Investigating the Role of Modifiers in Trinucleotide Repeat Diseases

Student Name: Heather Ging

Institution: University College London (UCL)

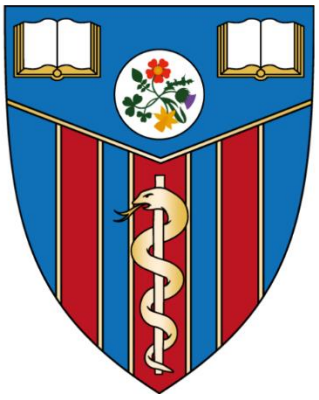
Thesis Degree: Doctor of Philosophy, PhD

Student Status: UCL Institute of Neurology Research Student

Student Number: 14088709

Primary Supervisor: Prof Paola Giunti

Secondary Supervisor: Prof Jernej Ule



Declaration

I, Heather Ging, 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.

Signature:	Date:
Heather Ging	30.11.2020

Acknowledgements

This thesis is dedicated to two inspiring women from my past and present. First to the late Breda Ging, known for her red lipstick, ambition and love of her family, who sparked my interest in the field, and to Lauren Byrne, who provided me with the professional and personal support to complete this thesis. Besides your vast and in-depth knowledge, your passion alone for Huntington's disease is truly humbling and inspirational.

My thanks to my supervisors, Prof Paola Giunti and Prof Jernej Ule, for giving me the opportunity to drive my project under their guidance. I would like to especially thank my "unofficial supervisor" Dr Suran Nethisinghe who has supported me from near and far, over the course of my MSc project turned PhD project. Your knowledge of the field is a credit to you, along with your expert technical skills and attention to detail (or OCD, who knows!). I truly believe that I am the researcher I am today because of you. I am extremely grateful to Prof Gill Bates whose incredible support led to the completion of this thesis. A big thank you to the past and current members of the Giunti Team and wider department, Dr Léa R'Bibo, Mr Chris Lovejoy, Dr Alexander Brown, Dr Gilbert Thomas-Black, Dr Hector Garcia-Moreno, Dr Rosella Abeti, Dr Anna Zeitlberger, Ms Natalie Guzikowski and Dr Zofia Fleszar, all of whom were a pleasure to work with.

To the one I love, Jack, thank you for being my escape and for making me laugh on a daily basis. To my powerhouses and confidants, Nikki, Caitrìona, Aisling and Aysha, thank you for being there through every experience. Thank you to my twin Laurence, for flying me home in times of need. In saving the best for last, my biggest and most sincere thanks to my Mam and Dad, who not only suffered me as a student for nearly 8 years, but their belief in me never faltered. Your continued support, motivation and the knowledge that I wouldn't be any less of a person if I didn't want to do/continue the PhD was my ultimate driving force. If not for you two, then for who?

Abstract

Friedreich's Ataxia (FRDA) and Huntington's disease (HD) are trinucleotide repeat diseases, resulting from homozygous expanded GAA and heterozygous expanded CAG repeats, respectively. The pathogenic repeat length is inversely correlated to disease onset and severity; the longer the repeat, the earlier the onset and the more severe the phenotype. However, repeat size does not fully explain the phenotypic variability. This report investigates if repeat interruptions act as disease modifiers. Small GAA repeat interruptions were common in the FRDA cohort in contrast to large interruptions, yet we did not expose the base pair configuration. Clone sequencing of the CAG repeat revealed that the age at onset is more accurately predicted based on the length of the pure CAG repeat. The penultimate synonymous CAA interruption was determined to modify onset with its loss hastening onset and an additional CAA interruption delaying onset. These results have been substantiated by recent reports (GEM-HD, 2019; Wright et al., 2019). Complementing clone sequencing, the base pair configuration of the HD pathogenic region was determined using next- and third-generation sequencing technologies, revealing that Illumina MiSeq sequencing was most applicable to our samples based on DNA quantity and quality. In contrast, Pacific Biosciences single molecule real time sequencing and Oxford Nanopore sequencing were limited by sample concentration.

A prominent characteristic of HD is somatic mosaicism, which mirrors the specific neurodegeneration. To understand the contribution of CAG repeat instability to HD pathogenesis, the somatic mosaicism profile in six HD *post-mortem* brains was analysed by Illumina MiSeq, which determines the proportion of common variants (small CAG repeat changes) and SP-PCR, which quantifies large CAG repeat changes. Illumina MiSeq revealed that the striatum contained the highest level of instability. In contrast, SP-PCR determined that the cortical regions displayed the greatest levels of instability. These results complement the somatic mosaicism profiles previously determined in Kennedy *et al*., 2003, and supports the hypothesis that cells with the largest CAG repeat sizes are primarily lost. Emerging evidence highlights DNA repair pathway genes as modifiers of CAG repeat instability and HD phenotype (GEM-HD, 2019). In our HD cohort, who were genotyped for the implicated disease modifying DNA repair pathway SNPs, we similarly show that some of the phenotypic variability can be attributed to these genetic variants, specifically in *FAN1*. The future use of more physiological disease models, such as induced pluripotent stem cells, will aid in deciphering the exact role of DNA repair pathway genes as modifiers of instability and thus, disease progression.

Impact Statement

Friedreich's ataxia (FRDA) and Huntington's disease (HD) are trinucleotide repeat expansion diseases. Depending on the mutation size, these diseases often manifest with markedly varied phenotypes, the reason for which is currently unknown. This report focuses on examining the genotype-phenotype relationship within these disorders and additionally aims to identify potential modifiers of disease progression. The delineation of such modifiers will not only advance our understanding of disease progression, but will also aid in the development of validated therapeutic targets that can be translated across trinucleotide repeat disorders as a whole.

As the trinucleotide repeat length is the dominant determinant of disease onset, this report investigates the presence of pathogenic sequence interruptions as disease modifiers. The results do not identify large repeat sequence interruptions as disease modifiers in FRDA, however, this is based on methods that do not determine the sequence configuration. The age at disease onset in HD was better predicted when the length of the pure CAG repeat was considered instead of the polyglutamine length, which includes the penultimate synonymous CAA interruption. More specifically, the absence of the CAA interruption in HD patients significantly advances disease onset, whereas the presence of an additional interruption delays onset. This report therefore highlights the modifying role of CAA interruptions, which has been further substantiated by recent reports (GEM-HD, 2019; Wright et al., 2019). It is clear from the results presented here that the repeat size alone is not sufficient to accurately predict disease onset and further emphasises the need to translate sequencing technologies into routine diagnostic services, which will ultimately be of benefit for the genetic counselling of patients. However, until these technologies are optimised across all sample types and until they are made cost efficient, this will not be an immediate transition.

The instability of the pathogenic repeat in HD has been implicated in disease progression and accordingly, we have investigated DNA repair pathway genetic variants which have been previously identified as disease modifiers (GEM-HD, 2019). The identification of such modifiers will ultimately guide the development of therapeutic targets with the aim to reduce instability to a sufficient level that slows disease progression. Expanding on previous research, this report identifies DNA repair pathway genetic variants as modifiers of HD age at onset. This provides a foundation for further investigation and development into these variants as therapeutic targets that have the potential to be translated across the trinucleotide repeat diseases that display repeat instability.

Table of Contents

List of Figures

List of Tables

Abbreviations

1KGP	1000 Genomes Project
^{32}P	phosphorus 32
8-oxoG	7,8-dihydro-8-oxoguanine
AAO	age at onset
ADORA2A	adenosinergic A2A receptor
AAD	age at death
AP site	apurinic or apyrmidinic site
Atg7	autophagy-related protein 7
ATN1	atrophin 1
ATR	Rad3-related protein kinase
ATXN	ataxin
ATXN1	ataxin-1
ATXN3	ataxin-3
BDNF	brain-derived neurotrophic factor
BER	base excision repair
BMP	bone morphogenetic protein
bp	base pair
BPES	blepharophimosis ptosis and epicanthus inversus
C9orf72	chromosome 9 open reading frame 72
CAG	glutamine
$(CAG)_n$	CAG repeat length (of size "n")
CAT	histidine
CBP	cAMP-response elelment (CREB)-binding protein
CCD	cleidocranial dysplasia
CCHS	congenital central hypoventilation syndrome
CCS	circular consensus sequence
c-MYC	myc proto-oncogene c
crRNA	Cas9 guide RNA
Ctip2	COUP-TF-interacting protein 2
DARPP32	dopamine- and cAMP-regulated neuronal phosphoprotein
DBS	double-strand break
Dlx2	distal-less homeobox 2
DM	myotonic dystrophy
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dRP	5'-2-deoxyribose-5-phosphate
DRG	dorsal root ganglion
Drop-seq	droplet-based single-cell RNA-seq
DRPLA	dentatorubral pallidoluysian atrophy
EA	expanded allele
EGF	epidermal growth factor
En2	homeobox protein engrailed-2
eQTL	expression quantitative trait loci
ESC	embryonic stem cell
ExAC	Exome Aggregation Consortium
EXO1	exonuclease 1
FACS	fluorescence-activated cell sorting
FAN1	fanconi-associated nuclease 1
FECD	Fuchs endothelial corneal dystrophy;
Fen1	flap endonuclease 1

FGF	fibroblast growth factor
FLIM	fluorescence lifetime imaging microscopy
FMR1	fragile X mental retardation 1
FoxP	forkhead box
FRDA	Friedreich's ataxia
FRET	Förster resonance energy transfer
FRAXA	fragile X syndrome
FRAXE	fragile X mental retardation associated with *FRAXE* site
FXN	Frataxin gene
FXN	frataxin protein
FXTAS	fragile X tremor and ataxia syndrome
Gbx	gastrulation brain homeobox
GeM-HD	The Genetic Modifiers of Huntington's Disease Consortium
GG-NER	global genomic NER
GRIK2	GluR6 subunit of kainite receptor
Gsx2	Genetic-screened homeobox 2
GWA	genome-wide association
GWAS	genome-wide association study
HAP1	huntingtin-associated protein-1
HD	Huntington's disease
HDL2	Huntington's-disease-like 2
Hdh	Huntington's disease gene homolog
HFG	hand–foot–genital syndrome
HNK	human natural killer
HPE5	holoprosencephaly 5
HTT	Huntingtin gene
HTT	huntingtin protein
ICL	inter-strand crosslink
iPSC	induced pluripotent stem cell
IsO	isthmic organiser
ISSX	X-linked infantile spasm syndrome
Kirrel2	kirre like nephrin family adhesion molecule 2
KLF4	Kruppel-like factor 4
L7	Purkinje cell protein 2
LGE	lateral ganglionic eminence
Lhx5	LIM homeobox 5
LIG1	DNA ligase I
LIG3	DNA ligase III
LP-BER	long-patch BER
LRRFIP2	LRR binding FLII interacting protein 2
mHTT	mutant huntingtin protein
MEF	mouse embryonic fibroblast
MHB	midbrain-hindbrain boundary
min	minute/s
MJD	Machado-Joseph disease
MLH1	MutL E. coli homolog of 1
MLH3	MutL E. coli homolog of 3
MMLV	murine molony leukemia virus
MMR	mismatch repair
MRGH	mental retardation with isolated growth hormone deficiency
MSH2	MutS E. coli homolog of 2
MSH3	MutS E. coli homolog of 3

MSH6	MutS E. coli homolog of 6
MTMR10	myotubularin related protein 10
MTMR10	myotubulin-related protein 10
NBM	neurobasal medium
NCBI	National Center for Biotechnology Information
NER	nucleotide excision repair
Neph3	nephrin-like 3
NeuN	neuronal nuclear antigen
NGS	next generation sequencing
NHEJ	non-homologous end-joining
NK-kB	nuclear factor-kB
nNOS	nitric oxide synthase
NMD	nonsense-mediated mRNA decay
Nolz1	zinc finger protein 503
OCT-4	octamer-binding transcription factor 4
OGG1	8-oxoguanine DNA glycosylase
ONT	Oxford Nanopore Technologies
OPMD	oculopharyngeal muscular dystrophy
Otx2	orthodenticle homeobox 2
OXPHOS	oxidative phosphorylation
p53	tumour suppressor protein 53
PacBio	Pacific Biosciences
PAM	protospacer-adjacent motif
PCNA	proliferating cell nuclear antigen
PCR	polymerase chain reaction
PE	paired-end
PGC-1α	peroxisome proliferator-activated receptor-γ coactivator 1α
PMS1	postmeiotic segregation increased s.cerevisiae 1
PMS2	postmeiotic segregation increased s.cerevisiae 2
POLB	polymerase β
Polε	DNA polymerase ε
polyA	poly alanine
PolyP	polyproline
PolyQ	polyglutamine
Polδ	DNA polymerase δ
RNA	ribnucleic acid
RPA	replication protein A
RPA1	replication protein A1
RRM2B	ribonucleoside-diphosphate reductase subunit M2 B
RT	reverse transcription
RT-qPCR	real time quantitative polymerase chain reaction
SBS	sequencing-by-synthesis
SBMA	spinal and bulbar muscular atrophy
SCA	spinocerebellar ataxia
SDF1	stromal cell-derived factor 1
sec	second/s
SeV	Sendai virus
sgRNA	single guide RNA
SMAX/SBMA	spinal and bulbar muscular atrophy X-linked 1
SMRT	single molecule real time
SN-BER	single-nucleotide BER
SnmC-seq	single nucleus methylcytosine sequencing

SNP	single nucleotide polymorphism
sNuc-seq	single-nucleus RNA sequencing
SOX2	sex determining region Y box 2
SPD	synpolydactyly
SP1	specificity protein-1
SP-PCR	small pool PCR
TC-NER	transcription coupled NER
TGS	third-generation sequencing
TP-PCR	triplet primed PCR
UBR5	ubiquitin protein ligase E3 component N-recognin 5
UCHL1	ubiquitin C-terminal hydrolase 1
UTR	untranslated region
WT	wild-type allele
XPA	xeroderma pigmentosum complementation group A
ZMW	zero-mode waveguide

Chapter 1. Introduction

1.1 Trinucleotide Repeat Diseases

Trinucleotide repeat diseases are caused by triplet repeat expansions, which are a repeated DNA sequence of three nucleotides that exceeds the stable threshold in their associated disease-specific genes (Figure1.1). Currently, atypical trinucleotide repeats account for over 40 neurological diseases, which are characterised by the location of the repeat within the gene (Kovtun and McMurray, 2008). The two main categories consist of expanded trinucleotide repeats that are located outside of the gene coding region and those that are translated into a protein product. Expanded trinucleotide repeats located outside of the gene coding region cause loss of gene function by reducing or abolishing transcription such as the intronic GAA repeat expansions in Friedreich's ataxia (FRDA), which results in *frataxin* gene silencing. In contrast, the CAG repeat in Huntington's disease (HD) is translated into a mutant polyglutamine tract in the huntingtin protein, which leads to the dysfunction and degeneration of specific neuronal subpopulations (Reiner et al., 2011). Pathogenically expanded CAG repeats encoding polyglutamine tracts are responsible for a family of nine known diseases including HD, spinocerebellar ataxia (SCA) types 1, 2, 3 (Machado-Joseph disease), 6, 7, and 17, spinal and bulbar muscular atrophy (SBMA) and dentatorubral pallidoluysian atrophy (DRPLA) (La Spada and Taylor, 2003).

Generally, the non-coding disease-associated trinucleotide repeat sequences are longer than the coding repeat sequences and once past the critical disease threshold, there is a strong inverse relationship between the length of the repeat and disease severity in that the longer the repeat, the earlier the age at onset (Jones et al., 2017). Expanded trinucleotide repeats have the propensity to adopt unusual structural features, which may variably disrupt the cellular replication, repair and recombination machineries in different conditions, and are predisposed to further expansion (Lee and McMurray, 2014; Mirkin, 2007). Repeat instability, when the repeat becomes unstable and increases or decreases in size, can occur in dividing and non-dividing cells and is tissue-, cell-, and disease-specific (Gomes-Pereira et al., 2014). One consequence of increased trinucleotide repeat expansion in the germline is genetic anticipation, which describes the earlier disease onset in successive generations due to inheriting a larger trinucleotide repeat expansion (Mirkin, 2007).

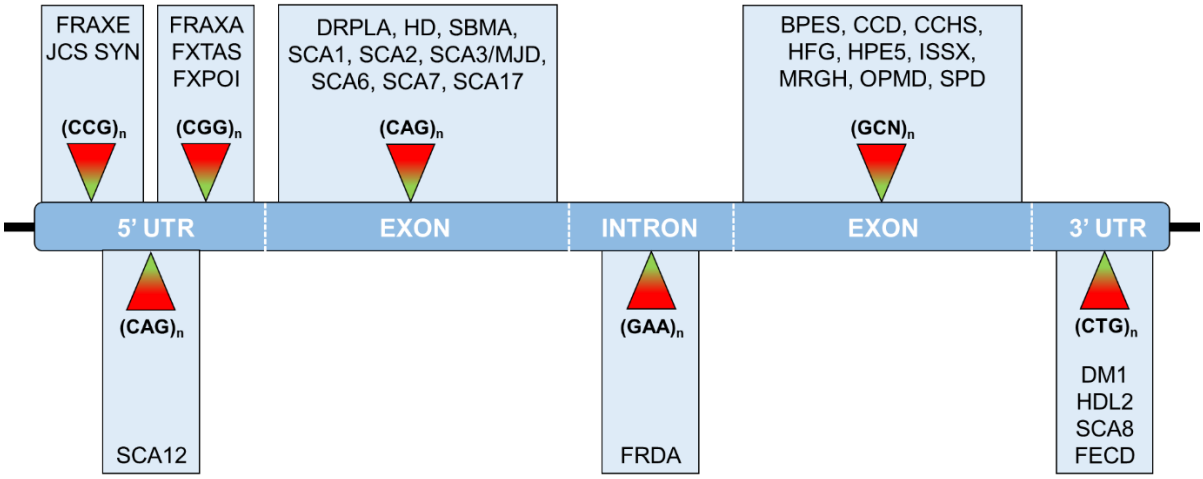


Figure 1.1. Trinucleotide repeats and their associated diseases

Trinucleotide repeat diseases are displayed by the corresponding location of their causative triplet repeat expansion within the gene. BPES: blepharophimosis ptosis and epicanthus inversus; CCD: cleidocranial dysplasia; CCHS: congenital central hypoventilation syndrome; DM: myotonic dystrophy; DRPLA: dentatorubral–pallidoluysian atrophy; FECD: Fuchs endothelial corneal dystrophy; FRAXA: fragile X syndrome; FRAXE: fragile X mental retardation associated with *FRAXE* site; FRDA: Friedreich's ataxia; FXPOI: fragile X-associated primary ovarian insufficiency; FXTAS: fragile X tremor and ataxia syndrome; HD: Huntington's disease; HDL2: Huntington's-disease-like 2; HFG: hand–foot–genital syndrome; HPE5: holoprosencephaly 5; ISSX: X-linked infantile spasm syndrome; JCS SYN: Jacobsen syndrome; MRGH: mental retardation with isolated growth hormone deficiency; OPMD: oculopharyngeal muscular dystrophy; SBMA: spinal and bulbar muscular atrophy; SCA: spinocerebellar ataxia; SPD: synpolydactyly; n: repeat length number; UTR: untranslated region; triangles: illustrative of wild-type and expanded repeats.

