenic/ [Accessed 2017].

PLENGE, R. M., SCOLNICK, E. M. & ALTSHULER, D. 2013. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov,* 12**,** 581-94.

PLUCIENNIK, A., BURDETT, V., BAITINGER, C., IYER, R. R., SHI, K. & MODRICH, P. 2013. Extrahelical (CAG)/(CTG) triplet repeat elements support proliferating cell nuclear antigen loading and MutLalpha endonuclease activation. *Proc Natl Acad Sci U S A,* 110**,** 12277-82.

POLLARD, K. S., HUBISZ, M. J., ROSENBLOOM, K. R. & SIEPEL, A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res,* 20**,** 110-21.

POTTER, N. T. 1996. The relationship between (CAG)n repeat number and age of onset in a family with dentatorubral-pallidoluysian atrophy (DRPLA): diagnostic implications of confirmatory and predictive testing. *J Med Genet,* 33**,** 168-70.

PULST, S. M. 1999. Genetic linkage analysis. *Arch Neurol,* 56**,** 667-72.

PULST, S. M., SANTOS, N., WANG, D., YANG, H., HUYNH, D., VELAZQUEZ, L. & FIGUEROA, K. P. 2005. Spinocerebellar ataxia type 2: polyQ repeat variation in the CACNA1A calcium channel modifies age of onset. *Brain,* 128**,** 2297-303.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK:

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

QUARRELL, O. W., RIGBY, A. S., BARRON, L., CROW, Y., DALTON, A., DENNIS, N., FRYER, A. E., HEYDON, F., KINNING, E., LASHWOOD, A., LOSEKOOT, M., MARGERISON, L., MCDONNELL, S., MORRISON, P. J., NORMAN, A., PETERSON, M., RAYMOND, F. L., SIMPSON, S., THOMPSON, E. & WARNER, J. 2007. Reduced penetrance alleles for Huntington's disease: a multi-centre direct observational study. *Journal of Medical Genetics,* 44**,** e68.

QUARRELL, O. W. J. 2014. Juvenile Huntington's disease. *In:* BATES, G. P., TABRIZI, S. J. & JONES, L. (eds.) *Huntington's Disease.* Oxford University Press.

RAMASAMY, A., TRABZUNI, D., GUELFI, S., VARGHESE, V., SMITH, C., WALKER, R., DE, T., CONSORTIUM, U. K. B. E., NORTH AMERICAN BRAIN EXPRESSION, C., COIN, L., DE SILVA, R., COOKSON, M. R., SINGLETON, A. B., HARDY, J., RYTEN, M. & WEALE, M. E. 2014. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat Neurosci,* 17**,** 1418-1428.

RAMOS, E. M., LATOURELLE, J. C., LEE, J. H., GILLIS, T., MYSORE, J. S., SQUITIERI, F., DI PARDO, A., DI DONATO, S., HAYDEN, M. R., MORRISON, P. J., NANCE, M., ROSS, C. A., MARGOLIS, R. L., GOMEZ-TORTOSA, E., AYUSO, C., SUCHOWERSKY, O., TRENT, R. J., MCCUSKER, E., NOVELLETTO, A., FRONTALI, M., JONES, R., ASHIZAWA, T., FRANK, S., SAINT-HILAIRE, M. H., HERSCH, S. M., ROSAS, H. D., LUCENTE, D., HARRISON, M. B., ZANKO, A., MARDER, K., GUSELLA, J. F., LEE, J. M., ALONSO, I., SEQUEIROS, J., MYERS, R. H. & MACDONALD, M. E. 2012. Population stratification may bias analysis of PGC-1alpha as a modifier of age at Huntington disease motor onset. *Hum Genet*.

RANUM, L. P., CHUNG, M. Y., BANFI, S., BRYER, A., SCHUT, L. J., RAMESAR, R., DUVICK, L. A., MCCALL, A., SUBRAMONY, S. H., GOLDFARB, L. & ET AL. 1994. Molecular and clinical correlations in spinocerebellar ataxia type I: evidence for familial effects on the age at onset. *Am J Hum Genet,* 55**,** 244-52.

RAVINA, B., ROMER, M., CONSTANTINESCU, R., BIGLAN, K., BROCHT, A., KIEBURTZ, K., SHOULSON, I. & MCDERMOTT, M. P. 2008. The relationship between CAG repeat length and clinical progression in Huntington's disease. *Mov Disord,* 23**,** 1223-7.

RAWLINS, M. 2010a. Huntington's disease out of the closet? *The Lancet,* 376**,** 1372-1373.

RAWLINS, M. 2010b. Huntington's disease out of the closet? *Lancet,* 376**,** 1372-3.

REACTOME. 2016. *Reactome Pathway Database* [Online]. Available: http://www.reactome.org/ [Accessed].

REDDY, K., ZAMIRI, B., STANLEY, S. Y., MACGREGOR, R. B. & PEARSON, C. E. 2013. The disease-associated r(GGGGCC)n repeat from the C9ORF72 gene forms tract length-dependent uni- and multi-molecular RNA G-quadruplex structures. *J Biol Chem*.

REEDEKER N, B. J., VAN DUIJN E, GILTAY EJ, ROOS RAD, ROOS C, VAN DER MAST RC 2011. Incidence, Course, and Predictors of Apathy in Huntington's Disease: A Two-Year Prospective Study. *J Neuropsychiatry Clin Neurosci,* 23**,** 434-441.

RENTON, A. E., MAJOUNIE, E., WAITE, A., SIMON-SANCHEZ, J., ROLLINSON, S., GIBBS, J. R., SCHYMICK, J. C., LAAKSOVIRTA, H., VAN SWIETEN, J. C., MYLLYKANGAS, L., KALIMO, H., PAETAU, A., ABRAMZON, Y., REMES, A. M., KAGANOVICH, A., SCHOLZ, S. W., DUCKWORTH, J., DING, J., HARMER, D. W., HERNANDEZ, D. G., JOHNSON, J. O., MOK, K., RYTEN, M., TRABZUNI, D., GUERREIRO, R. J., ORRELL,

R. W., NEAL, J., MURRAY, A., PEARSON, J., JANSEN, I. E., SONDERVAN, D., SEELAAR, H., BLAKE, D., YOUNG, K., HALLIWELL, N., CALLISTER, J. B., TOULSON, G., RICHARDSON, A., GERHARD, A., SNOWDEN, J., MANN, D., NEARY, D., NALLS, M. A., PEURALINNA, T., JANSSON, L., ISOVIITA, V. M., KAIVORINNE, A. L., HOLTTA-VUORI, M., IKONEN, E., SULKAVA, R., BENATAR, M., WUU, J., CHIO, A., RESTAGNO, G., BORGHERO, G., SABATELLI, M., HECKERMAN, D., ROGAEVA, E., ZINMAN, L., ROTHSTEIN, J. D., SENDTNER, M., DREPPER, C., EICHLER, E. E., ALKAN, C., ABDULLAEV, Z., PACK, S. D., DUTRA, A., PAK, E., HARDY, J., SINGLETON, A., WILLIAMS, N. M., HEUTINK, P., PICKERING-BROWN, S., MORRIS, H. R., TIENARI, P. J. & TRAYNOR, B. J. 2011. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron,* 72**,** 257-68.