The clinical differences between the trinucleotide repeat diseases are thought to be due to the function and expression pattern of the repeat-containing proteins, however, the substantial phenotypic variation observed within each disease has yet to be fully explained (Jones et al., 2017). Factors influencing the trinucleotide repeat disorder phenotypes include the size of the repeat, which is the primary determinant for age at onset, the trinucleotide repeat sequence base pair configuration and genetic modifiers (GEM-HD, 2019; Menon et al., 2013). Due to the inverse relationship between the length of the repeat and disease onset and severity, the size of the repeat is used to clinically diagnose the disorder and to predict the age at onset (Duyao et al., 1993; Filla et al., 1996). However, using the size of the repeat as the sole age at onset predictor does not account for all of the phenotypic variability observed.

Multiple components of DNA repair pathways including mismatch repair, base excision repair, and nucleotide excision repair have been implicated in generating further repeat expansions in which the increased somatic repeat expansion rates have been linked to a more severe disease phenotype and an earlier age at onset (Gomes-Pereira, 2004; Manley et al., 1999; van den Broek et al., 2002). Ablation of the *Msh2* gene in HD mouse models, which is involved in mismatch repair, eliminates CAG repeat instability thus preventing further CAG repeat expansions and abrogated striatal mutant huntingtin causing a significant delay in nuclear huntingtin mutant protein accumulation (Manley et al., 1999; Wheeler et al., 2003). This reinforces the role of somatic repeat instability as a disease modifier in addition to DNA repair pathway components (Swami et al., 2009). A genome wide association study (GWAS) of HD patients identified variants at a number of loci in or near genes encoding DNA repair pathway components, with many involved in mismatch repair, that had significant associations with age at disease onset (GEM-HD, 2019, 2015). Bettencourt *et al*., 2016 extended this further to include other polyglutamine diseases including HD and SCA1, 2, 3, 6, 7, and 17, which identified many of the same DNA repair genetic modifiers associated with age at onset (Bettencourt et al., 2016). This suggests a common pathogenic mechanism at the level of the somatic CAG trinucleotide repeat.

Repeat sequence interruptions are additional factors influencing phenotypic variability in some of the trinucleotide repeat diseases. Both synonymous and non-synonymous CAG repeat sequence interruptions, which describes an alteration in the triplet repeat sequence that leads to the same or a different amino acid being translated, respectively, have been reported as disease modifiers (GEM-HD, 2019; Menon et al., 2013; Wright et al., 2019).

Individuals who carry SCA1 CAG repeat expanded alleles with sizes ranging from 6 to 44 CAGs, and are interrupted by at least one histidine (CAT) trinucleotide when the tract exceeds 21 CAGs, are phenotypically normal, even though the pathogenic threshold is 39 CAGs (Chung et al., 1993; Nethisinghe et al., 2018; Quan et al., 1995). Additionally in HD, individuals with an expanded pure CAG repeat have an earlier disease onset than individuals with an interrupting CAA (glutamine) codon, despite the same overall polyglutamine length (GEM-HD, 2019; Wright et al., 2019). This reinforces that the specific codon composition is an additional factor to repeat length in determining the age at onset and highlights the importance of elucidating the exact sequence of the repeat at the base pair level.

The length of the trinucleotide repeat is usually determined by polymerase chain reaction (PCR) to amplify the genomic region of interest in which the repeated DNA motif is sized by various methods including capillary electrophoresis, gel electrophoresis and southern blot analysis (Haddad et al., 1996; Hsiao et al., 1999; Lyon et al., 2010). However, these methods are not without their limitations in that they are typically time-consuming and labour-intensive, cannot be performed in high-throughput screening studies, and have difficulty in long read determination and sequencing GC-rich repeat regions (Liu et al., 2017). As the trinucleotide repeat length in patients can exceed the length of the "sequenceable" repeat sizes and can have greater than 80% GC contents, current long read sequencing technologies by Pacific Biosciences and Oxford Nanopore Technologies have overcome these limitations and offer the determination of greater than 10,000 base pairs of genomic DNA sequence (Liu et al., 2017). Ultimately, the precise determination of repeat length and sequence composition will allow a better understanding of the genotype-phenotype correlation that will further lead to an improved understanding of the disease, which is essential for diagnosis and prognosis.

1.2 Friedreich's Ataxia

FRDA is an autosomal recessive, adolescent-onset neurodegenerative disorder and is the most common form of inherited ataxia. The prevalence of FRDA in Western populations varies between one in 20,000 and one in 725,000 individuals, with a prevalence gradient in Europe (in descending order) from South of France, North of Spain and Ireland to Scandinavia and Russia (Vankan, 2013). FRDA is characterised as multi-systemic, encompassing not only the neurological features: poor balance, impaired coordination, dysarthria, weakness, ocular fixation instability, deep sensory loss, and visual and hearing impairment; but also, diverse non-neurological features including hypertrophic cardiomyopathy, diabetes mellitus, kyphoscoliosis, and foot deformities (Parkinson et al., 2013; Reetz et al., 2015). Approximately 66% of FRDA patients have cardiomyopathy and up to 30% have diabetes mellitus. Apart from a multidisciplinary approach to manage the presenting symptoms, there is currently no disease-modifying treatment to alter FRDA disease progression.

The pathological consequence of the mutation in FRDA is the deficiency of the frataxin protein which results in the accumulation of intra-mitochondrial iron, defective mitochondrial respiration and an over production of oxygen free radicals (Campuzano et al., 1996). The level of gene silencing correlates with the length of GAA1 (Ohshima et al., 1998). Specifically, *FXN* mRNA levels were determined to be reduced to 19.4%, 50.4% and 53% in FRDA patients, late-onset FRDA patients and FRDA carriers, respectively (Saccà et al., 2011). However, a considerable overlap in mRNA and frataxin levels between FRDA patients, carriers and controls was also confirmed, which suggests that reduced mRNA and frataxin levels are not the sole factors in determining FRDA disease manifestation (Saccà et al., 2011).

The most promising treatments for FRDA focus on anti-oxidant therapy and improving mitochondrial function. Omaveloxolone, developed by Reata Pharmaceuticals to increase the transcription of nuclear factor erythroid-derived 2-related factor 2 (NrF2) and induce its antioxidant target genes, is hypothesised to improve mitochondrial function by reducing oxidative stress and preventing lipid peroxidation (Reisman et al., 2019). In part one of a Phase II study, omaveloxolone has been reported to improve the modified FRDA Rating Scale compared to placebo treated FRDA patients, which indicates a slowing of disease progression (Lynch et al., 2019). Currently in part two of the Phase II study, the safety and efficacy of omaveloxolone long term in FRDA patients is being assessed (ClinicalTrials.gov Identifier: NCT02255435).

1.2.1 Genetics

FRDA results from an unstable GAA repeat expansion situated in intron 1 of the *frataxin* (*FXN*) gene, which is located on the proximal long arm of chromosome 9 (Montermini et al., 1995). Approximately 96% of FRDA patients have homozygous GAA repeat expansions ranging from 44 to 1,700 GAAs, with 600 to 900 GAAs being most common (Campuzano et al., 1996; Parkinson et al., 2013; Schmucker and Puccio, 2010). The shorter expanded allele is commonly referred to as GAA1 with the longer allele referred to as GAA2. The remaining 4% of FRDA patients are compound heterozygous for one GAA repeat expansion and a second *FXN* mutation, such as nonsense, frameshift, missense or splice site mutations (van den Ouweland et al., 2012). Additionally, intragenic deletion of *FXN* exons 2 and 3, exon 5a, and complete *FXN* deletion have been described in heterozygous FRDA patients (van den Ouweland et al., 2012). Heterozygous individuals with one GAA expansion but without any other abnormality within *FXN* are not thought to be clinically affected (Parkinson et al., 2013). In unaffected individuals, the length of the GAA repeat ranges from 6 to 27 GAAs, however rare cases of 33 to 130 GAAs have been identified. Fully penetrant alleles contain 66 or more GAAs, with 44 GAAs being the shortest repeat length associated with disease (Cossée et al., 1997; Montermini et al., 1997). FRDA individuals carrying 100 to 500 GAAs in GAA1 commonly present with a late-onset atypical FRDA phenotype.

1.2.2 Neuropathology

FRDA neuropathology is characterised by the degeneration of the dorsal root ganglia, peripheral nerves, the spinal cord and the dentate nucleus in the cerebellum (Koeppen et al., 2009). Specifically, and confirmed *in vivo* by magnetic resonance imaging, there is mild cerebellar atrophy due to the loss of the dentate nucleus and its efferent fibres, which results in superior cerebellar peduncle atrophy (Parkinson et al., 2013). The dorsal root ganglia decrease in size and the dorsal spinal roots become thin and grey. Atrophy of the dorsal columns lead to reduced spinal cord quality, specifically in the thoracic region. Degeneration is also visible in the spinocerebellar and corticospinal tracts (Koeppen, 2011; Parkinson et al., 2013). Prior to full manifestation of FRDA the cerebellum and brainstem are minimally affected. In contrast, FRDA patients show a significant reduction in the total cerebellar volume, specifically affecting the posterior lobe. Pathological and imaging studies have identified that the dentate nuclei are the most affected structures with a subtle mitochondrial iron accumulation preceding degeneration and atrophy (Cocozza et al., 2020).

1.2.3 FRDA Age at Onset

The mean age at onset in FRDA is 15.5 years, with most cases developing before 25 years of age. Early onset FRDA is defined as onset before 10 years of age whereas FRDA individuals with disease onset after 25 and 40 years of age have phenotypes of late-onset and very late-onset, respectively (Dürr et al., 1996). FRDA patients with an early age at onset have a faster disease progression as determined by the Scale for the Assessment and Rating of Ataxia (SARA), which assesses eight impairments associated with cerebellar ataxia (Dürr et al., 1996). FRDA patients with disease onset before 8 years of age progress most rapidly compared to those with onset after 15 years of age (Figure 1.2) (Pandolfo, 2020). The age at onset is inversely correlated with the length of GAA1, the shorter GAA expanded allele, with a prediction of a two to three year earlier onset for every 100 GAAs added to GAA1. Thus, the smaller of the two alleles is the primary determinant for FRDA age at onset (Parkinson et al., 2013; Reetz et al., 2015). The combination of onset before 20 years and cardiac involvement is associated with a more severe disease progression (De Michele et al., 1996). In contrast, FRDA patients with late onset or very late onset have a slower and milder disease progression. Non-neurological symptoms including cardiomyopathy, diabetes, or skeletal deformities are less common in these later phenotypes, which are characterised as more spastic in nature (Ragno et al., 1997). However, there is an extreme phenotypic variability in FRDA, with GAA1 accounting for only 36% to 56% of the variation in age at onset (McDaniel et al., 2001). This suggests that there are additional factors influencing the age at onset such as somatic mosaicism, interruptions in the repeat sequence, and other modifying genes or environmental aspects (Filla et al., 1996; Pandolfo, 2009; Reetz et al., 2015).

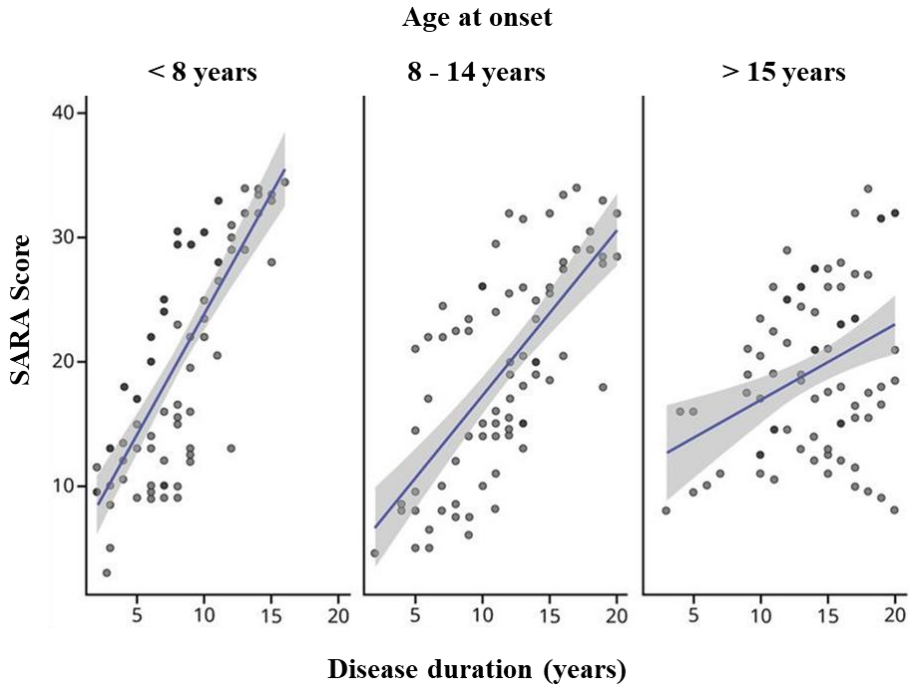


Figure 1.2. SARA score and disease progression in FRDA patients adapted from (Pandolfo, 2020) copyright licence obtained: 4887521069890

Linear regression analysis of FRDA patients with onset before 8 years of age determined a more rapid disease progression compared to those with onset between 8 and 14 years of age ($p = < 0.001$) and those with onset after 15 years of age ($p = < 0.001$), which was statistically significant between all comparisons.

1.2.3 GAA Repeat Sequence

The pathogenic GAA repeat mutations have previously been characterised only in terms of their overall repeat size, rather than actual sequence content. Short GAA expansions up to approximately 130 GAAs have facilitated full sequencing to determine the sequence composition, which subsequently revealed interruptions within the GAA repeat tract (Cossée et al., 1997; Montermini et al., 1997; Ohshima et al., 1999). Such interruptions included $(GAGGAA)_{5\text{-}9}$ or $(GAAGGA)_{65}$ sequences, which were associated with either the absence of disease phenotype or an atypical mid late-onset or very late-onset disease phenotype (Cossée et al., 1997; Ohshima et al., 1999; Sakamoto et al., 2001b; Stolle et al., 2008). Studies have shown that the pure GAA repeat tract is stable up to 44 GAAs and after this, instability occurs. Somatic instability of the GAA1 allele was determined in two FRDA patients who carried small GAA1 alleles of 44 or 66 GAAs and a large GAA2 allele, while their sibling carried a GAA1 allele of 37 GAAs and was clinically asymptomatic (Sharma et al., 2004). Additionally, small-pool PCR (SP-PCR) analysis in FRDA carrier blood samples determined that 107 pure GAAs were unstable compared to 114 GAAs that were interrupted, $(GAA)_{76}(GAGGGA)(GAA)_{18}(GAGGAA)_{5}(GAA)_{8}$, which reinforces the stabilising role of the $(GAGGAA)_{n}$ hexanucleotide interruption that is common in FRDA alleles with greater than 27 GAAs (Pollard, 2004). Similarly, a $(GAA)_{90}(GAAAGAA)_{2}(GAA)_{20}$ interruption within a $(GAA)_{112}$ repeat was stably transmitted through two generations (Cossée et al., 1997). The investigation of GAA repeat sequence interruptions has mostly been carried out in the GAA1 allele, as it is technically too difficult to obtain an accurate sequence of the entire repeat in the longer GAA2 allele. In addition, *in vitro* studies have shown that interrupted GAA repeats inhibit non-B form DNA secondary structure formation, which alleviates transcription inhibition and reduces repeat instability (Ohshima et al., 1999; Sakamoto et al., 2001b). Therefore, this suggests that interruptions have the potential to influence *FXN* expression levels and reduce the instability of the GAA repeat, thus impacting upon FRDA disease progression (Al-Mahdawi et al., 2018).

1.3 Huntington's Disease

HD is an adult-onset neurodegenerative genetic disorder that is progressive and ultimately fatal. It is characterised by movement disturbances, cognitive decline and behavioural symptoms (Bates et al., 2015). HD affects approximately 17.2 in 100,000 people of European ancestry (Bates et al., 2015). The current treatments available are limited to therapies to treat symptoms only, as no treatment thus far has been successful to prevent or slow disease progression. As the mutation in HD results in an expanded polyglutamine stretch in the huntingtin protein, which is considered to be the principal toxic agent, therapies are being developed to target the transcription and translation of the *huntingtin* (*HTT*) gene. Therapies aimed at reducing the levels of the huntingtin protein by targeting its mRNA include antisense oligonucleotides (ASOs) and RNA interference, while those directly targeting *HTT* DNA include zinc finger transcriptional repressors and clustered regularly interspaced short palindromic repeats (CRISPR) and the accompanying CRISPR-associated system (Cas) genome editing constructs (CRISPR-Cas) (Wild and Tabrizi, 2017). The most promising clinical trial to date inhibits *HTT* expression with a second-generation 2'-O-(2-methoxyethyl) ASO targeted to *HTT* mRNA, RG6042 (ClinicalTrials.gov Identifier: NCT03842969). It has previously been reported that the complete inactivation of the mouse homologue of the *HTT* gene (*Htt*) in the forebrain and testis resulted in a progressive degenerate neuronal phenotype and sterility (Dragatsis et al., 2000). However, the partial huntingtin reduction of 50% or more is well tolerated across animal models, with the longest primate study showing no toxicity after 6 months of partial suppression in the striatum (Grondin et al., 2012). The safety, tolerability, pharmacokinetics and pharmacodynamics of ascending doses of the RG6042 ASO (10 mg, 30 mg, 60 mg, 90 mg, 120 mg), were assessed in 46 HD patients, of which 34 were randomly dosed with ASO and 12 with placebo (Tabrizi et al., 2019). ASO or placebo was administered as a bolus intrathecal injection every 4 weeks for four doses, which resulted in a dose-dependent reduction of mutant huntingtin in the cerebrospinal fluid (Tabrizi et al., 2019). The next phase of this study will evaluate the long-term safety and tolerability of RG6042 in HD patients (ClinicalTrials.gov Identifier: NCT03842969).

1.3.1 Genetics

HD is inherited in an autosomal dominant manner, has a single gene aetiology and is fully penetrant when 40 or more CAGs are present in exon 1 of the *HTT* gene (MacDonald, 1993). Translation of the expanded CAG repeat leads to the formation of a pathogenic polyglutamine stretch in the mutant huntingtin protein, which acquires a toxic gain of function with the propensity to misfold and form intra-nuclear and cytoplasmic aggregates in neuronal cells. Wild-type alleles can contain up to 35 CAGs, with 27 to 35 CAGs considered to be in the intermediate range. The majority of HD individuals worldwide have between 40 and 55 CAGs (Bates et al., 2015). Alleles containing 60 or more CAGs are associated with onset before age 21, defined as juvenile-onset HD (Quigley, 2017), with onset at or before 10 years often being described as childhood-onset HD (Quarrell et al., 2013). Alleles containing 36 to 39 CAGs present with reduced penetrance and have an increasing chance of causing disease within a normal life-span (Rubinsztein et al., 1996). The smallest CAG repeat that has been associated with the HD phenotype is 36 CAGs, yet elderly asymptomatic individuals with 36 to 39 CAGs also indicate that this CAG repeat range is incompletely penetrant over the average human lifespan (Kay et al., 2016). Additionally, due to the meiotic instability of the CAG repeat in the germline driving anticipation, offspring of individuals carrying 27 to 35 CAGs are at risk of inheriting longer CAGs that may enter the reduced penetrance or full penetrance range and become pathologically relevant. By convention, the clinical diagnosis and age at onset of HD are based on the motor phenotype with 50% to 70% of patients presenting primarily with chorea, which is an involuntary movement characterised by brief, abrupt, unpredictable, and irregular movements. However, the remaining 30% to 50% of patients first present with cognitive or mood changes (Bates et al., 2014).