REVA, B. A., ANTIPIN, Y. A. & SANDER, C. 2007. Determinants of protein function revealed by combinatorial entropy opitimization. *Genome Biol,* 8**,** R232.

RHODES, L. E., FREEMAN, B. K., AUH, S., KOKKINIS, A. D., LA PEAN, A., CHEN, C., LEHKY, T. J., SHRADER, J. A., LEVY, E. W., HARRIS-LOVE, M., DI PROSPERO, N. A. & FISCHBECK, K. H. 2009. Clinical features of spinal and bulbar muscular atrophy. *Brain,* 132**,** 3242-51.

ROGACHEVA, M. V., MANHART, C. M., CHEN, C., GUARNE, A., SURTEES, J. & ALANI, E. 2014. Mlh1-Mlh3, a Meiotic Crossover and DNA Mismatch Repair Factor, Is a Msh2-Msh3-stimulated Endonuclease. *Journal of Biological Chemistry,* 289**,** 5664-5673.

ROLFS, A., KOEPPEN, A. H., BAUER, I., BAUER, P., BUHLMANN, S., TOPKA, H., SCHOLS, L. & RIESS, O. 2003. Clinical features and neuropathology of autosomal dominant spinocerebellar ataxia (SCA17). *Ann Neurol,* 54**,** 367-75.

ROOS, R. A. 2014. Clinical Neurology. *In:* BATES, G. P., TABRIZI, S. J. & JONES, L. (eds.) *Huntington's Disease.* 4th Edition ed. Oxford Oxford.

ROSAS, H. D., REUTER, M., DOROS, G., LEE, S. Y., TRIGGS, T., MALARICK, K., FISCHL, B., SALAT, D. H. & HERSCH, S. M. 2011. A tale of two factors: what determines the rate of progression in Huntington's disease? A longitudinal MRI study. *Mov Disord,* 26**,** 1691-7.

ROSENBLATT, A., KUMAR, B. V., MO, A., WELSH, C. S., MARGOLIS, R. L. & ROSS, C. A. 2012. Age, CAG repeat length, and clinical progression in Huntington's disease. *Mov Disord,* 27**,** 272-6.

ROSHYARA, N. R. & SCHOLZ, M. 2014. fcGENE: a versatile tool for processing and transforming SNP datasets. *PloS one,* 9**,** e97589.

ROSS, C. A., AYLWARD, E. H., WILD, E. J., LANGBEHN, D. R., LONG, J. D., WARNER, J. H., SCAHILL, R. I., LEAVITT, B. R., STOUT, J. C., PAULSEN, J. S., REILMANN, R., UNSCHULD, P. G., WEXLER, A., MARGOLIS, R. L. & TABRIZI, S. J. 2014. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nat Rev Neurol,* 10**,** 204-16.

ROSS, C. A. & TABRIZI, S. J. 2011. Huntington's disease: from molecular pathogenesis to clinical treatment. *The Lancet Neurology,* 10**,** 83-98.

ROUBROEKS, J. A. Y., SMITH, R. G., VAN DEN HOVE, D. L. A. & LUNNON, K. 2017. Epigenetics and DNA methylomic profiling in Alzheimer's disease and other neurodegenerative diseases. *J Neurochem,* 143**,** 158-170.

RUB, U., SCHOLS, L., PAULSON, H., AUBURGER, G., KERMER, P., JEN, J. C., SEIDEL, K., KORF, H. W. & DELLER, T. 2013. Clinical features, neurogenetics and

neuropathology of the polyglutamine spinocerebellar ataxias type 1, 2, 3, 6 and 7. *Prog Neurobiol,* 104**,** 38-66.
RUBIN, D. B. 2008. *Multiple Imputation for Nonresponse in Surveys,* New York, Wiley.
RUBINSZTEIN, D. C., LEGGO, J., CHIANO, M., DODGE, A., NORBURY, G., ROSSER, E. & CRAUFURD, D. 1997. Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. *Proc Natl Acad Sci U S A,* 94**,** 3872-6.
RUNNE, H., KUHN, A., WILD, E. J., PRATYAKSHA, W., KRISTIANSEN, M., ISAACS, J. D., REGULIER, E., DELORENZI, M., TABRIZI, S. J. & LUTHI-CARTER, R. 2007. Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood. *Proc Natl Acad Sci U S A,* 104**,** 14424-9.
SALEH, N., MOUTEREAU, S., DURR, A., KRYSTKOWIAK, P., AZULAY, J. P., TRANCHANT, C., BROUSSOLLE, E., MORIN, F., BACHOUD-LEVI, A. C. & MAISON, P. 2009. Neuroendocrine disturbances in Huntington's disease. *PLoS One,* 4**,** e4962.
SANCHEZ-PERNAUTE, R., GARCIA-SEGURA, J. M., DEL BARRIO ALBA, A., VIANO, J. & DE YEBENES, J. G. 1999. Clinical correlation of striatal 1H MRS changes in Huntington's disease. *Neurology,* 53**,** 806-12.
SCHAEFER, C. F., ANTHONY, K., KRUPA, S., BUCHOFF, J., DAY, M., HANNAY, T. & BUETOW, K. H. 2009. PID: the Pathway Interaction Database. *Nucleic Acids Research,* 37**,** D674-D679.
SCHMIDT, M. H. & PEARSON, C. E. 2016. Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst),* 38**,** 117-26.
SCHMUTTE, C., SADOFF, M. M., SHIM, K. S., ACHARYA, S. & FISHEL, R. 2001. The interaction of DNA mismatch repair proteins with human exonuclease I. *J Biol Chem,* 276**,** 33011-8.
SCHNEIDER, S. A., VAN DE WARRENBURG, B. P., HUGHES, T. D., DAVIS, M., SWEENEY, M., WOOD, N., QUINN, N. P. & BHATIA, K. P. 2006. Phenotypic homogeneity of the Huntington disease-like presentation in a SCA17 family. *Neurology,* 67**,** 1701-3.
SCHNEIDER, S. A., WALKER, R. H. & BHATIA, K. P. 2007. The Huntington's disease-like syndromes: what to consider in patients with a negative Huntington's disease gene test. *Nat Clin Pract Neurol,* 3**,** 517-25.
SCHWARZ, J. M., RODELSPERGER, C., SCHUELKE, M. & SEELOW, D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods,* 7**,** 575-6.
SEMAKA, A., KAY, C., DOTY, C., COLLINS, J. A., BIJLSMA, E. K., RICHARDS, F., GOLDBERG, Y. P. & HAYDEN, M. R. 2013. CAG size-specific risk estimates for intermediate allele repeat instability in Huntington disease. *J Med Genet,* 50**,** 696-703.
SEONG, I. S. 2005. HD CAG repeat implicates a dominant property of huntingtin in mitochondrial energy metabolism. *Human Molecular Genetics,* 14**,** 2871-2880.
SEREDENINA, T. & LUTHI-CARTER, R. 2012. What have we learned from gene expression profiles in Huntington's disease? *Neurobiol Dis,* 45**,** 83-98.
SHAM, P. C. & PURCELL, S. M. 2014. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet,* 15**,** 335-46.
SHI, H., KICHAEV, G. & PASANIUC, B. 2016. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet,* 99**,** 139-53.

SHIN, J. & LEE, C. 2015. A mixed model reduces spurious genetic associations produced by population stratification in genome-wide association studies. *Genomics,* 105**,** 191-196.

SHOULSON, I. & FAHN, S. 1979. Huntington's disease: Clinical care and evaluation. *Neurology,* 31**,** 1333-1335.

SINNREICH, M., SORENSON, E. J. & KLEIN, C. J. 2004. Neurologic course, endocrine dysfunction and triplet repeat size in spinal bulbar muscular atrophy. *Can J Neurol Sci,* 31**,** 378-82.

SMITH, A. 1968. The symbol-digit modalities test: a neuropsychologic test of learning and other cerebral disorders. *In:* HELMUTH, J. (ed.) *Learning disorders.* Seattle:: Seattle: Special Child Publications.

SMITH, B. N., NEWHOUSE, S., SHATUNOV, A., VANCE, C., TOPP, S., JOHNSON, L., MILLER, J., LEE, Y., TROAKES, C., SCOTT, K. M., JONES, A., GRAY, I., WRIGHT, J., HORTOBAGYI, T., AL-SARRAJ, S., ROGELJ, B., POWELL, J., LUPTON, M., LOVESTONE, S., SAPP, P. C., WEBER, M., NESTOR, P. J., SCHELHAAS, H. J., ASBROEK, A. A., SILANI, V., GELLERA, C., TARONI, F., TICOZZI, N., VAN DEN BERG, L., VELDINK, J., VAN DAMME, P., ROBBERECHT, W., SHAW, P. J., KIRBY, J., PALL, H., MORRISON, K. E., MORRIS, A., DE BELLEROCHE, J., VIANNEY DE JONG, J. M., BAAS, F., ANDERSEN, P. M., LANDERS, J., BROWN, R. H., JR., WEALE, M. E., AL-CHALABI, A. & SHAW, C. E. 2012. The C9ORF72 expansion mutation is a common cause of ALS+/-FTD in Europe and has a single founder. *Eur J Hum Genet*.

SMOGORZEWSKA, A., DESETTY, R., SAITO, T. T., SCHLABACH, M., LACH, F. P., SOWA, M. E., CLARK, A. B., KUNKEL, T. A., HARPER, J. W., COLAIACOVO, M. P. & ELLEDGE, S. J. 2010. A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Mol Cell,* 39**,** 36-47.

SNELL, R. G., MACMILLAN, J. C., CHEADLE, J. P., FENTON, I., LAZAROU, L. P., DAVIES, P., MACDONALD, M. E., GUSELLA, J. F., HARPER, P. S. & SHAW, D. J. 1993. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat Genet,* 4**,** 393-7.

SODING, J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics,* 21**,** 951-60.

SOKOLOVSKY, N., COOK, A., HUNT, H., GIUNTI, P. & CIPOLOTTI, L. 2010. A preliminary characterisation of cognition and social cognition in spinocerebellar ataxia types 2, 1, and 7. *Behav Neurol,* 23**,** 17-29.

SPADA, A. L. 2014. Spinal and Bulbar Muscular Atrophy.

ST-AMOUR, I., TURGEON, A., GOUPIL, C., PLANEL, E. & HEBERT, S. S. 2017. Co-occurrence of mixed proteinopathies in late-stage Huntington's disease. *Acta Neuropathol*.

STOREY, E., FORREST, S. M., SHAW, J. H., MITCHELL, P. & GARDNER, R. J. 1999. Spinocerebellar ataxia type 2: clinical features of a pedigree displaying prominent frontal-executive dysfunction. *Arch Neurol,* 56**,** 43-50.

STOREY, J. D. & TIBSHIRANI, R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A,* 100**,** 9440-5.

STOUT, J. C., JONES, R., LABUSCHAGNE, I., O'REGAN, A. M., SAY, M. J., DUMAS, E. M., QUELLER, S., JUSTO, D., SANTOS, R. D., COLEMAN, A., HART, E. P., DURR, A., LEAVITT, B. R., ROOS, R. A., LANGBEHN, D. R., TABRIZI, S. J. & FROST, C. 2012.

Evaluation of longitudinal 12 and 24 month cognitive outcomes in premanifest and early Huntington's disease. *J Neurol Neurosurg Psychiatry,* 83**,** 687-94.

STUART, J. M., SEGAL, E., KOLLER, D. & KIM, S. K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science,* 302**,** 249-55.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

SUBRAMONY, S. 2012. The Ataxias. *In:* WOOD, N. W. (ed.) *Neurogenetics. A guide for Clinicians.* Cambridge University Press.

SUGAWARA, N., PAQUES, F., COLAIACOVO, M. & HABER, J. E. 1997. Role of Saccharomyces cerevisiae Msh2 and Msh3 repair proteins in double-strand break-induced recombination. *PNAS,* 94**,** 9214 - 9219.

SUH, E., LEE, E. B., NEAL, D., WOOD, E. M., TOLEDO, J. B., RENNERT, L., IRWIN, D. J., MCMILLAN, C. T., KROCK, B., ELMAN, L. B., MCCLUSKEY, L. F., GROSSMAN, M., XIE, S. X., TROJANOWSKI, J. Q. & VAN DEERLIN, V. M. 2015. Semi-automated quantification of C9orf72 expansion size reveals inverse correlation between hexanucleotide repeat number and disease duration in frontotemporal degeneration. *Acta Neuropathol,* 130**,** 363-72.