1.3.2 Neuropathology

The neuropathological profile in HD is characterised by bilateral symmetrical neuronal loss in the striatum of patients, which is caused by the extensive degeneration of GABAergic medium spiny neurons. These medium spiny neurons are the primary targets of striatal input and provide the main efferent output of the striatum (Rüb et al., 2016). *Post-mortem* examination in the human HD brain revealed a 30% reduction of total brain weight and in addition to the striatal neuronal loss, enlarged lateral ventricles are the most striking features (de la Monte et al., 1988). As the disease progresses, there is subsequent volumetric reduction of the globus pallidus, neocortex, thalamus, subthalamic nucleus, substantia nigra, white matter and the cerebellum. Ultimately, there is widespread

neuropathological changes by end-stage disease. A grading system of 0 to 4 details the gradual striatal neuronal loss and reactive gliosis, which has an ordered and topographic distribution (Vonsattel et al., 2011). In ascending order of severity, grade 0 correlates with no discernible neuropathological abnormalities, suggesting that the anatomical changes are slower than the clinical abnormalities. Neuropathological changes observed microscopically, denoted as grade 1, include neuronal loss and gliosis, which are most distinct in the caudate nucleus. Neuronal counts in the caudate nucleus reveal that 50% are lost in grade 1, which increases dramatically to 95% by grade 4, and correspondingly, the number of astrocytes are greatly increased in grades 2 to 4. Atrophy of the caudate nucleus at the macroscopic level describes grade 2, with microscopic examination revealing marked neuronal loss and astrocytosis of the caudate nucleus. Grade 3 is characterised by decreased volume of the caudate nucleus and by neuronal loss and gliosis in the putamen. The caudate nucleus is severely degraded in grade 4, and the putamen is extremely atrophic (Vonsattel et al., 1985).

1.3.3 Age at Onset and Phenotypic Variability

Defining disease onset is often difficult as the transition from premanifest to manifest is gradual, and the psychiatric and cognitive changes are often not concurrent with motor onset (Moss et al., 2017; Ross et al., 2014). The symptomatology of HD patients has similarly been shown to be inconsistent, even among those with identical CAG repeat sizes (Thu et al., 2010; Tippett et al., 2007; Wexler et al., 2004). The reason for this phenotypic variability is currently unknown. Symptom onset and disease progression are primarily correlated with the number of CAGs; longer CAGs cause an earlier onset that is usually accompanied by a more severe phenotype (Bates et al., 2015). However, CAG repeat length only accounts for approximately 44% of the age at onset variability in HD patients harbouring the commonest expansion lengths. Previous reports determined the cumulative probability of symptom onset for a given CAG repeat length at five year intervals, which revealed that as the CAG repeat length increased, there was a simultaneous increase in the probability of onset for a given age (Brinkman et al., 1997). In a cohort of 866 HD patients with 39 to 50 CAGs, the study determined that an individual with 40 CAGs only had a 13% probability of having onset by 45 years, which increased to 32% for individuals with 42 CAGs, 73% for individuals with 44 CAGs and 100% for individuals with 46 CAGs (Brinkman et al., 1997). This emphasises the variability that exists within the most common CAG repeat lengths.

Age at onset variation is also evident within patients containing the same CAG repeat length. In 443 heterozygous HD individuals with 40 to 86 CAGs, the length of the expansion was plotted against age at onset (Figure 1.3 (A)) (Wexler et al., 2004). Overall, the curvilinear relationship between these two variables represents their negative correlation and determines that the size of the CAG repeat accounts for 72% of the variance in this cohort's age at onset. For the HD individuals with 40 to 55 CAGs, which comprises the majority of the cohort, the CAG repeat is less strongly correlated with age at onset, accounting for approximately 44% of the variance (Wexler et al., 2004). Langbehn *et al*., 2004 similarly portrays the inverse relationship between the CAG repeat length and the mean age at onset by estimating HD age at onset distributions using a database of 2,913 patients contributed by forty HD centres worldwide, and modelling this against expansions of 36 to 60 CAGs (Figure 1.3 (B)) (Langbehn et al., 2004). The graph cannot be used to predict any individuals age at onset with absolute certainty, instead the data provides an estimated mean age at onset based on the CAG repeat length (Langbehn et al., 2004). After the CAG repeat length has been controlled for, the remaining variance in age at onset was not accounted for by the size of the wild-type allele, any of the parental alleles, or gender. The residual phenotypic variability has been attributed to additional modifier genes and environmental factors (Aziz et al., 2018; Wexler et al., 2004).

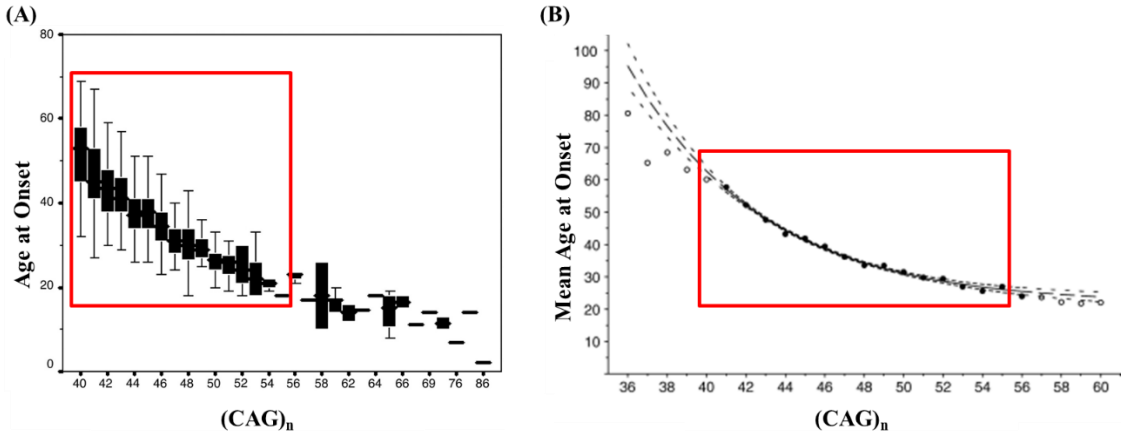


Figure 1.3. The relationship between age at onset and CAG repeat length, adapted from (Wexler et al., 2004), Copyright 2004 National Academy of Sciences and (Langbehn et al., 2004), copyright licence obtained: 4503471225719

(A) The box plot displays the curvilinear relationship between age at onset and CAG repeat length. (B) The Langbehn *et al*., 2004 model uses survival analysis to estimate HD onset distributions. The distribution of CAG repeat length between 41 and 56 CAGs were modelled against age at onset using a nonstandard parametric survival model, which also offered extrapolations for the repeat range of 36 to 40 CAGs. The solid line and solid circles indicate the range and means, respectively, of the data that was used to fit the exponential curves. The dashed line and unfilled circles indicate the CAG repeat length for which the model's predictions were extrapolated. The 95% confidence interval is shown by the small dashed lines, where the larger spaces between dashes indicate the region where the model's predictions were extrapolated (Langbehn et al., 2004). Red box: HD individuals with the most common clinical CAG repeat lengths (40 to 55 CAGs); $(CAG)_n$: CAG repeat of length n (n = number).

1.3.4 Modifiers of HD Age at Onset

Understanding the influence of the CAG repeat base pair configuration on disease pathogenesis in HD is important as changes to the genotype have relevant clinical implications. The most common sequence composition found in greater than 95% of European ancestry HD chromosomes is $(CAG)_n(CAA)(CAG)$ (GEM-HD, 2019). Sequence variants have previously been reported in HD in both the CAG repeat and the directly flanking CCG repeat (Table 1.1). The carboxyl-terminal side of the CAG repeat consists of two pure proline (CCG) tracts (denoted from here on as CCG1 and CCG2, respectively), which are separated by a leucine-proline rich region (Caron et al., 2013). CCG1 has been reported to vary in length from 7 to 12 CCGs. In addition, the CCT repeat following CCG1 is most often reported as $(CCT)_2$ with $(CCT)_3$ being more rare (Pêcheux et al., 1995). Clone sequencing of the CAG repeat determined that the absence of the penultimate CAA trinucleotide in HD patients is associated with marked intergenerational instability and predisposes patients to an earlier age at onset (Goldberg et al., 1995). Wright *et al*., 2019 further demonstrated that in a cohort of 16 manifest HD patients with the CAA to CAG transition, causing the loss of interruption and a pure CAG repeat tract, 75% carried alleles in the reduced penetrance range (36 to 39 CAGs). This suggests that the effect of this loss of interruption is most evident in HD patients with CAG repeat lengths in the reduced penetrance range (Wright et al., 2019). The loss of interruption variant in the CAG repeat is associated with an additional CCA to CCG transition in CCG1 and HD patients with these two transitions presented approximately 25 years earlier than patients with the CAA and CCA interruptions (Wright et al., 2019). In contrast to the loss of interruption variant, a duplication of the CAA-CAG motif was associated with a delayed age at onset of approximately 4.8 years in the HD patients examined in Wright *et al*., 2019.

Table 1.1. Previously reported *HTT* sequence alterations

Reference	Allele	Sequence
HTT Sequence		TTC (CAG)n CAA CAG CCG CCA (CCG)n $(CCT)_2$
Goldberg *et al.*, 1995	IA	TTC (CAG)n CA**A** CAG CCG CCA (CCG)n
	IA	TTC (CAG)n CA**G** CAG CCG CCA (CCG)n
	IA	TTC (CAG)n CA**G** CAG CCG CC**G** CCG_7
Pêcheux *et al.*, 1995*	HD/WT	TTC (CAG)n CAA CAG CCG CCA $(CCG)_7$ $(CCT)_2$
	HD/WT	TTC (CAG)n CA**A** CAG CAA CAG CCG CCA $(CCG)_7$ $(CCT)_3$
	HD/WT	TTC (CAG)n CAA CAG CCG CCA $(CCG)_{10}$ $(CCT)_2$
Gellera *et al.*, 1996	HD	TTC (CAG)n CA**G** CAG CCG CCA (CCG)n
Margolis *et al.*, 1999	HD	TT**G** (CAG)n CAA CAG CCG CCA (CCG)n
	WT	TTC (CAG)n CAA CAG CCG CC**G** (CCG)n
Yu *et al.*, 2000	HD/WT	TTC (CAG)n CA**A** CAG CAA CAG CCG CCA (CCG)n
	HD	TTC (CAG)n CAA CA**A** CAG CCG CCA (CCG)n

The *HTT* sequence was obtained from the National Centre for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/nuccore/NC_000004.12?report=fasta&from=3074681&to=3243960). Substitutions and trinucleotide variations previously detected (bold and underlined) in the CAG and CCG repeats in human samples. *: Pêcheux *et al.*, 1995 includes the sequence information for the CCG1 repeat region; WT: wild-type allele; HD: pathogenic allele with 40 or more CAGs; IA: intermediate allele, defined by Goldberg *et al.,* 1995 as containing 29-35 CAGs, in the general population and/or sporadic HD cases.

Targeted candidate gene studies have identified modifier loci in the following genes which accounts for some of the residual variability observed in HD age at onset after controlling for CAG repeat length; ubiquitin C-terminal hydrolase 1 (*UCHL1*), adenosinergic A2A receptor (*ADORA2A*), autophagy-related protein 7 (*Atg7*) and peroxisome proliferator-activated receptor-γ coactivator 1α (*PPARGC1A*) (Sun et al., 2017). HD individuals carrying the S18Y polymorphism in *UCHL1* and the G/G genotype of the rs7665116 SNP in *PPARGC1A* have a later age at onset (Weydt et al., 2009; Xu et al., 2009). In contrast, the T/T genotype of the rs5751876 SNP in *ADORA2A* and the V471A polymorphism in *Atg7* has been reported to advance age at onset in HD (Dhaenens et al., 2009; Metzger et al., 2010). Additional modifier loci have been identified in the following genes; huntingtin-associated protein-1 (*HAP1*), apolipoprotein E (*APOE*), and GluR6 subunit of kainite receptor (*GRIK2*), however, the extent of their modifying role on HD age at onset is controversial (Sun et al., 2017).

Recent studies have identified genetic modifiers in HD involved in DNA repair-related processes (Bettencourt et al., 2016; GEM-HD, 2019). A GWAS was carried out on over 9,000 HD individuals using the difference between age at onset predicted by CAG repeat length and actual age at onset of motor symptoms, referred to as residual age at onset (GEM-HD, 2019). This work expands on the previous GWAS of 4,082 HD patients with 40 to 55 CAGs, which reported three significant modifier signals at one loci on chromosome 8 (*RRM2B/UBR5*) and two loci on chromosome 15 (*FAN1*) for which two independent opposing effects were identified (GEM-HD, 2015). The largest effect size was from rs146353869 in *FAN1* resulting in a 6.1 year earlier age at onset in HD patients GEM-HD, 2015). A follow up study additionally identified a genome-wide significant signal on chromosome 3 (*MLH1*) (Lee et al., 2017). Increasing the power of the GWAS identified infrequent modifier alleles of strong effect and more common modifiers with a modest impact at new loci. This GWAS also highlighted that the length of the pure CAG repeat is the rate determining driver for age at onset in HD (GEM-HD, 2019). The aforementioned significant modifier signals were recapitulated in addition to new loci identified on chromosome 2 (*PMS1*), 5 (*MSH3/DHFR* and *TCERG1*), 7 (*PMS2*), 11 (*CCDC82*) and 19 (*LIG1*). With the exception of *TCERG1* and *CCDC82*, all of the modifier signals are located in genes associated with DNA repair. In relation to the pure CAG repeat acting as the primary determinant for age at onset and the identification of DNA repair genes as modifiers, it suggests that these DNA repair genes influence age at onset by a DNA-level effect on somatic instability of the CAG repeat.

1.3.5 Somatic Mosaicism in HD

Somatic mosaicism, which describes the presence of genetically distinct cells, occurs from a post-zygotic mutation and, in contrast to inherited mutations, can affect only a portion of the body and are not transmitted to progeny (Campbell et al., 2015; Gonitel et al., 2008; Kraus-Perrotta and Lagalwar, 2016). In the trinucleotide repeat diseases, the repeat is genetically unstable and can undergo size variations, both in the germline and soma. Determining the mutation profiles is essential for the understanding of the contribution of repeat instability to disease susceptibility and severity (Veitch et al., 2007). SP-PCR quantifies the degree of repeat-length variation in a given sample by PCR amplification of the target region in multiple small pools of input DNA. The PCR products are resolved by agarose gel electrophoresis and detected by Southern blot hybridisation using methods that identify products from single-input DNA molecules (Monckton et al., 1995). However, SP-PCR methods used to determine somatic instability are labour intensive and are not designed for high-throughput analyses. Therefore, subsequent techniques were developed to combat this.

A novel method was designed to quantify trinucleotide repeat sizes from bulk genomic DNA, which generates multiple PCR products that can be viewed using GeneMapper software (Lee et al., 2010). The PCR products are represented as a cluster of peaks differing by a single repeat unit. To distinguish signal peaks from background noise, a threshold factor of 20% of the largest peak height is applied with any peak heights below this being excluded. The most common method currently used for somatic instability quantification in HD is the instability index, which represents the mean CAG repeat length change from the modal allele per cell in a given tissue (Figure 1.4) (Lee et al., 2010). After the 20% threshold is applied, the peak heights are normalised by dividing the peak height of each peak by the sum of the heights from all the peaks. The change in CAG repeat length per peak is deduced from the constitutive CAG repeat length of the highest peak, which represents the modal allele. The normalised peak heights are multiplied by the change from the main allele and these values are summed to calculate the instability index (Lee et al., 2010). Similar to the instability index is the expansion ratio, which is calculated using the area under all expanded CAG repeat peaks that are greater than the most prominent peak relative to the area under the most prominent peak (Wright et al., 2019). However, this method does not account for repeat contractions.

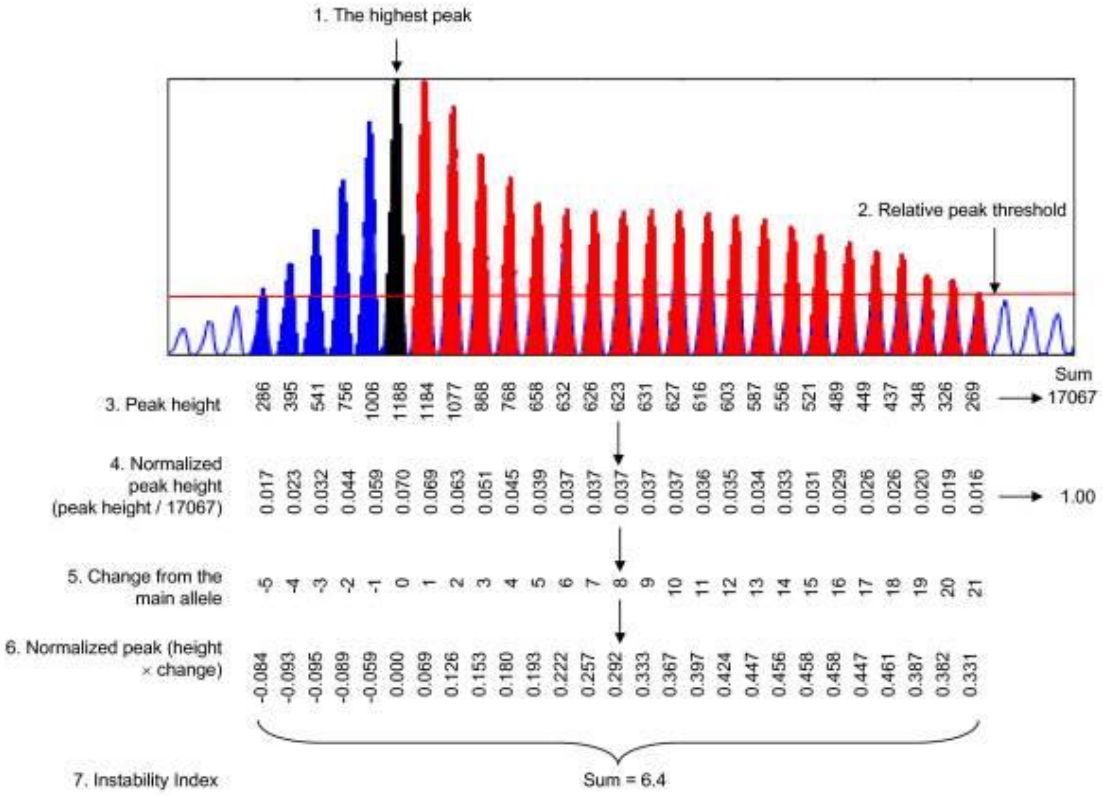


Figure 1.4. Somatic instability index (Lee et al., 2010), Creative Commons Attribution License

Repeat instability was analysed from GeneMapper traces of *Hdh*$^{Q111/+}$ mouse striatum at 5 months of age with 100 ng input genomic DNA. The open, blue, black, and red peaks represent background, contracted alleles, main allele from tail analysis of same mouse, and expanded alleles, respectively. Peak height was used to determine a relative threshold of 20% in which peaks falling below this threshold were excluded. Peak heights were normalised to the total of all peak heights and multiplied by the change in CAG repeat length of each peak relative to the highest peak (main allele). These values were summed to generate an instability index.

Early studies on the somatic mosaicism of the HD CAG repeat was reported by Telenius *et al.* 1994, who hypothesised that the variation in the size of the CAG repeat in different tissues of the same individual may have a role in the tissue-specific effects of the HD gene (Telenius et al., 1994). This study concluded that the somatic mosaicism pattern is not random as it is consistently most obvious in the brain regions which show the most marked neuropathological changes in HD. The basal ganglia and cerebral cortex displayed the largest CAG repeat expansions as compared to the cerebellar cortex, which displayed the lowest degree of CAG instability (Telenius et al., 1994). Additional studies report that the expanded CAG repeats display allele length-, age- and cell type-dependent somatic instability (Gonitel et al., 2008; Kennedy, 2003; Shelbourne et al., 2007; Veitch et al., 2007). CAG repeat length profiles in HD patient *post-mortem* brains have been determined using SP-PCR (Kennedy, 2003). The results from a HD patient who died at 40 years of age with 41 CAGs and Vonsattel grade 0 neuropathology revealed high levels of instability in the striatum compared to the cortex and hypothalamus. Estimated CAG repeat sizes within the striatum of this patient revealed that 10% to 15% of cells contained increases of greater than 20 CAGs, and approximately 2% of cells contained in excess of 200 CAGs. The largest expansion determined was greater than 1,000 CAGs, which is approximately 24-fold greater than the inherited progenitor allele CAG repeat size (Kennedy, 2003). The results suggest that dramatic repeat expansions occur in the most vulnerable brain regions and transpire as an early event in HD pathogenesis, potentially preceding symptom onset.

Swami *et al*., 2009 explored the role of CAG repeat somatic instability as a modifier of HD. Somatic instability was determined by SP-PCR within the *post-mortem* cortex from a cohort of HD patients who exhibited phenotypic extremes of early and late age at onset as predicted by their cerebellar CAG repeat length (Langbehn et al., 2004; Swami et al., 2009). Extreme early and extreme late phenotypes were designated by age at onset less than 0.5 standard deviation below the mean and greater than 0.5 standard deviation above the mean age at onset, respectively. The two cohorts consisted of 24 HD patients each and had estimated mean age at onsets differing by approximately 30 years. The frontal cortex was the region targeted for this study as in contrast to the striatum, which displays low levels of somatic instability in end-stage disease, the frontal cortex has previously been shown to display relatively high levels of somatic instability in HD *post-mortem* brains (Kennedy, 2003). SP-PCR analysis determined various degrees of somatic instability with a dominant expansion bias, which was more prominent in the HD patients

with an earlier age at onset. The results revealed that on average, approximately 50% of pathogenic alleles expanded further by at least one CAG, 22% expanded further by at least five CAGs and 11% had further expansions of at least 10 CAGs (Swami et al., 2009). More rare observations included increases of at least 35 CAGs and as great as 68 CAGs. Additionally, a marked statistical difference was revealed in the magnitude of the average maximum expansion for each cohort with the early age at onset group having a mean of 42 CAGs compared to the late age at onset group with a mean of 29 CAGs. This data suggests that somatic expansions of the CAG repeat contributes to HD disease progression (Swami et al., 2009).

Swami *et al*., 2009 subsequently used skewness, a measurement of the degree of symmetry of a distribution determined by fragment analysis, to measure the levels of somatic instability. Early age at onset HD patients exhibited a greater right skewness than those with late age at onset, which is in keeping with the assumption that as the CAG repeat length changes are bias towards expansion, their distributions are skewed to the right. A negative correlation was determined between skewness and residual age at onset, which reinforces the association between greater somatic expansion and an earlier age at onset (Swami et al., 2009). Similarly, regression analysis revealed that skewness was a significant predictor of residual age at onset, with an increase in right skewness being associated with a lower residual age at onset (Swami et al., 2009). Overall, the results demonstrate that larger somatic expansion of the cortical CAG repeats are significantly associated with an earlier age at onset, which is independent of any effects of constitutive CAG repeat length on both somatic instability and age at onset (Swami et al., 2009).

The loss of interruption variant identified in Wright *et al*., 2019, which results in a pure CAG repeat tract, was similarly reported to be associated with increased CAG repeat instability in both somatic and germline tissues. The CAG repeat somatic expansion ratio, which is calculated using the area under all expanded CAG repeat lengths relative to the modal CAG repeat length, was determined in the blood of HD patients with the loss of interruption variant and those with the canonical sequence, $(CAG)_n(CAA)(CAG)$. After correcting for CAG repeat length and age at onset, the somatic expansion ratio was significantly higher in the HD patients with the loss of interruption (Wright et al., 2019). Additional SP-PCR analysis in sperm revealed that loss of interruption HD patients had greater CAG repeat lengths in sperm and increased germline CAG repeat instability compared to those with the canonical sequence. This further supports the role of somatic instability in HD progression (Wright et al., 2019).

1.3.6 DNA Repair in HD

The recent GWAS in HD has highlighted DNA repair genes as modifiers of HD phenotype and previous reports in HD cell and mouse models, and patient clinical samples, has identified a progressive level of DNA damage in HD (Castaldo et al., 2018; Lu et al., 2014; Maiuri et al., 2016). To combat DNA damaging insults, cells have developed a specialised DNA repair system, which is sub-divided into several distinct mechanisms tailored to remove specific types of DNA lesions (Figure 1.5). Mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER) and double-strand break repair (DSBR) are examples of such distinctive DNA repair mechanisms and comprise both non-homologous end-joining and homologous recombination (Dexheimer, 2013). These DNA repair pathways collectively repair all types of DNA damage through the DNA damage response, which involves a highly co-ordinated cascade of steps; lesion recognition and repair factor recruitment, DNA strand breakage through excision of the lesion, DNA end processing, and DNA synthesis to complete the repair (Dexheimer, 2013; Yuan et al., 2012). Trinucleotide repeat regions are mutational hotspots in the genome that can readily form secondary DNA structures including slipped strands, hairpin loops, G-quadruplexes and R-loops (Mirkin, 2007). These secondary structures can lead to trinucleotide repeat instability through aberrant DNA processing and result in repeat length variation, which can produce phenotypic differences. In relation to certain CAG repeat diseases, these differences have the potential to act as disease modifiers. As has been shown in several GWAS studies, components of the DNA repair pathways have been genetically proven to significantly modify HD age at onset (Bettencourt et al., 2016; Flower et al., 2019; GEM-HD, 2019, Lee et al., 2017).

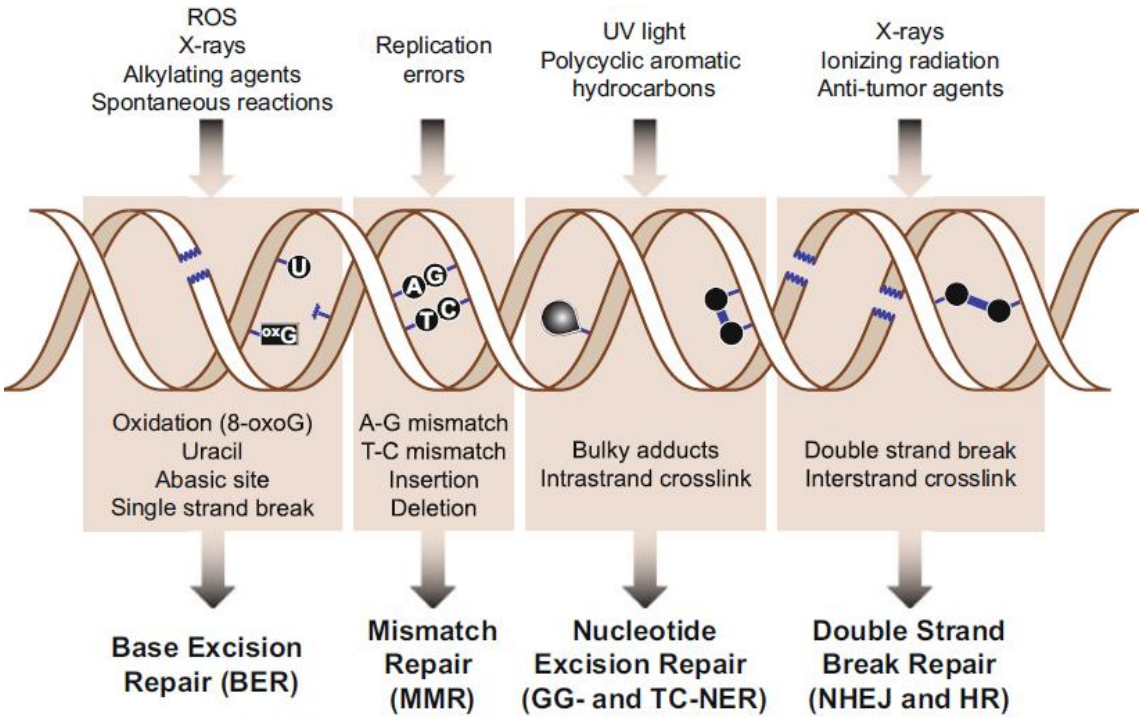


Figure 1.5 DNA damage and associated DNA repair pathways (Dexheimer, 2013), copyright licence obtained: 4533080518165

Common causes of DNA damage are named above, along with their associated lesions and the relevant DNA repair pathway responsible for removing the DNA lesions. DNA damage is induced spontaneously during cellular metabolism or by environmental agents. Spontaneous DNA damage can occur from reactive oxygen species (ROS), generating oxidised DNA bases and DNA breaks, which activates BER, and dNTP misincorporation during DNA replication, which subsequently recruits the MMR components. Additionally, environmental DNA damage can be generated by physical and chemical sources including ultraviolet (UV) light, ionising radiation (IR), and anti-tumour agents. UV light can induce bulky adducts, such as pyrimidine dimers (generally thymine dimers), which are recognised and repaired by NER. IR generates SSBs and DSBs that are restored by DSBR through the gap repairing mechanisms of non-homologous end joining and homologous repair (Ciccia and Elledge, 2010). BER: base excision repair; MMR: mismatch repair; NER: nucleotide excision repair; SSBs: single-strand breaks; DSBs: double-strand breaks; DSBR: double strand break repair; NHEJ: non-homologous end joining; HR: homologous repair.

The MMR pathway is represented by several SNPs in the genes revealed by the GWAS including *MSH3, MLH1, PMS1,* and *PMS2*. MMR is essential for the maintenance of genomic stability. It functions by correcting DNA replication errors, attenuating chromosomal rearrangements, and mediating the cellular response to certain types of DNA damage (Gorbunova et al., 2007). Data from various expanded CAG repeat mouse models has revealed an essential role for components of the MMR pathway, including Msh2, Msh3, Mlh1, Mlh3 and Pms2 (Morales et al., 2016). The HD mouse model *Hdh*(Q111), which carries 111 CAGs at the mouse *Htt* homolog, was bred on both the C57BL/6 (B6.*Hdh*(Q111)) and 129 genetic background. Somatic instability in the striatum, liver and tail of these mice was assessed by the somatic instability index (Lee et al., 2010). In comparison to the 129 genetic background, higher levels of somatic instability were observed in the striatum and liver of *Hdh*(Q111) mice bred on the C57BL/6 genetic background (Pinto et al., 2013). While the CAG repeat length was attributed to some of the variance in the instability index between the two genetic backgrounds, it could not account for it all. This indicated the possibility of additional genetic modifiers. The striatal somatic instability index was subsequently used as a quantitative phenotype for linkage mapping as the distinction between the instability index of the two strains was most evident in the striatum compared to the liver. The strain specific difference was attributed to a single quantitative trait locus identifying the MMR gene *Mlh1* as the phenotypic modifier. *Mlh1* dimerises with *Mlh3* forming the MutLγ complex, which is thought to be preferentially recruited to the site of DNA damage to carry out MMR. B6.*Hdh*(Q111) mice were crossed with *Mlh1* null mice, which eliminated the instability seen previously in the striatum and liver and demonstrated that *Mlh1* is essential for CAG repeat instability. Additionally, crossing B6.Hdh(Q111) mice with *Mlh3* null mice abolished CAG repeat instability, which highlights the MutLγ complex as a key driver of somatic expansion in this HD mouse model (Pinto et al., 2013).

A common characteristic between the DNA repair pathways is functional redundancy, which describes the ability of some of the repair pathway components to participate in multiple independent pathways (Zhao et al., 2009). Components of the MMR pathway have been shown to interact with those involved in inter-strand cross-link repair, with FAN1 compensating for the loss of the EXO1 exonuclease, which mediates DNA excision during MMR activity (Desai and Gerson, 2014). *FAN1* was subsequently identified as one of the most significant hits by the recent GWAS having at least two independent signals shown to modify HD age at onset, with one advancing and one

delaying onset (GEM-HD, 2019). Goold *et al*., 2018 investigated how *FAN1* expression modifies the HD phenotype and found that reduced expression or function results in a hastened onset and increased expression leads to delayed onset with slower disease progression (Goold et al., 2018). Specifically, the lowering of *FAN1* expression in the U20S cell line expressing mutant HTT exon 1, in HD-patient derived induced pluripotent stem cells and in differentiated medium spiny neurons, increased CAG repeat expansions in a CAG repeat length-dependent manner (Goold et al., 2018). This highlights the protective role of *FAN1* in HD and the influence of DNA repair pathway components as HD modifiers.

1.4 Sequencing Technologies

To fully comprehend the genotype-phenotype correlation in trinucleotide repeat diseases, it is important to detect repeat sizes accurately and determine the sequence configuration. Pathogenic allele repeat length is inversely associated with the severity of trinucleotide repeat diseases and the age at symptom onset. Additionally, sequence interruptions have been reported as modifiers of disease phenotype (GEM-HD, 2019; Menon et al., 2013). DNA sequencing technologies have been continuously developing with advancements from Sanger sequencing to next generation sequencing and the most currently released third generation sequencing platforms (Table 1.2). The development of first generation sequencing, more commonly referred to as Sanger sequencing, marked the breakthrough for DNA sequencing technology using the chain-termination method or the dideoxy technique (Sanger et al., 1977). Subsequent advancements to Sanger sequencing included the replacement of phosphor- or tritrium-radiolabelling with fluorometric based detection and improved detection through capillary based electrophoresis. These advancements led to the development of increasingly automated DNA sequencing machines. Shotgun sequencing was established to sequence longer fragments, greater than 1kb, in which overlapping DNA fragments were cloned and sequenced separately and finally assembled into one long contiguous sequence or contig *in silico* (Heather and Chain, 2016). PCR and recombinant DNA technologies aided in the generation of high concentrations of DNA. Eventually, the ABI PRISM range by Applied Biosystems allowed the simultaneous sequencing of hundreds of samples, which were used in the Human Genome Project (Hood and Rowen, 2013). The use of Sanger sequencing in the Human Genome Project required long run times, was expensive and provided limited throughput. The advancement to second- or next-generation sequencing subsequently reduced run times and costs and increased throughput. Next generation sequencing technologies use the sequencing by synthesis method in which a polymerase is used and a signal (fluorophore or a change in ionic concentration) identifies the incorporation of a nucleotide into an elongating strand. The parallelisation of this technology is facilitated by the millions of individual sequencing by synthesis reaction centres, each with its own DNA template, from which a sequencing platform collects information from many millions of DNA molecules simultaneously (Goodwin et al., 2016).

Table 1.2. Comparison of sequencing technologies

	Sanger sequencing	NGS	TGS
Sample Preparation	Fragmentation, PCR, fluorescently end-labelled bases	Clonally amplified templates, single DNA molecule templates	CRISPR-Cas systems
Physical Sequencing	Capillary electrophoresis	SBS/CRT, SNA	SMRT
Re-assembly	Reference genome	Reference genome	CCS
Read Length	800 to 1000 bp	150 to 300 bp	Up to 10,000 bp
Read Accuracy	High	High	Moderate
Throughput	Low	High	Moderate
Cost	High	Low	Low-Moderate

NGS: next generation sequencing; TGS: third generation sequencing; SBS: sequencing by synthesis; CRT: cyclic reversible termination; SNA: single nucleotide addition; SMRT: single molecule real time; CCS: circular consensus sequence; bp: base pairs.

1.4.1 Illumina Sequencing

The Illumina technology of sequencing by synthesis combined with bridge amplification of template molecules is the dominant next generation sequencing platform worldwide (Leggett and Clark, 2017). In contrast to sequencing a single DNA fragment, next generation sequencing extends this process across millions of fragments in a massively parallel fashion (Behjati and Tarpey, 2013). In brief, DNA polymerase catalyses the incorporation of fluorescently labelled deoxyribonucleotide triphosphates into a DNA template strand during sequential cycles of DNA synthesis. At the point of incorporation during each cycle, the nucleotides are identified by fluorophore excitation. Paired-end sequencing is most commonly used, which sequences both ends of the DNA fragments in a library and aligns the forward and reverse reads as read pairs. This produces twice the number of reads for the same time and effort in library preparation (Illumina). More accurate read alignment is enabled by paired-end sequencing and it is also sensitive enough to detect insertions/deletions (indels), which cannot be identified in single-read data. The analysis of differential read-pair spacing can remove PCR duplicates, a common artefact resulting from PCR amplification during library preparation. Paired-end sequencing produces a higher number of single nucleotide variant calls following read-pair alignment (Illumina). Some areas of the human genome are left unresolvable as the Illumina technology amplifies DNA templates, extends with single fluorescent nucleotides, and images each step, which limits the read lengths. The short 100 to 400 base pairs read lengths obtained are due to inevitable phasing when the templates in a polymerase colony lose synchronicity (Leggett and Clark, 2017). These short read lengths make genome, transcriptome, and metagenome assembly more challenging. DNA regions specifically suffering from this limitation include extreme GC sequences, tandem repeat sequences and interspersed repeats (Nakano et al., 2017).

1.4.2 Third Generation Sequencing

The advancement to third generation sequencing technologies no longer relies on PCR amplification, instead, it directly targets single DNA molecules to enable real-time sequencing. The improvements offered by third generation sequencing over the previous methods include increased read lengths, reduced sequencing time and the reduction or elimination of sequencing biases introduced by PCR (Lu et al., 2016). The PacBio RS II instrument was the first commercialised third generation sequencer and uses single molecule real time (SMRT) technology, which enables the direct observation of DNA synthesis by the DNA polymerase. PacBio's subsequent improvement of sequencing chemistries and the release of the new sequencer, the Sequel System, generates approximately ten-fold more sequence data at the cost of two-fold less than that of the RS II instrument (van Dijk et al., 2018). Molecules of up to 2 kb can now be sequenced and the circular consensus sequencing and increased polymerase processivity strongly improves the overall sequencing accuracy. These advantages allow the resolution and analysis of hard-to-sequence regions (Nakano et al., 2017).

Oxford Nanopore Technologies (ONT) sequentially released a third generation sequencing platform using nanopore sequencers. The Nanopore MinION device contains 512 channels, which allows up to 512 independent DNA molecules to be sequenced simultaneously. Specialist MinKNOW software is run on the host computer to carry out data acquisition, real-time analysis and feedback, data streaming and sample identification and tracking (Lu et al., 2016). The read length profile offered by ONT is comparable to that of PacBio, with a maximum length of up to a few 100,000 base pairs. To improve flexibility in throughput, ONT released the PromethION, which can provide a total of 144,000 channels available per run. This results in a theoretical maximum of 15 Tb of sequence data to be obtained per 48-hour run. Depending on the needs of the user, ONT also released the GridION X5 containing 2,560 channels, which can generate up to 100 Gb of data, and the SmidgION, which is even smaller than the MinION and can be controlled by a smart phone (van Dijk et al., 2018). In contrast to PacBio, Nanopore technology is not capable of sequencing the same strand multiple times. For increased accuracy, ONT sequences both strands of a double-stranded DNA molecule by a $1D^2$ process, which uses an adapter with a specialised sequence that promotes the entry of the second DNA strand into the pore after the first strand (van Dijk et al., 2018).