SVEINBJORNSSON, G., ALBRECHTSEN, A., ZINK, F., GUDJONSSON, S. A., ODDSON, A., MASSON, G., HOLM, H., KONG, A., THORSTEINSDOTTIR, U., SULEM, P., GUDBJARTSSON, D. F. & STEFANSSON, K. 2016. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet,* 48**,** 314-7.

SWAMI, M., HENDRICKS, A. E., GILLIS, T., MASSOOD, T., MYSORE, J., MYERS, R. H. & WHEELER, V. C. 2009. Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum Mol Genet,* 18**,** 3039-47.

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A., TSAFOU, K. P., KUHN, M., BORK, P., JENSEN, L. J. & VON MERING, C. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res,* 43**,** D447-52.

TABRIZI, S. J., LANGBEHN, D. R., LEAVITT, B. R., ROOS, R. A., DURR, A., CRAUFURD, D., KENNARD, C., HICKS, S. L., FOX, N. C., SCAHILL, R. I., BOROWSKY, B., TOBIN, A. J., ROSAS, H. D., JOHNSON, H., REILMANN, R., LANDWEHRMEYER, B. & STOUT, J. C. 2009a. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol,* 8**,** 791-801.

TABRIZI, S. J., LANGBEHN, D. R., LEAVITT, B. R., ROOS, R. A., DURR, A., CRAUFURD, D., KENNARD, C., HICKS, S. L., FOX, N. C., SCAHILL, R. I., BOROWSKY, B., TOBIN, A. J., ROSAS, H. D., JOHNSON, H., REILMANN, R., LANDWEHRMEYER, B., STOUT, J. C. & INVESTIGATORS, T.-H. 2009b. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol,* 8**,** 791-801.

TABRIZI, S. J., REILMANN, R., ROOS, R. A., DURR, A., LEAVITT, B., OWEN, G., JONES, R., JOHNSON, H., CRAUFURD, D., HICKS, S. L., KENNARD, C., LANDWEHRMEYER, B., STOUT, J. C., BOROWSKY, B., SCAHILL, R. I., FROST, C. & LANGBEHN, D. R. 2012.

Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *Lancet Neurol,* 11**,** 42-53.

TABRIZI, S. J., SCAHILL, R. I., DURR, A., ROOS, R. A., LEAVITT, B. R., JONES, R., LANDWEHRMEYER, G. B., FOX, N. C., JOHNSON, H., HICKS, S. L., KENNARD, C., CRAUFURD, D., FROST, C., LANGBEHN, D. R., REILMANN, R. & STOUT, J. C. 2011. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol,* 10**,** 31-42.

TABRIZI, S. J., SCAHILL, R. I., OWEN, G., DURR, A., LEAVITT, B. R., ROOS, R. A., BOROWSKY, B., LANDWEHRMEYER, B., FROST, C., JOHNSON, H., CRAUFURD, D., REILMANN, R., STOUT, J. C. & LANGBEHN, D. R. 2013a. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *The Lancet Neurology,* 12**,** 637-49.

TABRIZI, S. J., SCAHILL, R. I., OWEN, G., DURR, A., LEAVITT, B. R., ROOS, R. A., BOROWSKY, B., LANDWEHRMEYER, B., FROST, C., JOHNSON, H., CRAUFURD, D., REILMANN, R., STOUT, J. C. & LANGBEHN, D. R. 2013b. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurol,* 12**,** 637-49.

TAHERZADEH-FARD, E., SAFT, C., AKKAD, D. A., WIECZOREK, S., HAGHIKIA, A., CHAN, A., EPPLEN, J. T. & ARNING, L. 2011. PGC-1alpha downstream transcription factors NRF-1 and TFAM are genetic modifiers of Huntington disease. *Mol Neurodegener,* 6**,** 32.

TAI, Y. F., PAVESE, N., GERHARD, A., TABRIZI, S. J., BARKER, R. A., BROOKS, D. J. & PICCINI, P. 2007. Microglial activation in presymptomatic Huntington's disease gene carriers. *Brain,* 130**,** 1759-1766.

TANAKA, F., REEVES, M. F., ITO, Y., MATSUMOTO, M., LI, M., MIWA, S., INUKAI, A., YAMAMOTO, M., DOYU, M., YOSHIDA, M., HASHIZUME, Y., TERAO, S., MITSUMA, T. & SOBUE, G. 1999. Tissue-specific somatic mosaicism in spinal and bulbar muscular atrophy is dependent on CAG-repeat length and androgen receptor--gene expression level. *Am J Hum Genet,* 65**,** 966-73.

TELENIUS H, K. H., THEILMANN J, ANDREW SE, ALMQVIST E, ANVRET M, GREENBERG C, GREENBERG J, LUCOTTE G, SQUITIERI F, STARR E, GOLDBERG YP, HAYDEN MR 1993. Molecular analysis of juvenile Huntington disease: the major influence on (CAG)n repeat length is the sex of the affected parent. *Human Molecular Genetics,* 2**,** 1535 - 1540.

TELENIUS, H., KREMER, B., GOLDBERG, Y. P., THEILMANN, J., ANDREW, S. E., ZEISLER, J., ADAM, S., GREENBERG, C., IVES, E. J., CLARKE, L. A. & ET AL. 1994. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nat Genet,* 6**,** 409-14.

TEZENAS DU MONTCEL, S., DURR, A., BAUER, P., FIGUEROA, K. P., ICHIKAWA, Y., BRUSSINO, A., FORLANI, S., RAKOWICZ, M., SCHOLS, L., MARIOTTI, C., VAN DE WARRENBURG, B. P., ORSI, L., GIUNTI, P., FILLA, A., SZYMANSKI, S., KLOCKGETHER, T., BERCIANO, J., PANDOLFO, M., BOESCH, S., MELEGH, B., TIMMANN, D., MANDICH, P., CAMUZAT, A., CLINICAL RESEARCH CONSORTIUM FOR SPINOCEREBELLAR ATAXIA (CRC-SCA), T., NETWORK, E., GOTO, J., ASHIZAWA, T., CAZENEUVE, C., TSUJI, S., PULST, S. M., BRUSCO, A., REISS, O.,

BRICE, A. & STEVANIN, G. 2014a. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain*.

TEZENAS DU MONTCEL, S., DURR, A., BAUER, P., FIGUEROA, K. P., ICHIKAWA, Y., BRUSSINO, A., FORLANI, S., RAKOWICZ, M., SCHOLS, L., MARIOTTI, C., VAN DE WARRENBURG, B. P., ORSI, L., GIUNTI, P., FILLA, A., SZYMANSKI, S., KLOCKGETHER, T., BERCIANO, J., PANDOLFO, M., BOESCH, S., MELEGH, B., TIMMANN, D., MANDICH, P., CAMUZAT, A., GOTO, J., ASHIZAWA, T., CAZENEUVE, C., TSUJI, S., PULST, S. M., BRUSCO, A., RIESS, O., BRICE, A. & STEVANIN, G. 2014b. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain,* 137**,** 2444-55.