1.6.2.1 Pacific Biosciences

PacBio provides SMRT sequencing, which is a sequencing by synthesis technology based on real-time imaging of fluorescently phospho-tagged nucleotides as they are synthesised along individual DNA template molecules. This method is based on zero-mode waveguide (ZMW) technology (Figure 1.6) (Levene, 2003). A ZMW chamber, tens of nanometers in diameter, prevents visible laser light from passing through completely. Instead, the light exponentially decays and by exposing the ZMW chamber to laser illumination, only the bottom 30 nm lights up. A single DNA polymerase is anchored to the bottom of the ZMW chamber and nucleotides that are phosphate chain-labelled with a fluorophore corresponding to a specific base are flooded above (Schadt et al., 2010). The labelled nucleotides travel into the ZMW chamber by diffusion, surround the DNA polymerase, and exit. Fluorescence only occurs when the correct nucleotide is detected and incorporated. During the incorporation of a nucleotide the fluorescent label emits coloured light, the phosphate chain is cleaved, and the dye molecule is freed leaving a natural piece of DNA. The signal-to-noise ratio is based on the time difference which has a higher signal intensity for incorporated versus unincorporated nucleotides (Schadt et al., 2010). The average read length is approximately 3000 base pairs, which is due to using a DNA polymerase to drive the reaction. As it images single molecules, there is no degradation of signal over time. Instead, the reaction ends when the template and polymerase dissociate (Roberts et al., 2013). The accuracy in SMRT sequencing is achieved by sequencing the same molecule multiple times and deriving a highly accurate circular consensus sequence for each read, thus performing self-error correction. The circular consensus sequences are formed from multiple passes around the circularised DNA molecule (SMRTbell), which can be used to identify real replication errors (Potapov and Ong, 2017).

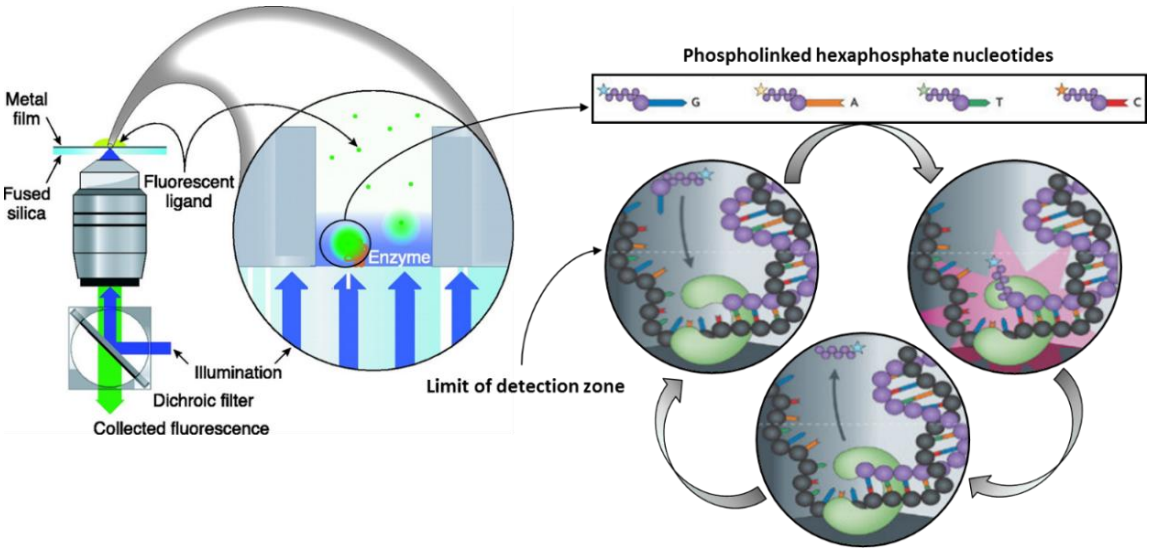


Figure 1.6. SMRT sequencing based on ZMW technology, adapted from (Levene, 2003; Metzker, 2010)

The four colour SMRT sequencing method has a ZMW design that reduces the observation volume, thus reducing the number of stray fluorescently labelled molecules. The ZMW consists of small chambers within a metal film on top of a microscope coverslip. These chambers, in which there can be millions on a single coverslip, allow for the massive parallelism of third generation sequencing. The DNA polymerase is fixed to the bottom of the ZMW chamber in the presence of fluorescently tagged nucleotides. Illumination occurs through the microscope objective from below where the fluorescence is collected back through the same objective. The ZMW detectors allow the DNA polymerase to perform efficiently when the nucleotides are present in the micromolar concentration range. The rate of catalysis governs the residence time of the phospholinked nucleotides in the active site, and corresponds to a recorded fluorescence pulse, as only the bound dye-labelled nucleotide occupies the ZMW chamber on this timescale. The fluorescence signal is reduced to background levels when the dye-labelled pentaphosphate by-product is released and diffuses away. Template translocation marks the interphase period before binding and incorporation of the next phospholinked nucleotide.

1.6.2.2 Oxford Nanopore Technologies

ONT sequencing involves a biological nanopore which is built into an electrically resistant artificial membrane with a voltage applied across the membrane. DNA molecules are prepared with standard library preparation protocols, which include attaching a lead adaptor and motor protein to one strand of DNA. In contrast to PacBio's SMRT sequencing, nanopore sequencing does not sequence the same DNA strand multiple times, instead, both strands of the double-stranded DNA molecule are sequenced. Originally, two-directional (2D) sequencing was developed in which the second DNA strand was sequenced after the first due to a hairpin adapter located on one extremity (van Dijk et al., 2018). The method preferred for obtaining the longest reads is one-directional (1D) sequencing which was developed next using a regular adapter with a specialised sequence promoting the entry of the second strand into the pore directly after the first strand has been passed through. The main advantages of this method are reduced library preparation time and increased yield due to single strand sequencing of each molecule. $1D^2$ libraries, which are more accurate than 1D, are currently dominating this technique where the two strands of the DNA molecule are delivered to the pore noncovalently linked (Leggett and Clark, 2017). Subsequent to the sequencing of the template strand, $1D^2$ relies on the complement strand remaining near the pore and being captured by the pore immediately after the template strand. The sequencing runs last up to 48 hours, with the first 24 hours producing much higher yields due to the gradual decline of the flow cell. Every 8 hours mux scans, which plot total event yield versus time, are performed by the system in order to choose the highest performing nanopore in each channel's group of four (Ip et al., 2015). The nanopore sequencing quality is the same at the beginning and end of the DNA molecule, therefore read length is dependent on the DNA extraction and preparation. During sequencing, the motor protein unzips dsDNA and guides the ssDNA through the pore one base at a time. A deflection in current across the pore is caused by the presence of the DNA molecule, and this current change can be related to the exact bases present in the pore at that moment (Figure 1.7) (Leggett and Clark, 2017).

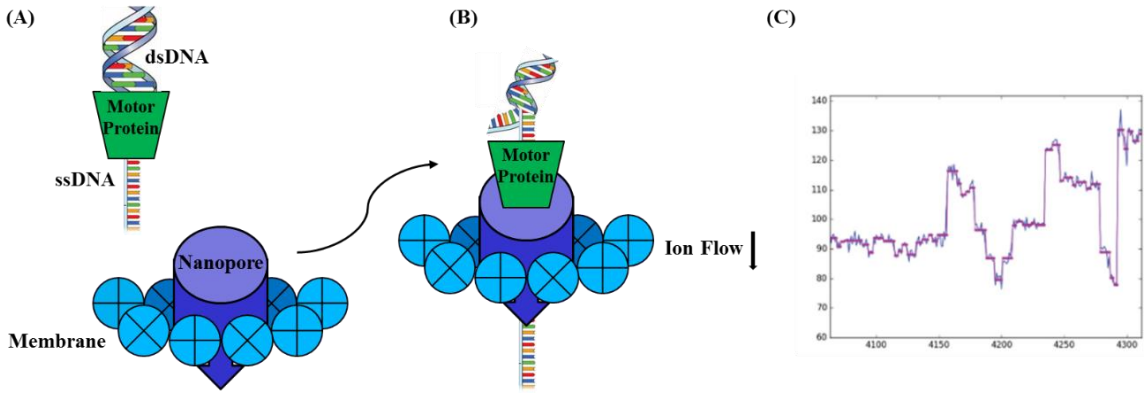


Figure 1.7. Overview of ONT sequencing technology, adapted from (Leggett and Clark, 2017)

ONT technology uses nanopores as biosensors due to the fact that an ionic current passing through a nanopore is dependent on the specific identity of nucleic acid bases. These bases interact with and transition through the nanopore, and the speed of which can be enzymatically controlled allowing a translocation speed on a millisecond time scale (Ip et al., 2015). (A) A biological nanopore is inserted into an electrically resistant synthetic membrane, where a potential is applied across the membrane, resulting in a flow of ions. For a localised library concentration, library DNA molecules preferentially locate to the membrane due to adaptors with aliphatic tethers (not shown). (B) The DNA molecule is passed through the pore when the motor protein bound to the other adaptor docks with it. (C) Disruptions in the current are due to bases in the nanopore which are characteristic of their sequence (blue line). In other basecallers, the signal is further refined to events which correspond to distinct pore kmers, a measurement referring to all possible subsequence's of length K from a given read (purple line).

1.5 Scope of Thesis

In the trinucleotide repeat diseases, the wild-type alleles of the associated disease genes contain either very short repetitive runs, or longer runs with several stabilising interruptions. Pathogenic expansions occur when a repetitive DNA tract exceeds a disease-specific threshold, often because of the loss of these stabilising interruptions (Mirkin, 2007). Once this threshold is overcome, further expansions become progressively more likely, resulting in somatic mosaicism. It is therefore of interest to deduce the exact sequence configuration of the pathogenic repeat region as sequence interruptions have been identified as disease modifiers, which can account for some of the phenotypic variability observed between patients with similarly sized pathogenic alleles (GEM-HD, 2019; Menon et al., 2013; Wright et al., 2019). Similarly, the accurate characterisation of the population of repeat sizes present will aid in the understanding of somatic mosaicism and its relationship with the disease phenotype. The variability in the degree of somatic instability between tissues and individuals is not solely explained by age and progenitor allele size, and highlights the role of genetic factors, such as the DNA repair pathway genes, as additional disease modifiers.

1.5.1 Thesis Aims

- To identify sequence interruptions within the GAA repeat in Friedrich's ataxia and examine the relationship between the sequence composition and age at onset in a large cohort of patients.
- To determine if interruptions in the CAG repeat sequence configuration in a cohort of Huntington's disease patients with similarly sized pathogenic CAG repeats can account for the extremely varied age at onsets.
- To investigate which sequencing technology was best suited to not only size the CAG repeat but to elucidate the sequence at the base pair level.
- To examine the impact of known DNA repair pathway modifiers on the phenotypic and age at onset variability exhibited by our Huntington's disease patient cohort.
- To quantify the somatic mosaicism profile in Huntington's disease patient *post-mortem* brains and assess the relative contribution of small and large CAG repeat length changes to Huntington's disease pathogenesis.

Chapter 2. Materials and Methods

2.1 Ethical Statement

The Friedreich's Ataxia research was conducted under the ethical committee approval of the European Union Seventh Framework Programme [FP7/2007-2013], reference: 242193/EFACTS. The Huntington's disease research was conducted under the ethical committee approval of the London Queen Square NHS Research Ethics Committee at the National Hospital for Neurology and Neurosurgery, reference: 09/H0716/53.

2.2 Patient Samples

2.2.1 Friedreich's Ataxia

Peripheral blood DNA was obtained from 253 samples comprising of 246 FRDA patients and seven carriers. Dr Francesca Cavalcanti and Dr Mark Pook contributed 92 FRDA samples to our 161 UCL FRDA sample cohort. The size of the GAA repeat expansion for the UCL FRDA patient cohort was determined for both alleles by the Neurogenetics Unit (*National Hospital for Neurology and Neurosurgery, Queen Square, London*). These samples were part of the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS) patient London site database (Table 2.1). This study was carried out in collaboration with Dr Mark Pook (*Ataxia Research Group, Department of Life Sciences, Division of Biosciences, Brunel University London, Uxbridge*).

Table 2.1. FRDA patient and carrier samples

Sample #	Code	GAA1/2	Sample #	Code	GAA1/2
BRUNEL Samples			**UCL Samples**		
1	FA1	1023/1258	35	44293	1200/1200
2	FA11	720/760	36	44655	134/1134
3	FA15	**WTC**/+	37	47084	767/1000
4	FA16	**WTC**/720	38	47553	267/1100
5	FA12	**WTC**/520	39	47689	750/912
6	FA13	10/10	40	48978	750/850
7	FA14	10/500	41	49823	300/700
8	FA17	720 /720	42	51041	800/1000
9	FA18	500/720	43	44295	867/1100
10	FA19	**WTC**/900	44	52046	+/+
11	FA20	**WTC**/+	45	52999	200/1000
12	SCA121	730/1040	46	53084	767/967
13	FA31	750/900	47	53085	700/1100
14	FA35	630/730	48	53964	680/880
15	FA36	630/1040	49	54278	645/845
16	FA47	680/840	50	55057	1167/1500
17	FA49	763/1043	51	55070	580/745
18	FA53	850/1000	52	55718	600/967
19	FA61	**WTC**/+	53	55749	500/1000
20	FA62	+/+	54	55830	480/780
21	FA63	567/752	55	55837	667/900
22	FA64	+/+	56	56603	845/845
23	FA66	500/730	57	56994	834/1100
24	FA 75	+/+	58	56999	734/1067
25	FA 76	+/+	59	57261	920/1120
26	FA 77	+/+	60	57683	767/900
27	FA78	765/765	61	58035	785/1020
28	FA79	460/765	62	58666	800/867
29	FA85	+/+	63	59258	450/980
30	FA88	765/765	64	59345	1020/1250
31	FA90	765/1100	65	59923	820/820
32	FA96	430/1245	66	59992	885/1050
33	FA98	1250/1465	67	60541	500/667
34	FA102	+/+	68	67580	612/912
35	FA103	550/10	69	69777	+/+
36	FA104	782/782	70	71074	312/780
37	FA106	1040/1040	71	71891	645/880
38	FA107	600/163	72	72843	712/900
39	FA108	**WTC**/1000	73	73047	645/812
40	FA109	760/890	74	73066	785/785
41	FA110	890/890	75	73341	+/+
42	FA113	1045/1045	76	74809	680/745
43	FA114	765/1065	77	75641	+/+
44	FA115	940/112	78	75836	+/+
45	FA121	905/965	79	76333	+/+
46	FA123	700/1040	80	FRDA 6	167/500
47	FA131	900/1300	81	FRDA 11	720/920
48	FA132	10/330	82	FRDA 14	583/1183
49	FA134	930/930	83	FRDA 15	+/+
50	FA142	350/750	84	FRDA 18	1100/1134
51	FA150	500/800	85	FRDA 22	634/767
52	FA152	1070/1460	86	FRDA 23	167/834
53	FA153	400/1000	87	FRDA 26	100/1100
54	FA154	633/760	88	FRDA 27	412/850
55	FA156	740/1200	89	FRDA 28	380/780
56	FA163	700/1000	90	FRDA 33	834/1034
57	FA164	108/1040	91	FRDA 37	585/1250

58	FA165	765/1045	92	FRDA 39	785/785
59	FA167	1000/1000	93	FRDA 40	400/834
60	FA173	230/10	94	FRDA 41	100/500
61	FA174	1000/1000	95	FRDA 42	780/980
62	FA176	780/780	96	FRDA 43	334/900
63	FA178	249/559	97	FRDA 50	467/667
64	FA179	906/906	98	FRDA 54	650/850
65	FA181	536/809	99	FRDA 55	700/1000
66	FA188	10/1180*	100	FRDA 56	800/1000
67	FA191	150/573	101	FRDA 57	1100/1234
68	FA195	77/127	102	FRDA 58	200/1000
69	FA196	328/1194	103	FRDA 61	834/1200
70	FA197	478/1257	104	FRDA 74	800/867
71	SCA70	358/358	105	FRDA 76	1000/1200
72	SCA211	160/1040	106	FRDA 78	720/920
73	SCA321	696/800	107	FRDA 81	720/1020
74	SCA322	800/1013	108	FRDA 84	567/1000
75	SCA372	766/1046	109	FRDA 86	850/1150
76	SCA380	390/390	110	FRDA 87	850/850
77	SCA502	10/+	111	FRDA 88	685/1120
78	SCA596	10/+	112	FRDA 89	750/850
79	SCA597	765/1045	113	FRDA 92	380/520
80	SCA612	+/+	114	FRDA 93	720/885
81	SCA671	485/485	115	FRDA 94	900/1200
82	SCA694	800/800	116	FRDA 97	450/985
83	SCA743	483/905	117	FRDA 98	920/920
84	SCA814	+/+	118	FRDA 99	450/985
85	SCA922	765/1100	119	FRDA 100	850/1150
86	SCA937	700/1000	120	FRDA 101	1185/1185
87	SCA1013	400/400	121	FRDA 102	985/1120
88	SCA1120	+/10	122	FRDA 104	1050/1050
89	SCA1305	1010/1207	123	FRDA 105	10/867
90	SCA1306	1085/1165	124	FRDA 106	800/1134
91	SCA1311	1100/1400	125	FRDA 107	834/834
92	SCA1404	347/1301	126	FRDA 108	700/800
	UCL Samples		127	FRDA 109	834/1100
1	6336	200/200	128	FRDA 110	200/1100
2	9780	1020/1220	129	FRDA 111	No GAA
3	9940	350/1020	130	FRDA 112	734/900
4	10100	+/+	131	FRDA 113	720/720
5	10325	350/885	132	FRDA 114	400/667
6	10466	850/1050	133	FRDA 115	600/834
7	10722	400/1000	134	FRDA 116	634/1100
8	10905	683/983	135	FRDA 117	767/1134
9	11437	67/1100	136	FRDA 119	700/1000
10	11912	+/+	137	FRDA 122	785/850
11	12451	834/834	138	FRDA 123	700/1200
12	12941	600/767	139	FRDA 124	467/967
13	13037	520/850	140	FRDA 125	567/900
14	14805	850/1180	141	FRDA 126	767/867
15	15657	467/667	142	FRDA 127	600/1100
16	16852	967/1100	143	FRDA 128	434/600
17	17494	834/1167	144	FRDA 129	734/900
18	**17652**	**480/880**	145	FRDA 132	667/767
19	20886	1167/1167	146	FRDA 133	745/945
20	53297	30/612	147	FRDA 134	1080/1080
21	26162	400/534	148	FRDA 135	445/780
22	27643	667/1100	149	FRDA 137	780/880
23	27884	867/1134	150	FRDA 138	745/845
24	29897	150/850	151	FRDA 140	No GAA
25	30670	150/534	152	FRDA 141	645/780

26	34215	367/1100	153	FRDA 144	412/645
27	34655	1020/1220	154	FRDA 147	212/845
28	35594	567/834	155	FRDA 148	245/912
29	39232	1000/1000	156	FRDA 149	780/1180
30	40908	1100/1200	157	FRDA 150	45/745
31	41805	450/720	158	FRDA 151	650/980
32	42181	685/920	159	17786	+/+
33	43286	1067/1167	160	18204	+/+
34	44134	520/1050	161	65331	+/+

#: number; GAA1/2: GAA repeat sizes in allele 1 and allele 2; +: expanded allele of undetermined size but above 66 GAAs, which is considered as the pathogenic threshold; WTC: wild-type allele of undetermined size in a carrier (highlighted in bold); yellow highlight: samples were TP-PCR negative for the GAA repeat mutation. *: included in (Al-Mahdawi et al., 2018); red bold: sample failed for both *Mbo*II digestion and TP-PCR, and therefore removed from this study.