THE NETWORK AND PATHWAY ANALYSIS SUBGROUP OF THE PSYCHIATRIC GENOMICS CONSORTIUM 2015. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nat Neurosci,* 18**,** 199-209.

THE UNIPROT, C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res,* 45**,** D158-D169.

THEDA, C., HWANG, S. H., CZAJKO, A., LOKE, Y. J., LEONG, P. & CRAIG, J. M. 2018. Quantitation of the cellular content of saliva and buccal swab samples. *Scientific Reports,* 8**,** 6944.

THOMPSON, R., JOHNSTON, L., TARUSCIO, D., MONACO, L., BEROUD, C., GUT, I. G., HANSSON, M. G., T HOEN, P. B., PATRINOS, G. P., DAWKINS, H., ENSINI, M., ZATLOUKAL, K., KOUBI, D., HESLOP, E., PASCHALL, J. E., POSADA, M., ROBINSON, P. N., BUSHBY, K. & LOCHMULLER, H. 2014. RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research. *J Gen Intern Med,* Suppl 3**,** S780-7.

THORNTON, T., CONOMOS, M. P., SVERDLOV, S., BLUE, E. M., CHEUNG, C. Y., GLAZNER, C. G., LEWIS, S. M. & WIJSMAN, E. M. 2014. Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing. *BMC Proceedings,* 8**,** 1-7.

TOME, S., MANLEY, K., SIMARD, J. P., CLARK, G. W., SLEAN, M. M., SWAMI, M., SHELBOURNE, P. F., TILLIER, E. R., MONCKTON, D. G., MESSER, A. & PEARSON, C. E. 2013a. MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. *PLoS Genet,* 9**,** e1003280.

TOME, S., SIMARD, J. P., SLEAN, M. M., HOLT, I., MORRIS, G. E., WOJCIECHOWICZ, K., TE RIELE, H. & PEARSON, C. E. 2013b. Tissue-specific mismatch repair protein expression: MSH3 is higher than MSH6 in multiple mouse tissues. *DNA Repair (Amst),* 12**,** 46-52.

TOMITA, H., VAWTER, M. P., WALSH, D. M., EVANS, S. J., CHOUDARY, P. V., LI, J., OVERMAN, K. M., ATZ, M. E., MYERS, R. M., JONES, E. G., WATSON, S. J., AKIL, H. & BUNNEY, W. E., JR. 2004. Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain. *Biol Psychiatry,* 55**,** 346-52.

TRÄGER, U., ANDRE, R., MAGNUSSON-LIND, A., MILLER, J. R. C., CONNOLLY, C., WEISS, A., GRUENINGER, S., SILAJDŽIĆ, E., SMITH, D. L., LEAVITT, B. R., BATES, G. P., BJÖRKQVIST, M. & TABRIZI, S. J. 2015. Characterisation of immune cell function in fragment and full-length Huntington's disease mouse models. *Neurobiology of Disease,* 73**,** 388-398.

TRANG, H., STANLEY, S. Y., THORNER, P., FAGHFOURY, H., SCHULZE, A., HAWKINS, C., PEARSON, C. E. & YOON, G. 2015. Massive CAG repeat expansion and somatic

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

TRAPNELL, C., PACHTER, L. & SALZBERG, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics,* 25**,** 1105-11.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc,* 7**,** 562-78.

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 association study. *The Lancet Neurology,* 15**,** 1248-1256.

TROTTIER, Y., DEVYS, D., IMBERT, G., SAUDOU, F., AN, I., LUTZ, Y., WEBER, C., AGID, Y., HIRSCH, E. C. & MANDEL, J.-L. 1995. Cellular localization of the Huntington's 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 molecular genetics. *Adv Neurol,* 79**,** 399-409.

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., CEDERQUIST, K. & SAVONTAUS, M. L. 1998. High prevalence of Kennedy's disease in Western Finland -- is the syndrome underdiagnosed? *Acta Neurol Scand,* 98**,** 128-33.

UKBEC. 2015. *UK Brain Expression Consortium (UKBEC) Braineac Database* [Online]. Available: http://www.braineac.org/ [Accessed 2014].

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., SCHEFFER, H., BRUNT, E. R., IPPEL, P. F., MAAT-KIEVIT, J. A., DOOIJES, D., NOTERMANS, N. C., LINDHOUT, D., KNOERS, N. V. & KREMER, H. P. 2002. Spinocerebellar ataxias in the Netherlands: prevalence and age at onset variance analysis. *Neurology,* 58**,** 702-8.

VAN DEN BROEK, W. J. A. A., NELEN, M. R., WANSINK, D. G., COERWINKE, M. M., RIELE, H., GROENEN, P. J. T. A., WIERINGA, G. & WIERINGA, B. 2002. Somatic expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. *Hum Mol Gen,* 11**,** 191-198.

VAN DER BURG, J. M., BJORKQVIST, M. & BRUNDIN, P. 2009. Beyond the brain: widespread pathology in Huntington's disease. *Lancet Neurol,* 8**,** 765-74.

VAN LEEUWEN, E. M., KANTERAKIS, A., DEELEN, P., KATTENBERG, M. V., THE GENOME OF THE NETHERLANDS, C., SLAGBOOM, P. E., DE BAKKER, P. I. W., WIJMENGA, C., SWERTZ, M. A., BOOMSMA, D. I., VAN DUIJN, C. M., KARSSEN, L. C. & HOTTENGA, J. J. 2015. Population-specific genotype imputations using minimac or IMPUTE2. *Nat. Protocols,* 10**,** 1285-1296.

VEITCH, N. J., ENNIS, M., MCABNEY, J. P., SHELBOURNE, P. F. & MONCKTON, D. G. 2007. Inherited CAG.CTG allele length is a major modifier of somatic mutation length variability in Huntington disease. *DNA Repair (Amst),* 6**,** 789-96.

VELAZQUEZ PEREZ, L., CRUZ, G. S., SANTOS FALCON, N., ENRIQUE ALMAGUER MEDEROS, L., ESCALONA BATALLAN, K., RODRIGUEZ LABRADA, R., PANEQUE HERRERA, M., LAFFITA MESA, J. M., RODRIGUEZ DIAZ, J. C., RODRIGUEZ, R. A., GONZALEZ ZALDIVAR, Y., COELLO ALMARALES, D., ALMAGUER GOTAY, D. & JORGE CEDENO, H. 2009. Molecular epidemiology of spinocerebellar ataxias in Cuba: insights into SCA2 founder effect in Holguin. *Neurosci Lett,* 454**,** 157-60.

VINTHER-JENSEN, T., NIELSEN, T. T., BUDTZ-JORGENSEN, E., LARSEN, I. U., HANSEN, M. M., HASHOLT, L., HJERMIND, L. E., NIELSEN, J. E. & NORREMOLLE, A. 2016. Psychiatric and cognitive symptoms in Huntington's disease are modified by polymorphisms in catecholamine regulating enzyme genes. *Clin Genet,* 89**,** 320-7.

VISSCHER, P. M. 2008. Sizing up human height variation. *Nat Genet,* 40**,** 489-90.

VISSCHER, P. M., MEDLAND, S. E., FERREIRA, M. A. R., MORLEY, K. I., ZHU, G., CORNES, B. K., MONTGOMERY, G. W. & MARTIN, N. G. 2006. Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLOS Genetics,* 2**,** e41.

WALL, J. D. & PRITCHARD, J. K. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet,* 4**,** 587-97.

WANG, K., LI, M. & BUCAN, M. 2007. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet,* 81**,** 1278-83.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WARBY, S. C., GRAHAM, R. K. & HAYDEN, M. R. 2014. Huntington Disease.

WARDLE, M., MORRIS, H. R. & ROBERTSON, N. P. 2009. Clinical and genetic characteristics of non-Asian dentatorubral-pallidoluysian atrophy: A systematic review. *Mov Disord,* 24**,** 1636-40.

WARREN, J. D., ROHRER, J. D. & ROSSOR, M. N. 2013. Clinical review. Frontotemporal dementia. *Bmj,* 347**,** f4827.

WEIR, B. S., ANDERSON, A. D. & HEPLER, A. B. 2006. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet,* 7**,** 771-80.

WELLCOME TRUST CASE CONTROL, C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

WEXLER, N. S., LORIMER, J., PORTER, J., GOMEZ, F., MOSKOWITZ, C., SHACKELL, E., MARDER, K., PENCHASZADEH, G., ROBERTS, S. A., GAYAN, J., BROCKLEBANK, D., CHERNY, S. S., CARDON, L. R., GRAY, J., DLOUHY, S. R., WIKTORSKI, S., HODES, M. E., CONNEALLY, P. M., PENNEY, J. B., GUSELLA, J., CHA, J. H., IRIZARRY, M., ROSAS, D., HERSCH, S., HOLLINGSWORTH, Z., MACDONALD, M., YOUNG, A. B., ANDRESEN, J. M., HOUSMAN, D. E., DE YOUNG, M. M., BONILLA, E., STILLINGS, T., NEGRETTE, A., SNODGRASS, S. R., MARTINEZ-JAURRIETA, M. D., RAMOS-ARROYO, M. A., BICKHAM, J., RAMOS, J. S., MARSHALL, F.,

SHOULSON, I., REY, G. J., FEIGIN, A., ARNHEIM, N., ACEVEDO-CRUZ, A., ACOSTA, L., ALVIR, J., FISCHBECK, K., THOMPSON, L. M., YOUNG, A., DURE, L., O'BRIEN, C. J., PAULSEN, J., BRICKMAN, A., KRCH, D., PEERY, S., HOGARTH, P., HIGGINS, D. S., JR. & LANDWEHRMEYER, B. 2004a. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc Natl Acad Sci U S A.* United States.

WEXLER, N. S., LORIMER, J., PORTER, J., GOMEZ, F., MOSKOWITZ, C., SHACKELL, E., MARDER, K., PENCHASZADEH, G., ROBERTS, S. A., GAYAN, J., BROCKLEBANK, D., CHERNY, S. S., CARDON, L. R., GRAY, J., DLOUHY, S. R., WIKTORSKI, S., HODES, M. E., CONNEALLY, P. M., PENNEY, J. B., GUSELLA, J., CHA, J. H., IRIZARRY, M., ROSAS, D., HERSCH, S., HOLLINGSWORTH, Z., MACDONALD, M., YOUNG, A. B., ANDRESEN, J. M., HOUSMAN, D. E., DE YOUNG, M. M., BONILLA, E., STILLINGS, T., NEGRETTE, A., SNODGRASS, S. R., MARTINEZ-JAURRIETA, M. D., RAMOS-ARROYO, M. A., BICKHAM, J., RAMOS, J. S., MARSHALL, F., SHOULSON, I., REY, G. J., FEIGIN, A., ARNHEIM, N., ACEVEDO-CRUZ, A., ACOSTA, L., ALVIR, J., FISCHBECK, K., THOMPSON, L. M., YOUNG, A., DURE, L., O'BRIEN, C. J., PAULSEN, J., BRICKMAN, A., KRCH, D., PEERY, S., HOGARTH, P., HIGGINS, D. S., JR. & LANDWEHRMEYER, B. 2004b. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc Natl Acad Sci U S A,* 101**,** 3498-503.

WEYDT, P., SOYAL, S. M., GELLERA, C., DIDONATO, S., WEIDINGER, C., OBERKOFLER, H., LANDWEHRMEYER, G. B. & PATSCH, W. 2009. The gene coding for PGC-1alpha modifies age at onset in Huntington's Disease. *Mol Neurodegener,* 4**,** 3.

WHEELER, V. C., KOVALENKO, M., GIORDANO, J., ANDREW, M., MENALLED, L. B., ALEXANDROV, V., THIEDE, C., WEIDNER, J., TEICHMANN, M., TOTTEY, W., CUMMING, S. A., CORREIA, K., BARKER, D., LAGER, B., FLYNN, G., FISCHER, D. F., TILLACK, K., MONCKTON, D. G., BRUNNER, D., RAMBOZ, S., KWAK, S. & HOWLAND, D. 2016. HTT CAG KNOCK-IN MICE WITH PURE AND INTERRUPTED REPEAT TRACTS PROVIDE INSIGHT INTO THE ROLE OF SOMATIC EXPANSION IN HD PATHOGENESIS. *JNNP,* 87(Suppl 1)**,** A1-A120.

WHEELER, V. C., LEBEL, L.-A., VRBANAC, V., TEED, A., RIELE, H. & MACDONALD, M. E. 2003. Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum. *Human Molecular Genetics,* 12**,** 273-281.

WHITNEY, A. R., DIEHN, M., POPPER, S. J., ALIZADEH, A. A., BOLDRICK, J. C., RELMAN, D. A. & BROWN, P. O. 2003. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci U S A,* 100**,** 1896-901.

WILD, E., MAGNUSSON, A., LAHIRI, N., KRUS, U., ORTH, M., TABRIZI, S. J. & BJÖRKQVIST, M. 2011. Abnormal peripheral chemokine profile in Huntington's disease. *PLoS Curr,* 3**,** RRN1231.