2.2.2 Huntington's Disease

Extracted lymphocyte DNA from 33 HD patients with 41 CAGs was provided by Prof Sarah Tabrizi (*Director of UCL Huntington's Disease Centre, Joint Head of Department of Neurodegenerative Disease, Dementia Research Institute, UCL Institute of Neurology and National Hospital for Neurology and Neurosurgery, Queen Square, London*) (Table 2.2). This cohort was chosen as although these patients have the same CAG repeat length, they displayed extreme phenotypic variability. The Neurogenetics Unit determined the size of the expanded allele by fragment analysis. All samples were sized with two sets of primers; HD3F/HDE and HD3F/HD5 (HD3F: 6-FAM-5'-CCTTCGAGTCCCTCAAGT-CCTT-3'; HDE: 5'-GGCGGTGGCGGCTGTTGCTGCTGCTGCTGC-3'; HD5: 5'-CGG-CTGAGGCAGCAGCGGCTGT-3'). The HDE primer covers the CAG repeat region. The HD5 primer includes the polymorphic CCG repeat region and is used to verify the HD3F/HDE results. CAG repeat length is diagnostically reported by the HD3F/HDE primer set. Age at onset was determined by the presence of the motor phenotype. Estimated mean age at onset relative to the CAG repeat length was determined using the model in Langbehn *et al*., 2004.

Six HD patient and six control *post-mortem* brains were obtained from The Queen Square Brain Bank (*UCL Queen Square Institute of Neurology, 1 Wakefield Street, London WC1N 1PJ*). DNA was extracted by LGC Genomics Ltd. (*Ostendstrasse 25, TGS Haus 8, 12459 Berlin, Germany*). The CAG repeat length was determined by fragment analysis of bulk DNA extracted from each brain region; frontal lobe, temporal lobe, occipital lobe, putamen, caudate nucleus, cerebellum, pons, and medulla (Table 2.3). Peripheral blood DNA was available for four out of the six HD *post-mortem* brains. Clinical information and details of disease onset and progression were obtained to aid in the understanding of the disease course in each patient (Table 2.4).

Table 2.2. HD patient blood samples with $(CAG)_{41}$

HD patient number	$(CAG)_n$	Age at onset		
		Actual	Estimated (Langbehn)	Residual
1	41	40.5	57.5	-17
2	41	42.5	57.5	-15
3	41	48.5	57.5	-9
4	41	50.5	57.5	-7
5	41	50.5	57.5	-7
6	41	51.5	57.5	-6
7	41	56.5	57.5	-1
8	41	56.5	57.5	-1
9	41	55.5	57.5	-2
10	41	58.5	57.5	1
11	41	55.5	57.5	-2
12	41	59.5	57.5	2
13	41	59.5	57.5	2
14	41	60.5	57.5	3
15	41	63	57.5	5.5
16	41	63.5	57.5	6
17	41	63.5	57.5	6
18	41	67.5	57.5	10
19	41	67.5	57.5	10
20	41	73.5	57.5	16
21	41	X	57.5	X
22	41	38	57.5	-19.5
23	41	P	57.5	X
24	41	P	57.5	X
25	41	58	57.5	0.5
26	41	X	57.5	X
27	41	61.5	57.5	4
28	41	54	57.5	-3.5
29	41	61	57.5	3.5
30	41	62	57.5	4.5
31	41	62	57.5	4.5
32	41	66	57.5	8.5
33	41	68	57.5	10.5

$(CAG)_n$: CAG repeat of length n (n = number); Actual: HD patient's actual age at onset; Estimated (Langbehn): estimated mean age at onset according to Langbehn *et al.*, 2004; Residual: actual age at onset minus estimated mean age at onset (Langbehn); X: no information available; P: premanifest.

Table 2.3. HD patient and control *post-mortem* brain fragment analysis

HD3F/ HDE	**FNT**	**TMP**	**OCC**	**PUT**	**CNU**	**CBM**	**PON**	**MED**	**Blood**
HD	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*
P40.97*	19-41	19-41	19-41	19-41	19-41	19-41	19-41	19-41	X
P2.03	18-43	18-43	18-43	18-43	18-43	18-43	18-43	18-43	21-44
P72.10	18-42	18-44	18-42	18-42	18-42	18-42	18-42	18-42	18-42
P3.92*	20-41	20-41	20-41	X	X	20-41	20-41	20-41	X
P7.96	18-41	18-41	18-41	18-41	18-41	18-41	18-41	18-41	18-42
P28.98	18-44	18-44	18-44	X	X	18-44	18-44	18-44	18-44
HD3F/ HD5	**FNT**	**TMP**	**OCC**	**PUT**	**CNU**	**CBM**	**PON**	**MED**	**Blood**
HD	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*	*WT-EA*
P40.97*	22-41	22-41	22-41	22-41	22-41	22-41	22-41	22-41	X
P2.03	21-43	21-43	21-43	21-43	21-43	21-43	21-43	21-43	21-43
P72.10	18-42	18-42	18-42	18-42	18-42	18-42	18-42	18-42	18-42
P3.92*	23-41	23-41	23-41	X	X	23-41	23-41	23-41	23-41
P7.96	18-41	18-41	18-41	18-41	18-41	18-41	18-41	18-41	18-41
P28.98	18-44	18-44	18-44	X	X	18-44	18-44	18-44	18-44
HD3F/ HDE	**FNT**	**TMP**	**OCC**	**PUT**	**CNU**	**CBM**	**PON**	**MED**	**Blood**
CTRL	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*
P1.11	18-19	18-19	18-19	18-19	18-19	18-19	18-19	18-19	18-19
P18.03	17-18	17-18	17-18	17-18	17-18	17-18	17-18	17-18	17-18
P32.09	15-18	15-18	15-18	15-18	15-18	15-18	15-18	15-18	15-18
P47.11	11-18	11-18	11-18	11-18	11-18	11-18	11-18	11-18	11-18
P82.10	18-22	18-22	18-22	18-22	18-22	18-22	18-22	18-22	18-22
P72.07	13-18	13-18	13-18	13-18	13-18	13-18	13-18	13-18	13-18
HD3F/ HD5	**FNT**	**TMP**	**OCC**	**PUT**	**CNU**	**CBM**	**PON**	**MED**	**Blood**
CTRL	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*	*WT-WT*
P1.11	18-19	18-19	18-19	18-19	18-19	18-19	18-19	18-19	18-19
P18.03	12-19	12-19	12-19	12-19	12-19	12-19	12-19	12-19	12-19
P32.09	15-18	15-18	15-18	15-18	15-18	15-18	15-18	15-18	15-18
P47.11	14-21	14-21	14-21	14-21	14-21	14-21	14-21	14-21	14-21
P82.10	18-22	18-22	18-22	18-22	18-22	18-22	18-22	18-22	18-22
P72.07	13-18	13-18	13-18	13-18	13-18	13-18	13-18	13-18	13-18

FNT: frontal lobe; TMP: temporal lobe; OCC: occipital lobe; PUT: putamen; CNU: caudate nucleus; CBM: cerebellum; PON: pons; MED: medulla; HD: Huntington's disease; CTRL: control; *WT*: wild-type allele CAG repeat size; *EA*: expanded allele CAG repeat size; *: peripheral blood DNA unavailable; X: sample unavailable.

Table 2.4. Additional information for HD patient and control *post-mortem* brains

HD	Gender	PMI	AAO	DD	EA	LAAO	Additional Clinical Information
P40.97*	Male	48	65	12	41	57.5	Epileptic fits, cognitive impairment, acute psychosis
P2.03	Male	74	47	18	43	47.5	Depression
P72.10	Male	37	55	15	42	52.5	Remained very active for 14 years after diagnosis at 55yrs until a haemorrhage due to serious head injury
P3.92*	Female	10	71	3	41	57.5	Hypomanic schizoaffective disorder. Parietal cystic lesion, malignant glioma
P7.96	Female	48	72	< 1	41	57.5	Senile onset chorea. Rapid progression
P28.98	Female	96	50	9	44	42.5	Paranoid delusions. Generalised cortical atrophy
CTRL	**Gender**	**PMI**	**AAD**	**WT**	**Additional Clinical information**		
P47.11	Female	89	79	21	Sjogren's syndrome, hypothyroidism		
P1.11	Female	89	93	19	Senile myocardial degeneration		
P18.03	Female	58	56	19	Polio, neurologically normal		
P32.09	Female	40	86	18	Colon cancer, multi-organ failure		
P82.10	Female	79	87	22	Metastatic colon carcinoma		
P72.07	Male	78	85	18	Multi-organ failure		

PMI: *post-mortem* index given in hours; AAO: age at onset in years; DD: disease duration given in years; EA: expanded allele CAG repeat size; LAAO: estimated mean age at onset in years according to Langbehn *et al.*, 2004; *: peripheral blood DNA unavailable; CTRL: control; AAD: age at death in years; WT: wild-type allele CAG repeat size.

2.3 *Mbo*II Digestion Analysis

Long-range PCR was performed with approximately 100 ng of DNA, using either Expand High Fidelity PCR System dNTPack (Roche), or Long-Range PCR Kit (Qiagen) together with GAA-B-F (5'-AATGGATTTCCTGGCAGGACGC-3') and GAA-B-R (5'-GCAT-TGGGCGATCTTGGCTTAA-3') primers as previously described (Holloway et al., 2011). The thermocycling conditions for each kit were; Roche Kit: 94°C for 2 min; 10 cycles of 94°C for 10 sec, 60°C for 30 sec, 68°C for 45 sec; 20 cycles of 94°C for 10 sec, 60°C for 30 sec, 68°C for 1 min with 20 sec increments; and a final cycle of 68°C for 10 mins; Qiagen Kit: 93°C for 3 min; 35 cycles of 93°C for 15 sec, 62°C for 30 sec, 68°C for 5 min, and a final cycle of 68°C for 10 min. The PCR products were run on 1% agarose gels and the FRDA positive samples were digested with *Mbo*II, which has a cleavage sequence of 5'-GAAGA(8/7)-3' (New England BioLabs). The PCR products were digested in a total reaction volume of 20 µL at 37°C for 1 hour. The digested DNA fragments were then heated at 95°C for 10 min followed by slow cooling to room temperature to prevent potential heteroduplex formation. Once cooled, the samples were separated by running on 2% agarose gels (1% Nusieve (Seakem Agarose GTG) and 1% Metaphor agarose (LONZA)). Pure GAA repeat sequences were fully cut leaving only two fragments from the uncut flanking sequences, 171/170 base pair upstream (referred to as 170 bp) and 117/118 base pair (referred to as 120 bp) downstream. If the GAA repeat contains an interruption, the sequences that were not cut by *Mbo*II leave either two bands with sizes that differ from the expected 170 bp and 120 bp bands and additional bands.

2.4 Triplet Primed PCR Analysis

The 3' end of the GAA repeat was assessed by triplet primed PCR (TP-PCR) with approximately 100 ng of genomic DNA as previously described (Ciotti et al., 2004). The TP-PCR master mix contained the following reagents: 10 µL Amplitaq Gold 360 Master Mix; 2 µL GC enhancer; 4 µL PCR grade H_2O; 1 µL primer FATP-P3 10 pmol/µL (5'-[6FAM]-TACGCATCCCAGTTTG-AGACG-3'); 1 µL primer mix FATP-P1 10 pmol/µL (5'-GCTGGGATTACAGGCGCGCGA-3') and FATP-P4 1 pmol/µL (5'-TAC-GCATCCCAGTTTGAGACGGAAGAAGAAGAAGAA-GAAGAA-3'). The following thermocycling conditions were used: 95°C for 10 min; 35 cycles of 95°C for 1 min, 58°C for 1 min, 72°C for 1 min; final extension of 72°C for 7 min. TP-PCR products were then analysed by capillary electrophoresis with 12 µL HiDi formamide and 0.03 µL GeneScan 500 LIZ® Size Standard per 1 µL TP-PCR product. The plate was heat-sealed, heated at

95°C for 3 min, followed by incubation on ice for 3 min before loading onto the ABI 3730xl DNA analyser. The resulting output was analysed using GeneMapper Software (version 5.0, Applied Biosystems). GeneScan 500 LIZ® Size Standard was used, which sizes DNA fragments in the 35 to 500 nucleotide range and provides 16 single-stranded labelled fragments of: 35, 50, 75, 100, 139, 150, 160, 200, 250, 300, 340, 400, 450, 490, and 500 nucleotides. Chromatographs were obtained and interrupted non-GAA sequences were identified by distinct gaps in the chromatographs. Seven potential reverse P1 primers were developed to create a TP-PCR assay for the 5' region of the GAA repeat (Table 2.5) in combination with FATP-P3(R), unaltered, and FATP-P4(R), which contains a reverse repeat sequence (5'-TACGCATCCCAGTTTGAGACGTTCTTCTT-CTTCTTCTTCTTC-3').

Table 2.5. Reverse TP-PCR P1 primers

Primer	Sequence 5'-3'
P1 (R0)	GACTAACCTGGCCAACATGGTG
P1 (R1)	GGAGTTCAAGACTAACCTGGCC
P1 (R2)	GCCAACATGGTGAAACCCAGTA
P1 (R3)	TGGTGAAACCCAGTATCTACTAAA
P1 (R4)	GTGAAACCCAGTATCTACTAAAAAATAC
P1 (R5)	GAAACCCAGTATCTACTAAAAAATACAAAAA
F Long P1	GGGATTGGTTGCCAGTGCTTAAAAGTTAG

P1: FATP-P1 primer; R0-5: reverse primer trial sequence 0 to 5.

2.5 Clone Sequencing of the CAG Repeat in HD Pathogenic Alleles

DNA was extracted from eight regions of six HD *post-mortem* brains, which were provided by the Queen Square Brain Bank, and DNA was extracted by LGC Genomics. HD alleles were amplified by slowdown PCR using Platinum SuperFi DNA polymerase (error rate of 10^{-8}/nucleotide incorporation) and primers; *HTT* 5163 Forward (5'-GCTGATGAAG-GCCTTCGAG-T-3') and *HTT* 5678 Reverse (5'-GAATTCAGGACA-GGCCCCAA-3'), which flank the CAG repeat and adjacent CCG repeats. The alleles were resolved on 3% (w/v) agarose gels, which were post-stained with a counterion-dye staining solution containing 0.0025% (w/v) crystal violet and 0.0005% (w/v) methyl orange in dH_2O (Yang et al., 2001). Individual expanded allele PCR products were gel purified according to the PureLink Quick Gel Extraction Kit (Life Technologies) and ligated into the pCR-Blunt vector (ThermoFisher Scientific). Ligations were transformed into chemically competent Stbl3 *E. coli* (Invitrogen) (genotype F− *mcr*B *mrr hsd*S20(r_B−,m_B−) *rec*A13 *sup*E44 *ara*-14 *gal*/K2 *lac*Y1 *pro*A2 *rps*L20(Str^R) *xyl*-5 λ−*leumtl*-1), and transformants containing plasmids with inserts were propagated. The plasmid DNA was isolated from the bacteria using the Monarch Plasmid Miniprep Kit (New England BioLabs). The isolated DNA was then sequenced by Source BioScience using the M13 primer and sequence analysis was performed on CodonCode Aligner.

2.6 Small Pool PCR Amplification

Bulk HD *post-mortem* brain DNA was serial diluted to 1 ng/µL, 250 pg/µL and 50 pg/µL. Repeat length variability was assessed by SP-PCR with the following primers, Hu_4 forward (5'-ATGGCGACCCTGGA-AAAGCTGATGAA-3') and Hu_3 reverse (5'-GG-CGGCTGAGGAAGCTGAGGA-3'). All dilutions of DNA were amplified in a 7 µL reaction containing 0.2 µM of each primer, 1X PCR buffer with added 2-mercaptoethanol, dimethyl sulfoxide, 0.2 U *Taq* DNA polymerase (Sigma) and nuclease-free H_2O. Amplification was performed with the following thermocycling conditions: 95°C for 2 min; with 28 subsequent cycles of denaturation at 95°C for 30 sec, annealing at 65°C for 30 sec and extension at 72°C for 3 min; 72°C for 10 min; hold at 10°C. The alleles were resolved on 1.5% (w/v) agarose gels in 0.5X TBE with 500 nM ethidium bromide. Before blotting, any excess gel was removed using a scalpel and the gel was flipped to orientate the DNA to the surface. This work was carried out in the lab of Prof Darren G Monckton (*Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow*).

2.7 Southern Blotting

The agarose gel was rinsed in dH_2O and transferred to a tray where it was washed with depurinating solution and left gently shaking for 10 min. Rinsing with dH_2O was repeated before adding the denaturing solution for 30 min and before adding the neutralising solution for 30 min. Hybridisation membrane was cut to the same size as the gel and equilibrated in the neutralisation solution before placing directly on top of the gel. Air bubbles were removed before adding three pieces of Whatman paper followed by paper towels and a glass plate to distribute approximately 500 g to 1000 g of weight and left for 3 hours to 16 hours to allow the transfer of the DNA from the gel onto the membrane by capillary action. The blot was then dismantled and the membrane reversed so that the DNA is on top. The membrane was dried at 80°C and the DNA fixed to the membrane by exposure to 1200 J/m^2 of UV light in a DNA crosslinker. The membrane was kept at room temperature until hybridisation. This work was carried out in the lab of Prof Darren G Monckton (*Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow*).

2.8 Hybridisation

To hybridise the membrane, it was wet in dH_2O, rolled and placed in a hybridisation bottle with the DNA side facing inwards. Hybridisation solution (5 mL) was added and the bottle incubated rotating at 65°C for 1 hour. During this time, the probe was made starting with 2 µL DNA ladder, 3 µL DM56 (DNA from DM1 patient), 18 µL dH_2O. A mixture of 6 µL (dATP-dGTP-dTTP), 15 µL random primer mix and 5 µL (α-phosphorus 32 [^{32}P]) dCTP was added to the probe and mixed. Subsequently, 1 µL of Klenow fragment (of DNA polymerase I) was added to the probe and mixed before incubation in a water bath at 25°C for 1 hour. The probe was then boiled for 5 min at 95°C and incubated on ice for 2 min before adding to the hybridisation bottle with the prehybridised membranes. Hybridisation was performed at 65°C overnight. Following hybridisation, the hybridisation solution was discarded in running water and the membrane rinsed (still in the bottle) with 15 mL of high-stringency washing solution (0.2% (w/v) SDS, 0.2X SSC) at room temperature to remove excess probe and free (α-^{32}P) dCTP. The membranes were then washed twice in 20 mL of the high-stringency washing solution for 20 min rotating at 65°C. Finally, the membrane was transferred to a large flat tray and rinsed gently shaking with the high-stringency washing solution for 30 min. The membranes were then transferred with the DNA side facing up to blotting paper and dried at 80°C for at least 2

hours. When dried, the membranes were directly exposed against X-ray film in an autoradiography cassette. The autoradiographs were developed after an exposure time of 4 hours to 3 days. This work was carried out in the lab of Prof Darren G Monckton (*Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow*).

2.9 Illumina NeuroChip Array Analysis of DNA Repair Pathway SNPs

The NeuroChip imputes over 5.3 million common SNPs from the latest release of the Haplotype Reference Consortium, which combines sequencing data from multiple neurodegenerative disease cohorts to create a large reference panel of human haplotypes (http://www.haplotype-reference-consortium.org/). The panel of SNPs used in this report were selected from the most significant genes (gene-wide $p < 0.1$) in the DNA repair pathway cluster from the GeM-HD consortium GWAS with SNPs from *RRM2B* and *UBR5* added due to their nominal significance and significant gene-wide p-values (GEM-HD, 2015) (Table 2.6). If the SNPs were not available on the NeuroChip array, proxy SNPs were identified in high linkage disequilibrium with the most significant SNP ($r^2 > 0.7$) using LDlinkR (Myers et al., 2020). To protect against the effects of population stratification, Hardy-Weinberg equilibrium was calculated using the Hardy-Weinberg package (Graffelman, 2020), and Bonferroni corrected. SNPs were excluded if they had a Hardy-Weinberg p value < 0.001. SNP genotyping was performed on DNA from our HD patient and control cohort using the Illumina NeuroChip array at UCL Genomics (*UCL Great Ormond Street Institute of Child Health*).