WILD, E. J., MUDANOHWO, E. E., SWEENEY, M. G., SCHNEIDER, S. A., BECK, J., BHATIA, K. P., ROSSOR, M. N., DAVIS, M. B. & TABRIZI, S. J. 2008. Huntington's disease phenocopies are clinically and genetically heterogeneous. *Mov Disord,* 23**,** 716-20.

WILD, E. J., TABRIZI S. T. 2007. Huntington's disease phenocopy syndromes. *Curr Opin Neurol,* 20**,** 681-687.

WILD, E. J. & TABRIZI, S. J. 2012. Huntington's disease. *In:* WOOD, N. W. (ed.) *Neurogenetics: a Guide for Clinicians.* Cambridge University Press.

WILLER, C. J., LI, Y. & ABECASIS, G. R. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics,* 26**,** 2190-1.

WILLIAMS, G. M. & SURTEES, J. A. 2015a. MSH3 Promotes Dynamic Behavior of Trinucleotide Repeat Tracts In Vivo. *Genetics,* 200**,** 737-+.
WILLIAMS, G. M. & SURTEES, J. A. 2015b. *MSH3* promotes dynamic behaviour of trinucleotide repeat tracts *in vivo*. *Genetics,* 11.
WOOD, N. 2012. *Neurogenetics: A Guide for Clinicians*, Cambridge University Press.
WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. & LIN, X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet,* 89**,** 82-93.
WYSS-CORAY, T. & ROGERS, J. 2012. Inflammation in Alzheimer disease-a brief review of the basic science and clinical literature. *Cold Spring Harb Perspect Med,* 2**,** a006346.
YANG, J., BAKSHI, A., ZHU, Z., HEMANI, G., VINKHUYZEN, A. A. E., LEE, S. H., ROBINSON, M. R., PERRY, J. R. B., NOLTE, I. M., VAN VLIET-OSTAPTCHOUK, J. V., SNIEDER, H., THE LIFELINES COHORT, S., ESKO, T., MILANI, L., MAGI, R., METSPALU, A., HAMSTEN, A., MAGNUSSON, P. K. E., PEDERSEN, N. L., INGELSSON, E., SORANZO, N., KELLER, M. C., WRAY, N. R., GODDARD, M. E. & VISSCHER, P. M. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet,* 47**,** 1114-1120.
YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. & VISSCHER, P. M. 2010a. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-9.
YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. & VISSCHER, P. M. 2010b. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet,* 42**,** 565-569.
YANG, J., FERREIRA, T., MORRIS, A. P., MEDLAND, S. E., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., WEEDON, M. N. & LOOS, R. J. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics,* 44**,** 369-375.
YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics,* 88**,** 76-82.
YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. & PRICE, A. L. 2014a. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet,* 46**,** 100-6.
YANG, J., ZAITLEN, N. A., GODDARD, M. E., VISSCHER, P. M. & PRICE, A. L. 2014b. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics,* 46**,** 100-106.
YEH, T. H., LAI, S. C., WENG, Y. H., KUO, H. C., WU-CHOU, Y. H., HUANG, C. L., CHEN, R. S., CHANG, H. C., TRAYNOR, B. & LU, C. S. 2012. Screening for C9orf72 repeat expansions in parkinsonian syndromes. *Neurobiol Aging*.
YUCE, O. & WEST, S. C. 2013. Senataxin, Defective in the Neurodegenerative Disorder Ataxia with Oculomotor Apraxia 2, Lies at the Interface of Transcription and the
DNA Damage Response. *Molecular and Cellular Biology,* 33**,** 406-417.

ZENG, W., GILLIS, T., HAKKY, M., DJOUSSE, L., MYERS, R. H., MACDONALD, M. E. & GUSELLA, J. F. 2006. Genetic analysis of the GRIK2 modifier effect in Huntington's disease. *BMC Neurosci,* 7**,** 62.

ZHANG, B., GAITERI, C., BODEA, L. G., WANG, Z., MCELWEE, J., PODTELEZHNIKOV, A. A., ZHANG, C., XIE, T., TRAN, L., DOBRIN, R., FLUDER, E., CLURMAN, B., MELQUIST, S., NARAYANAN, M., SUVER, C., SHAH, H., MAHAJAN, M., GILLIS, T., MYSORE, J., MACDONALD, M. E., LAMB, J. R., BENNETT, D. A., MOLONY, C., STONE, D. J., GUDNASON, V., MYERS, A. J., SCHADT, E. E., NEUMANN, H., ZHU, J. & EMILSSON, V. 2013. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell,* 153**,** 707-20.

ZHANG, B. & HORVATH, S. 2005. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol,* 4**,** Article17.

ZHANG, D., IYER, L. M., HE, F. & ARAVIND, L. 2012. Discovery of Novel DENN Proteins: Implications for the Evolution of Eukaryotic Intracellular Membrane Structures and Human Disease. *Front Genet,* 3**,** 283.

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 singal on gel. (Resuspend in less H2O if PCRs are weaker.) Vortex plate. Leave for 5 mins.
6) Votex 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 ddH20.
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 referes 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: [PDGnumber]_[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.

1. Flower M, Lomeikaite V, Ciosi M, Cumming S, Morales F, Lo K K, **Hensman Moss D J**, Jones L, Holmans P , The TRACK-HD Investigators, The OPTIMISTIC Consortium, Monckton D G and Tabrizi S J. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. **Brain**. 2020. 142(7); 1876-1886.
2. Goold R, Flower M, **Hensman Moss D**, Medway C, Wood-Kaczmar A, Andre R, Farshim P, Bates GP, Holmans P, Jones L, Tabrizi SJ. FAN1 modifies Huntington's disease progression by stabilizing the expanded HTT CAG repeat. **Hum Mol Genet**. 2019 Feb 15;28(4):650-661.
3. Lahr J, Minkova L, Tabrizi SJ, Stout JC, Klöppel S, Scheller E; TrackOn-HD Investigators: Coleman A, Decolongon J, Fan M, Koren T, Jauffret C, Justo D, Lehericy S, Nigaud K, Valabrègue R, Schoonderbeek A, 't Hart EP, Crawford H, Gregory S, **Hensman Moss D**, Johnson E, Read J, Owen G, Papoutsi M, Berna C, Razi A, Rees G, Scahill RI, Craufurd D, Reilmann R, Weber N, Stout J, Labuschagne I, Orth M, Landwehrmeyer GB, Langbehn D, Johnson H, Long J, Mills J.. Working Memory-Related Effective Connectivity in Huntington's Disease Patients. **Front Neurol**. 2018 Jun 4;9:370.
4. **Hensman Moss DJ***, Pardiñas AF*, Langbehn D, Lo K, Leavitt BR, Roos R, Durr A, Mead S, Holmans P, Jones L§, Tabrizi ST§, and the REGISTRY and the TRACK-HD investigators. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. ***The Lancet Neurology***. 2017. 16(9) 701-711. **Co-first author.*
5. **Hensman Moss DJ**, Robertson N, Farmer R, Scahill RI, Haider S, Tessari MA, Flynn G, Fischer DF, Wild EJ, Macdonald D, Tabrizi SJ. Quantification of huntingtin protein species in Huntington's disease patient leukocytes using optimised electrochemiluminescence immunoassays. ***PLOS ONE***. Dec 2017.
6. **Hensman Moss DJ**, Flower, MD, Lo KK, Miller JR, van Ommen G-J B, Hoen PAC, Stone TC, Guinee A, Langbehn DR, Jones L, Plagnol V, van Roon-Mom WMC, Holmans P, Tabrizi SJ. Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. ***Nature Scientific Reports***. 2017. 7, 44849.
7. **Hensman Moss DJ,** Tabrizi ST. A Newly Recognized HD-Phenocopy Associated with *C9orf72* Expansion: Case Studies in Movement Disorders Edited by Kailash P. Bhatia, Roberto Erro and Maria Stamelou. Cambridge University Press. April 2017. (Book chapter)
8. McColgan P, Gregory S, Razi A, Seunarine KK, Gargouri F, Durr A, Roos RA, Leavitt BR, Scahill RI, Clark CA, Tabrizi SJ, Rees G; Track On-HD Investigators, Coleman A, Decolongon J, Fan M, Petkau T, Jauffret C, Justo D, Lehericy S, Nigaud K, Valabrègue R, Choonderbeek A, Hart EP, **Hensman Moss DJ**, Crawford H, Johnson E, Papoutsi M, Berna C, Reilmann R, Weber N, Stout J, Labuschagne I, Landwehrmeyer B, Orth M, Johnson H. White matter predicts functional connectivity in premanifest Huntington's disease. Ann Clin Transl Neurol. 2017 Jan 16;4(2):106-118. doi: 10.1002/acn3.384.
9. Bettencourt C*, **Hensman Moss** D*, Flower M*, Wiethoff S*, Brice A, Goizet C, Stevanin G, Koutsis G, Karadima G, Panas M, Yescas-Gómez P, García-Velázquez LE, Alonso-Vilatela ME, Lima M, Raposo M, Traynor B, Sweeney M, Wood N, Giunti P; SPATAX Network, Durr A, Holmans P, Houlden H, Tabrizi SJ, Jones L. DNA repair pathways

underlie a common genetic mechanism modulating onset in polyglutamine diseases. ***Ann Neurol***. 2016 Jun;79(6):983-90. doi: 10.1002/ana.24656. **Co-first author.*

10. Hensman Moss DJ, Wood NW, Tabrizi SJ. Other genetic causes of cognitive impairment: Oxford Textbook of Cognitive Neurology and Dementia Edited by Masud Husain and Jonathan M. Schott. Oxford University Press. June 2016. (Book chapter: this book won first prize in the Neurology section at the 2017 British Medical Association book awards)

11. Miller JR, Lo KK, Andre R, **Hensman Moss** DJ, Träger U, Stone TC, Jones L, Holmans P, Plagnol V, Tabrizi SJ. RNA-Seq of Huntington's disease patient myeloid cells reveals innate transcriptional dysregulation associated with proinflammatory pathway activation. ***Hum Mol Genet***. 2016 May 11. pii: ddw142.

12. Minkova L, Scheller E, Peter J, Abdulkadir A, Kaller CP, Roos RA, Durr A, Leavitt BR, Tabrizi SJ, Klöppel S; TrackOn-HD Investigators**,** Coleman A, Decolongon J, Fan M, Petkau T, Jauffret C, Justo D, Lehericy S, Nigaud K, Valabrègue R, Choonderbeek A, Hart EP, **Hensman Moss DJ**, Crawford H, Johnson E, Papoutsi M, Berna C, Reilmann R, Weber N, Stout J, Labuschagne I, Landwehrmeyer B, Orth M, Johnson H. . Detection of Motor Changes in Huntington's Disease Using Dynamic Causal Modeling. ***Front Hum Neurosci***. 2015 Nov 25;9:634.

13. Klöppel S, Gregory S, Scheller E, Minkova L, Razi A, Durr A, Roos RAC, Leavitt BR et al and the Track-On investigators**,** Coleman A, Decolongon J, Fan M, Petkau T, Jauffret C, Justo D, Lehericy S, Nigaud K, Valabrègue R, Choonderbeek A, Hart EP, **Hensman Moss DJ**, Crawford H, Johnson E, Papoutsi M, Berna C, Reilmann R, Weber N, Stout J, Labuschagne I, Landwehrmeyer B, Orth M, Johnson H. . Compensation in Preclinical Huntington's Disease: Evidence from the Track-On HD Study. ***EBioMed***. Epub Aug 2015.

14. **Hensman Moss DJ**, Poulter M, Beck J, Hehir J, Polke JM, Campbell T, Adamson G, Mudanohwo E, McColgan P, Haworth A, Wild EJ, Sweeney MG, Houlden H, Mead S, Tabrizi SJ. C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. ***Neurology.*** 2014 Jan 28;82(4):292-9.

15. Fratta P, Poulter M, Lashley T, Rohrer JD, Polke JM, Beck J, Ryan N, **Hensman D**, Mizielinska S, Waite AJ, Lai MC, Gendron TF, Petrucelli L, Fisher EM, Revesz T, Warren JD, Collinge J, Isaacs AM, Mead S. Homozygosity for the C9orf72 GGGGCC repeat expansion in frontotemporal dementia. ***Acta Neuropathol***. 2013 Sep;126(3):401-9.

16. Beck J*, Poulter M*, **Hensman D**, Rohrer JD, Mahoney CJ, Adamson G, Campbell T, Uphill J, Borg A, Fratta P, Orrell RW, Malaspina A, Rowe J, Brown J, Hodges J, Sidle K, Polke JM, Houlden H, Schott JM, Fox NC, Rossor MN, Tabrizi SJ, Isaacs AM, Hardy J, Warren JD, Collinge J, Mead S. Large C9orf72 Hexanucleotide Repeat Expansions Are Seen in Multiple Neurodegenerative Syndromes and Are More Frequent Than Expected in the UK Population. ***Am J Hum Genet***. 2013 Mar 7;92(3):345-53. **Co-first author.*