The statistical analysis was carried out in RStudio (version 1.2.5033) using a script designed by Dr. Michael Flower, which was adapted from that used in Bettencourt *et al*., 2016. Linear regressions were performed on the natural log of the age at onset and age at death against the pure CAG repeat length. The regression parameters used were determined from the HD cohort used in this report and from the HD cohort used in Bettencourt *et al*., 2016. Both parameters were used to construct a predicted age at onset value for each patient, based on their CAG repeat length, which was then subtracted from their actual age at onset to give a residual value (GEM-HD, 2015). The Bettencourt *et al*., 2016 parameters were not used for the HD patient age at death data. The association of each SNP with age at onset and age at death was examined by performing a linear regression of the residual values on the number of minor alleles in R (Team, 2013). The primary analysis used in this report tested whether there was an overall association with

age at onset across the selected 22 SNPs. This was achieved by combining the association p values for each SNP using the Browns method (Brown, 1975) and the harmonic mean method (Wilson, 2019). The primary analysis used one-sided p-values for association in the same direction as that observed in GEM-HD, 2015. To assess the overall directionality of the association, the significance of the one-sided p values was compared against the significance obtained from analysis using two-sided p values, which were Bonferroni corrected. A similar analysis was used to test whether there was an overall association with age at death across the selected 22 SNPs. For the secondary analysis, the association with individual SNPs with age at onset and age at death were examined and Bonferroni corrected. To display the combined effect of the selected SNPs on the residual age at onset, a polygenic age at onset score was derived. It was defined as the sum of the number of minor alleles at each locus weighted by their effect size in this report using the Bettencourt *et al*., 2016 parameters. The polygenic score was plotted against the residual age at onset values. A similar analysis was used to derive the polygenic age at death score which used the regression parameters determined for the HD cohort in this report. Finally, the polygenic age at onset score was plotted against the relative rate of somatic expansion determined by Illumina MiSeq. The relative rate of somatic expansion was calculated as the proportion of reads in the sample with CAG repeat lengths greater than the progenitor allele relative to the number of reads with the progenitor allele CAG repeat size.

Table 2.6. DNA repair pathway SNP panel and corresponding proxy SNPs

SNP ID	Chr:position (bp) (GRCh38)	Gene
rs1037699	8:102238702	*RRM2B*
rs1037700	8:102238547	*RRM2B*
rs114136100	15:30905773	*FAN1*
rs115109737	5:80806625	*MSH3*
rs12531179*	7:5989056	*PMS2*
rs1382539[x]	5:80656335	*MSH3*
rs146353869	15:30834198	*HERC2P10*
rs150393409	15:30910758	*FAN1*
rs16869352	8:102293805	*UBR5*
rs175080*	14:75047125	*MLH3*
rs1799977	3:37012077	*MLH1*
rs1800937	2:47798625	*MSH6*
rs1805323*	7:5987311	*PMS2*
rs20579*	19:48165573	*LIG1*
rs3512*	15:30942802	*FAN1*
rs3735721	8:102205467	*RRM2B*
rs4150407*	2:127292055	*ERCC3*
rs5742933*	2:189784590	*PMS1*
rs5893603	8:102238612	*RRM2B*
rs6151792*	5:80761142	*MSH3*
rs61752302	8:102298925	*UBR5*
rs71636247	5:80823157	*MSH3*
rs72734283	14:75028356	*MLH3*
Reference SNP ID	**Proxy SNP ID (r^2 value)**	**Proxy Gene**
rs12531179*	rs852151 (0.811)	*EIF2AK1*
rs175080*	rs175084 (1.0)	*MLH3*
rs1805323*	rs12534423 (1.0)	*PMS2*
rs20579*	rs3730872 (0.732)	*LIG1*
rs3512*	rs11293 (1.0)	*FAN1*
rs4150407*	rs1566822 (1.0)	*ERCC3*
rs5742933*	rs3791767 (1.0)	*ORMDL1*
rs6151792*	rs6151816 (0.806)	*MSH3*

Proxy SNPs were identified in high linkage disequilibrium (LD) with the most significant SNP (r^2 > 0.7) using LDlink (https://analysistools.nci.nih.gov/LDlink/) in the British population. Chr: chromosome; dark red: not available on the NeuroChip array and/or proxy SNP r^2 < 0.7; *: proxy SNPs required; [x]: rs1382539 was analysed instead of rs557874766 as the genotyping data of rs1382539 are of higher quality, however the two are in high LD and tag the same association signal (Moss *et al.*, 2017).

2.10 DNA Library Preparation and Illumina MiSeq Sequencing

In collaboration with Prof Darren G Monckton (*Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow*), the HD patient blood and *post-mortem* brain samples were sequenced by Illumina MiSeq. The library preparation and MiSeq protocol was developed and completed by Dr Marc Ciosi (Ciosi et al., 2018). Briefly, the polyglutamine (CAG) and polyproline (CCG) tracts within exon one of *HTT* were amplified from genomic DNA using MiSeq-compatible PCR primers. The primers were designed to be *HTT* locus-specific with the complete sequence of a barcoded Illumina adapter, which allows the PCR product to be sequenced directly (Figure 2.1). A fraction of the post-PCR product was subjected to a clean-up reaction using AMPure XP beads (Beckman Coulter), which also removed any primer dimers that were present. Quality control of the cleaned sample was performed on the Qubit fluorometer, the DNA Bioanalyser and by qPCR. The library was sequenced adhering to the Illumina guidelines for an amplicon MiSeq run with MiSeq Reagent Kit v3 (Illumina) using a cluster density of 1000k cluster/mm2 supplemented with 5% PhiX spike-in (PhiX Control v3, Illumina), which allowed increasing nucleotide diversity during the run and serves as a sequencing control. The prepared sequencing library and PhiX were denatured according to the Illumina protocol and loaded onto the MiSeq instrument with a run time of 65 hours. The sequencing output was analysed using the MiSeq Control Software version 2.5.05, and the MiSeq Reporter software version 2.5.1 was used for demultiplexing the reads. The MiSeq raw data output files containing HD patient blood and *post-mortem* brain samples were genotyped using ScaleHD (version 0.315), developed by Mr Alastair Maxwell, which is an automated HD genotyping bioinformatics pipeline (http://scalehd.readthedocs.io/en/latest/). The raw data output files of both the forward and reverse sequence reads in the FastQ format are required for the pipeline. ScaleHD provides quality control, sequence alignment and genotyping of input files. The process of ScaleHD includes trimming the sequencing adapters and performing the Burrows-Wheeler Aligner Memory algorithm (BWA-MEM) alignment against 4,000 typical *HTT* references. Any reads that map to multiple references were filtered out and automated genotyping follows. The data was checked manually if a sample satisfied any of the ScaleHD behaviour queries (Table 2.7).

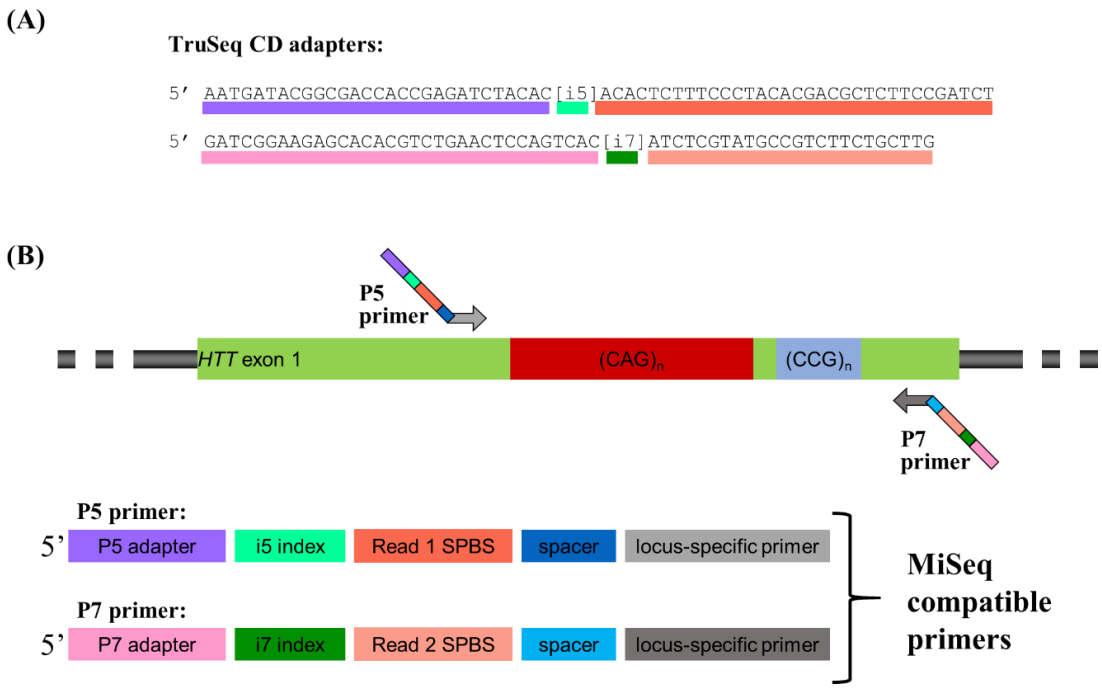


Figure 2.1. *HTT* locus-specific primers incorporating MiSeq adapters (gifted by Dr Marc Ciosi)

A: TruSeq combinatorial dual (CD) adapters with i5 and i7 indexes. Indexes are unique DNA sequences ligated to DNA library fragments with the downstream purpose of *in silico* sorting and identification. Illumina indexes can be pooled together, loaded into one lane of the sequencing flow cell, and sequenced in the same run. This allows the samples to be multiplexed. Individual reads are subsequently identified and sorted via bioinformatics. B: The designed *HTT* locus-specific primers incorporating MiSeq adapters. HS319F (5'-GCGACCCTGGAAAAGCTGATGA-3') and 33935.5 (5'-AGCAGCGGCTGTGCCTGC-3') are the two *HTT* locus specific primers, which respectively bind 26 bp 5' upstream of the CAG repeat and 26 bp 3' downstream of the CCG repeat. $(CAG)_n$: CAG repeat number; $(CCG)_n$: CCG repeat number; SPBS: sequencing primer binding site.

Table 2.7. ScaleHD behaviour queries

Behaviour Queries	
1	Check the FAIL samples. Most of them are probably due to a low number of mapped reads but there are some that will have failed for another reason and that can easily be genotyped manually like you've done on your Galaxy output. For the samples with a low number of mapped reads - the following rules regarding the number of mapped reads have been applied: a) < 100 reads mapped to the modal allele reference (n) = failed sample b) < 250 reads mapped to the modal allele reference = OK for genotyping but NOT OK for somatic mosaicism estimation
2	Check the homozygous haplotype samples (especially the ones with less than 80% forward mapped reads)
3	Check the samples with a confidence score < 55
4	Check the samples with $\geq CAG_{48}$
5	Check the samples for which the percentage of aligned forward reads is < 90%
6	Check samples with an exception raised
7	Check novel atypical sequences
8	Check the sample with SVM failure = TRUE
9	Check the samples with Differential Confusion = TRUE
10	Check the samples with homozygote CAG calls and homozygote CCG calls
11	Check the samples with atypical expanded calls

2.11 PacBio SMRT Sequencing

Five HD patient blood samples (HTT patient number 3, 5, 8, 13, and 24) and five HD *post-mortem* brain samples (P72.10 occipital lobe and medulla, and P3.92 temporal lobe, cerebellum and pons) were sent to PacBio (*Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025*) for amplification-free CRISPR-Cas9 targeted enrichment SMRT sequencing. All samples were processed with PacBio's CRISPR/Cas9 protocol for the RSII, except for HTT 8, P72.10 medulla, P3.92 temporal lobe and pons, which were processed with PacBio's CRISPR/Cas9 protocol for the Sequel System. Genomic DNA was fragmented with the high fidelity restriction enzymes, *Eco*RI and *Bam*HI. SMRTbell template libraries were prepared by ligation of capture adaptors carrying specific overhang sequences with *E.coli* DNA ligase. A genome complexity reduction step was added before library preparation to improve the observed on-target capture rate. In the presence of calf intestinal alkaline phosphatase, genomic DNA samples were pre-digested with the high fidelity restriction enzymes, *Kpn*I, *Mfe*I, *Spe*l, and *Eco*RV. Up to 1 µg of SMRTbell template library was used for the Cas9 nuclease digestion with up to four guide RNAs in the same digestion reaction. DNA samples were then ligated with a polyA hairpin adapter to obtain asymmetric SMRTbell templates from Cas9-digested target molecules. To enrich the asymmetric SMRTbell templates, PacBio MagBeads were used. The MagBead-DNA binding was carried out in high salt buffer for 2 hours at room temperature and washed once with low salt buffer (Figure 2.2). The bound DNA was then eluted in elution buffer at 65°C for 10 min. A standard PacBio sequencing primer lacking a polyA sequence was annealed to the eluted SMRTbell template and purified with AMPure beads for SMRT sequencing. A polymerase binding protocol with free hairpin adaptors in the binding buffer was used to bind the excess DNA polymerase. The sequencing data was collected on the PacBio RSII instrument using the following protocol; one-cell-per-well MagBead, P6/C4 sequencing chemistry, with a 4-hour collection time (Tsai et al., 2017). This protocol was optimised for the new Sequel System which added several steps including the use of many more restriction enzymes to remove background noise i.e. cutting up the unwanted DNA in order to increase the sequencing yield of the on-target DNA.

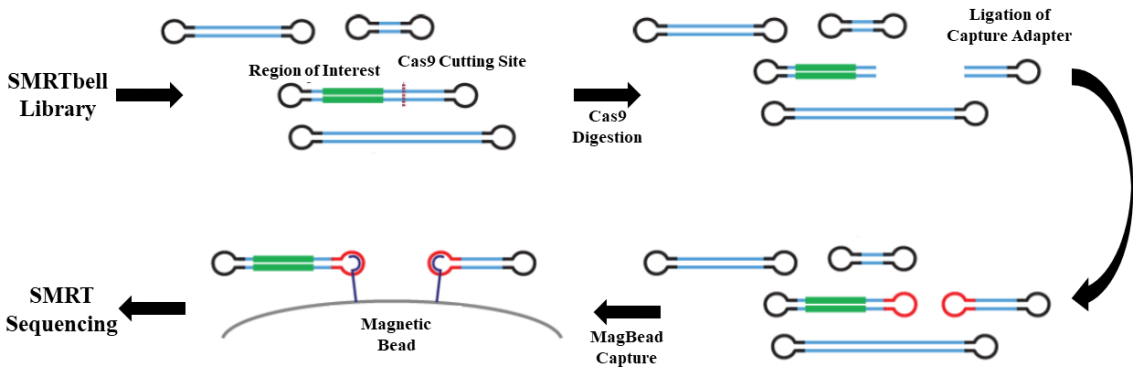


Figure 2.2. An overview of CRISPR-Cas9 targeted enrichment, adapted from (Tsai *et al*., 2017)

Native genomic DNA digested with *Eco*RI-HF and *Bam*HI-HF is used to generate the SMRTbell libraries. The SMRTbell templates containing the region of interest are cut open using Cas9 and a crRNA (CRISPR RNAs, transcribed from the CRISPR locus) designed to be complementary to a sequence adjacent to the region of interest. A capture adaptor (hairpin adaptor) is ligated to the digested templates and subsequently used as a handle in MagBead capture to enrich specifically for the region of interest. PacBio SMRT sequencing is used to sequence the captured SMRTbell templates.

2.12 Nanopore Sequencing

In total, 48 samples including control human *post-mortem* brain and HD patient blood and *post-mortem* brain samples were sent for Nanopore sequencing in collaboration with Dr Graham Taylor (*King's College London*) (Table 2.8). In combination with Nanopore long-read sequencing data, the estimation of CAG repeat length was determined in collaboration with Dr David Murphy (*Queen Square Genomics*), using an adjusted RepeatHMM as previously described (Liu et al., 2017). RepeatHMM is a computational tool that takes a set of reads and uses a split-and-align strategy to improve read alignments and perform error correction (Liu et al., 2017). This tool contains a hidden Markov model (HMM) and a peak calling algorithm, which is based on the Gaussian mixture model to infer repeat counts (Liu et al., 2017). For this report, the raw FASTA files received as the output of Nanopore sequencing from Dr Graham Taylor and team were converted to Sequencer Alignment/Map (SAM) files (https://github.com/lh3/minimap2). The SAM files were then converted to BAM files using Samtools, and Novosort sorted the BAM files to be further analysed. The BAM files were run through an adjusted RepeatHMM bioinformatics pipeline developed by Dr David Murphy and a bioinformatics script was used to align the CAG repeat reads to the human reference genome, Genome Reference Consortium Human Build 38 (GRCh38) (https://www.ncbi.nlm.nih.gov/assembly/GCA_000001405.28) (Script 2.1).

Table 2.8. Control and HD patient samples analysed by Nanopore sequencing

Sample number	Code	Tissue	Sample number	Code	Tissue
1	5318	Blood	25	59464	Blood
2	47875	Blood	26	58727	Blood
3	62710	Blood	27	68412	Blood
4	39032	Blood	28	P82.10 Frontal lobe	Brain*
5	14632	Blood	29	P82.10 Temporal lobe	Brain *
6	50310	Blood	30	P82.10 Occipital lobe	Brain *
7	55093	Blood	31	P82.10 Cerebellum	Brain *
8	65996	Blood	32	P82.10 Pons	Brain *
9	10037	Blood	33	P72.10 Frontal lobe	Brain
10	46584	Blood	34	P72.10 Temporal lobe	Brain
11	46024	Blood	35	P72.10 Occipital lobe	Brain
12	67935	Blood	36	P72.10 Putamen	Brain
13	52325	Blood	37	P72.10 Caudate nucleus	Brain
14	55777	Blood	38	P72.10 Cerebellum	Brain
15	30598	Blood	39	P72.10 Pons	Brain
16	47024	Blood	40	P72.10 Medulla	Brain
17	52251	Blood	41	P72.10 37025	Blood
18	48301	Blood	42	P3.92 Frontal lobe	Brain
19	60734	Blood	43	P3.92 Temporal lobe	Brain
20	72324	Blood	44	P3.92 Occipital lobe	Brain
21	57975	Blood	45	P3.92 Cerebellum	Brain
22	32038	Blood	46	P3.92 Pons	Brain
23	40254	Blood	47	P3.92 Medulla	Brain
24	73402	Blood	48	P3.92 6387	Brain

*: control human *post-mortem* brain.

Script 2.1. Alignment pipeline developed by David Murphy

```
#!/bin/bash

set -e

/hades/Software/NGS_Software/minimap2/minimap2 -ax map-ont /hades/dmurphy/pipeline/Homo _sapiens _assembly38.fasta  $1 > $1.sam samtools view -Sb $1.sam > $1.bam

rm $1.sam

/hades/Software/NGS_Software/novocraftV3.08.02/novosort --md --kt --ise -c 10 -t /hades/ pipelinetemp/ -f -i -o $1_sort.bam  $1.bam

rm $1.bam
```

Chapter 3. Investigating GAA Repeat Sequence Interruptions in Friedreich's Ataxia

3.1 Background

The FRDA GAA repeat is located on both alleles of the *frataxin* gene and the shorter of the two repeats is denoted as GAA1 and the subsequent as GAA2. The GAA repeats are sized by long-range PCR with primers that generate larger amplicons to alleviate selective amplification of GAA1 (Campuzano et al., 1996). However, PCR artefacts resembling large GAA expansions have been reported in FRDA negative individuals with two alleles of significantly different GAA repeat sizes, and similarly, in those with a single short repeat of < 200 GAAs (Poirier et al., 1999). In order to combat this, triplet repeat primed PCR (TP-PCR) was developed, which uses a fluorescently labelled locus specific primer flanking the repeat, with paired primers amplifying from multiple sites within the repeat and containing a common 5' tail (Warner et al., 1996). Yet, neither method informs on the sequence configuration. Pure GAA repeats have been determined to be stable up to a threshold of 44 GAAs, and surpassing this threshold results in instability (Sharma et al., 2004). An amplification and restriction enzyme based assay was developed to assess the purity of GAA repeats and the potential correlation of sequence configuration with repeat length (Holloway et al., 2011). This incorporated amplifying GAA repeats by PCR and restriction enzyme digestion using the endonuclease *Mbo*II. Previously, this method determined interrupted regions in one FRDA patient, in which part of the repeat was sequenced; $(GAA)_{21}(GGAGAA)_5(GGAGGAGAA)_{70}(GAA)_n$ (Holloway et al., 2011).

Several reports have identified GAA repeat interruptions, with suggestions that they protect against instability-promoting DNA secondary structures and therefore, have the potential to reduce the disease burden (Ohshima et al., 1999). Additionally, interruptions have been shown to increase the stability of the GAA repeat through intergenerational transmission; a $(GAAAGAA)_n$ interrupted $(GAA)_{112}$ repeat was reported to be stably transmitted through two generations (Cossée et al., 1997). To expand further upon this work, and to elucidate the commonality of interruptions in FRDA patients, we have investigated the sequence composition of the GAA repeat expansion in a large cohort of 253 FRDA patient and carrier DNA samples using TP-PCR analysis and long-range PCR amplification with *Mbo*II restriction enzyme digestion. Ultimately, understanding the role of sequence interruptions will provide a more accurate genotype-phenotype correlation for the improved genetic counselling of FRDA patients.

3.2 Results

3.2.1 TP-PCR Determines the Purity of the FRDA GAA Repeat Expansions

TP-PCR was performed on peripheral blood DNA obtained from 246 FRDA patient and seven carrier samples (N = 253) (Chapter 2, section 2.2.1 and *Supplementary Data Table 1*). The data were displayed in the form of chromatographs using GeneMapper software (version 5.0). If a sample was negative for the GAA expansion with < 44 GAAs, it was excluded. TP-PCR does not reveal the base pair configuration, therefore, sequence alterations were determined visually based on the resulting chromatograph profiles (Figure 3.1). Gaps in the TP-PCR chromatographs are representative of non-GAA sequences within one or both of the GAA repeats. Other patterns included; 'late starts', indicative of either sequence changes in the most 3' GAAs or insertions within the 3' flanking sequence; 'early starts', identifying deletions within the 3' flanking sequence, and 'double peaks', suggestive of one or two base pair insertions or deletions, which may be seen to resolve into single peaks again after further insertions or deletions.

Figure 3.1. TP-PCR chromatograph profiles (overleaf)

Each blue peak is representative of one GAA trinucleotide. Four different chromatograph profiles are represented; (A) depicts a sample starting at the expected 85 bp mark that has approximately 23 GAAs and is therefore FRDA negative ($GAA_n < 44$); (B) displays a typical GAA expanded profile of a positive FRDA patient starting at the expected 85 bp mark with an attenuating stretch of peaks ($GAA_n \geq 44$); (C) displays a sample with double peaks and the position of the first GAA is delayed by > 5 GAA's; (D) represents a sample that contains a gap (interruption) of 5 GAAs after the third GAA trinucleotide. Y axis: relative fluorescence units; X-axis: base pairs (bp); 80: 80 bp as normalised to the 500 LIZ size standard.

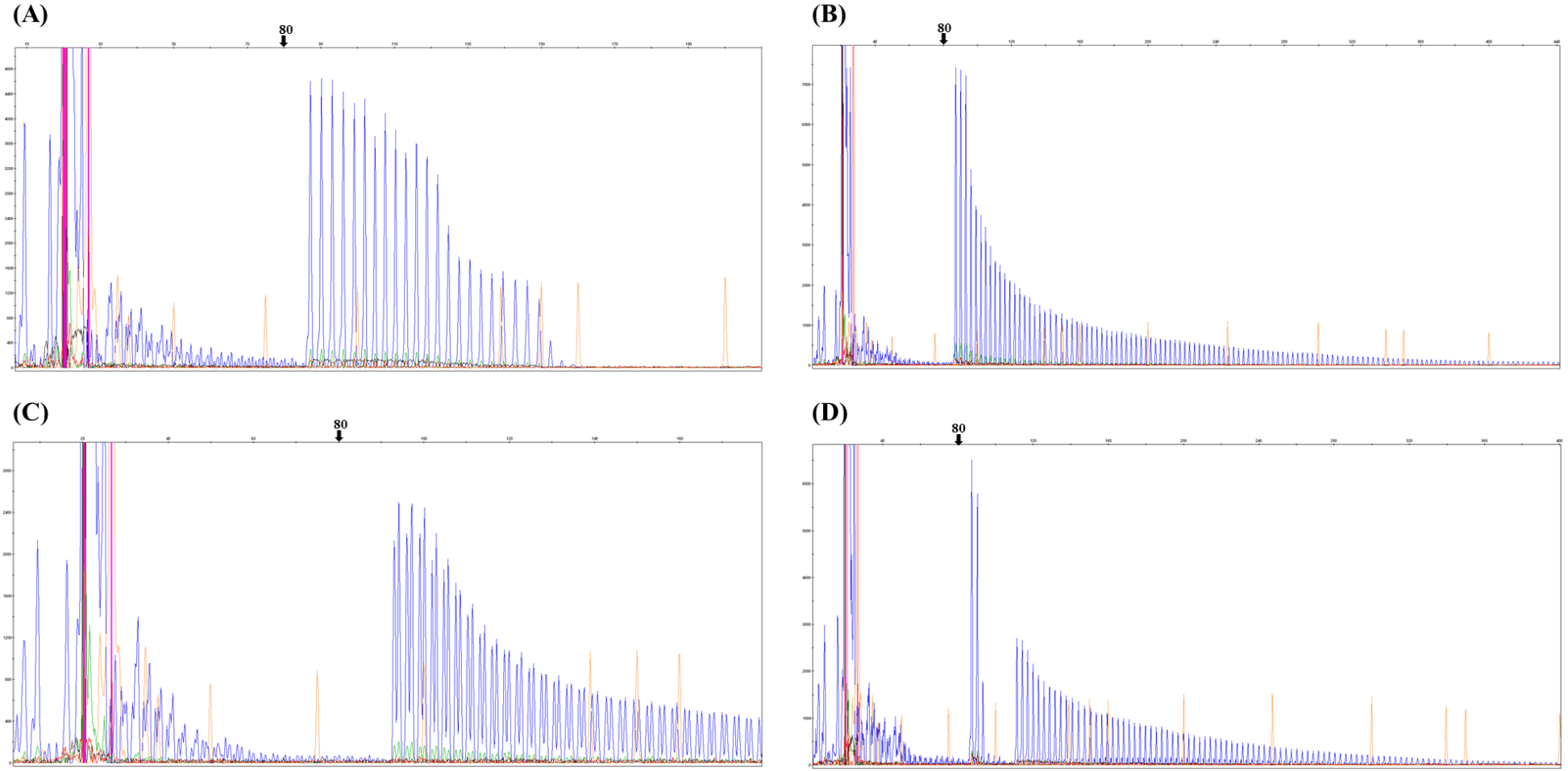


Figure 3.1. TP-PCR chromatograph profiles

TP-PCR of all 253 samples revealed that eight samples were negative for the GAA repeat expansion with < 44 GAAs, and two samples failed. These 10 samples were consequently excluded from the study. The remaining 236 FRDA patient and seven carrier samples displayed varying chromatograph profiles, indicating the presence of sequence interruptions in the 3' end of the GAA repeats. A full description of the chromatograph profiles for each sample is summarised in *Supplementary Data Table 1*. The TP-PCR data confirmed that 81 FRDA patients (34%) and four carriers had pure repeat sequences. In contrast, 155 FRDA patients (66%) and three carriers displayed sequence interruptions, which were located up to $(GAA)_{30}$ at the 3' end of the repeat (data not shown).

3.2.2 Reverse TP-PCR

As we have shown through TP-PCR that there are altered sequence profiles at the 3' end of the GAA repeat expansion in FRDA patients, a "reverse" TP-PCR protocol was developed to examine the 5' region of the repeat (Chapter 2, section 2.4). We hypothesized that in addition to the 3' GAA repeat interruptions, alterations at the 5' end of the expansion could act as phenotypic modifiers. Seven potential reverse P1 primers were developed to incorporate the 5' region of the GAA repeat (Chapter 2, Table 2.5). Promising results were obtained with the following combination of primers; F Long P1: 5'-GGGATTGGTTGCCAGTGCTTAAAAGTTAG-3', P3(R): [6FAM]–5'-T-ACGCA-TCCCAGTTTGAGACG-3' and P4(R): 5'-TACGATCCCAGTTTGAGACGTTCTTC-TTCTTCTTCTTCTTC-3', which were tested on a small selection of samples from our FRDA cohort (Figure 3.2). However, at the time of this study, the yield was too low and there was high background noise, which inhibited the application of this to the entire FRDA cohort. The low efficiency of this technique was considered to be due to a polyA stretch, which is located directly before the start of the GAA repeat.

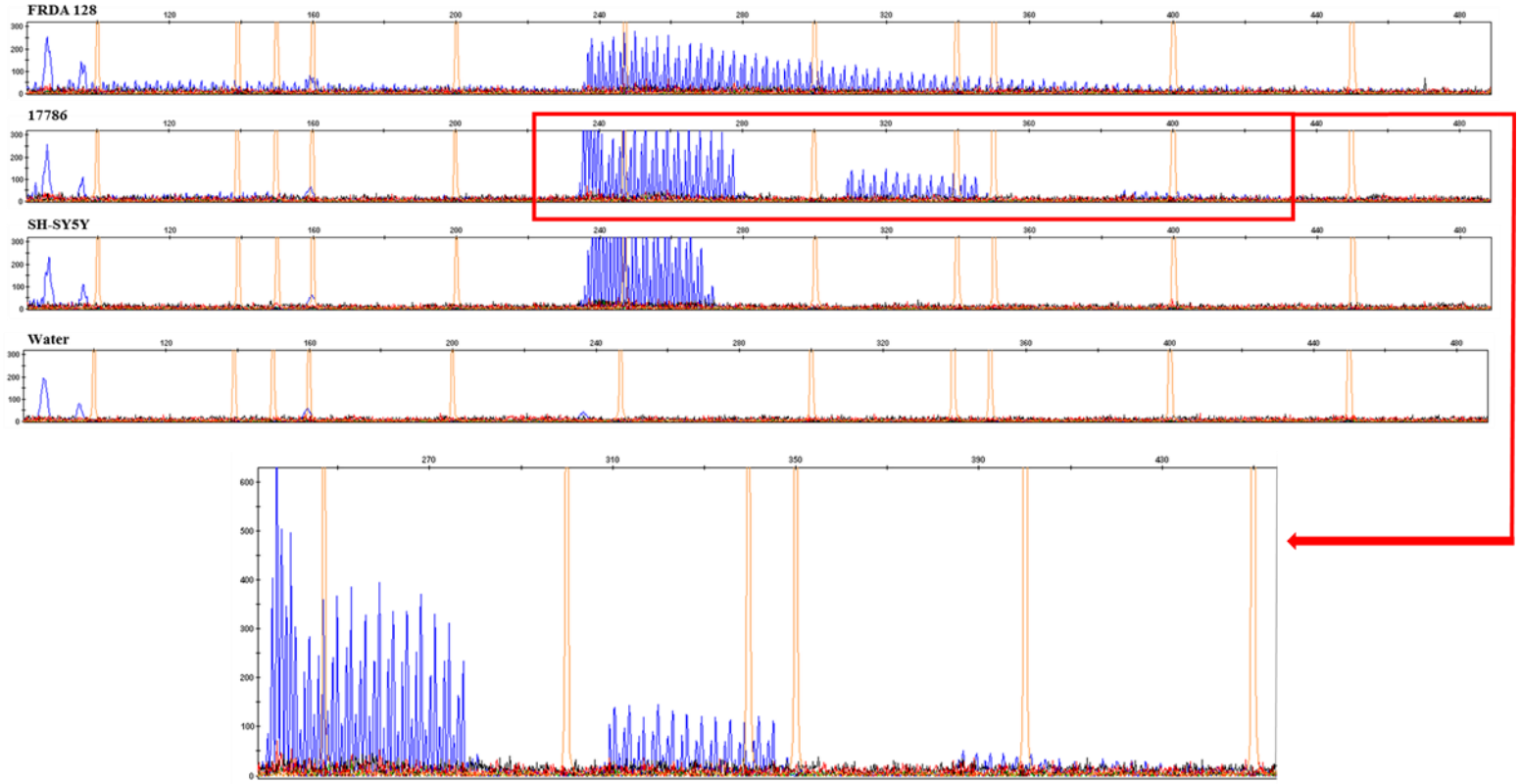


Figure 3.2. Reverse TP-PCR chromatographs

Chromatographs of alleles amplified by reverse TP-PCR, using an ABI 3730xl DNA sequencer and GeneMapper Software version 5.0. Each blue peak is representative of one GAA repeat. The numbers on top are the size in bp, relative to the 500 LIZ size standard. Red box highlights the enlarged area. The enlarged image shows that the GAA repeat continues along the chromatograph, however, the background noise is overshadowing it. This is more evident as the repeat expansion continues. Y axis: relative fluorescence units; X-axis: base pairs (bp); SH-SY5Y: DNA negative for the FRDA mutation, which was extracted from the human derived neuroblastoma SH-SY5Y cell line.

3.2.3 *Mbo*II Digestion Analysis Identifies Interrupted GAA Repeat Expansions

The DNA samples, at a concentration of 100 ng/µl, were amplified by long-range PCR with either the Expand High Fidelity PCR System, dNTPack (Roche), or the Long-Range PCR Kit (Qiagen) together with forward and reverse primers as previously described (Holloway et al., 2011). The PCR products, which contained the GAA repeat expansion with flanking sequences of 170 bp at the 5' end and 120 bp at the 3' end, were digested with *Mbo*II, which has a cleavage sequence of 5'-GAAGA(8/7)-3'. Out of the cohort of 236 confirmed FRDA patient and seven carrier samples, *Mbo*II digestion analysis was successful for 219 FRDA patient and seven carrier samples (N = 226). Pure GAA repeats were determined in 197 (87%) of the 226 samples, which presented as the two expected *Mbo*II bands at 120 bp and 170 bp. The remaining 13% of samples revealed alternative *Mbo*II band profiles (Figure 3.3). To investigate whether the results obtained from the *Mbo*II digestion profiles were replicable in the TP-PCR analysis, we compared samples with atypical *Mbo*II digestion profiles to their corresponding TP-PCR chromatographs (Figure 3.4). There was no relationship between the *Mbo*II digestion profiles and TP-PCR analysis of these samples.

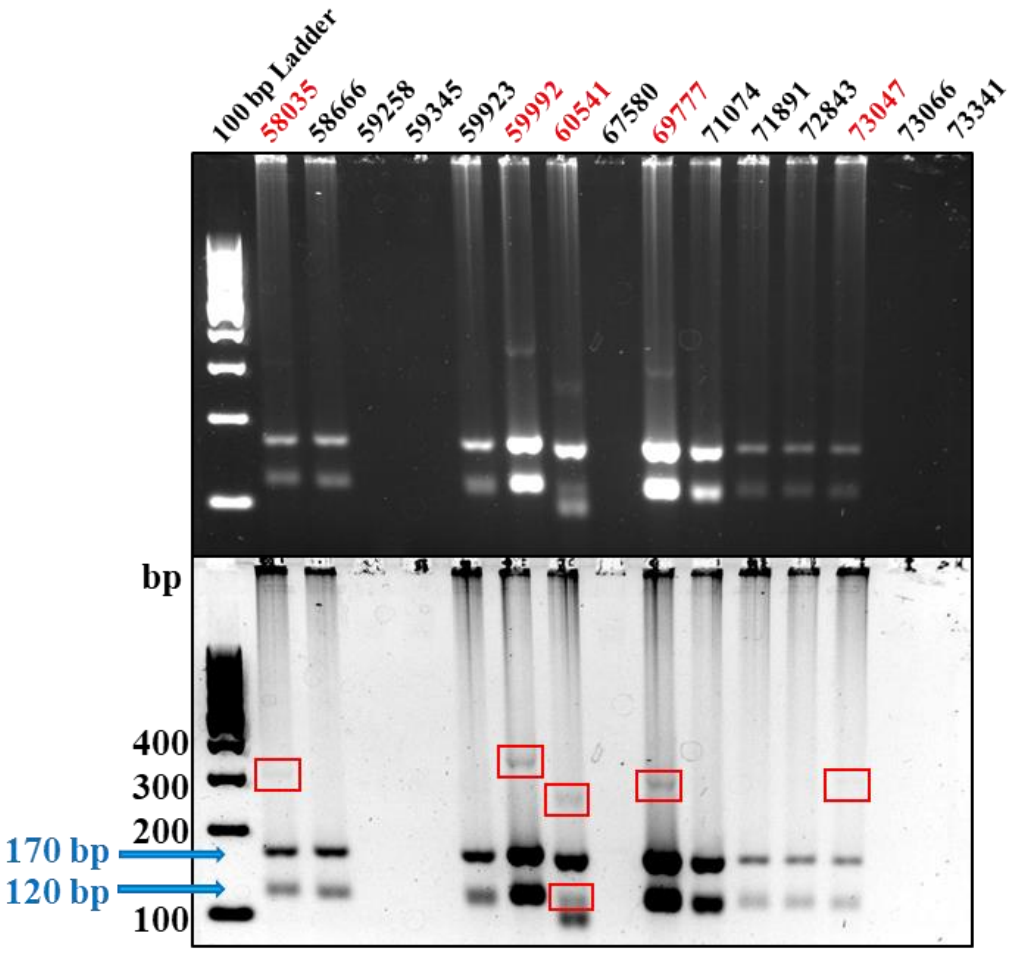


Figure 3.3. *Mbo*II digestion profiles

The bottom half of the image is the negative of the top half. The negative image allows an easier visualisation of atypical profiles. The blue 120 bp and 170 bp markers identify the two expected *Mbo*II digestion bands. Samples which show this two band profile are considered to have pure GAA repeats in relation to the sensitivity of the technique. An example of additional bands (red boxes) suggests the presence of GAA repeat interruptions within one or both of the GAA repeat expansions. In addition to the extra bands, sample 60541 displays a band profile of 100 bp and 170 bp, which indicates a deletion of approximately 20 bp in the 3' GAA flanking sequence.

Figure 3.4. Combined *Mbo*II and TP-PCR analysis for three of the FRDA samples (overleaf)

(A) The bottom half of the image is the negative of the top half. The negative image allows an easier visualisation of atypical profiles. The blue markers identify the two expected *Mbo*II digestion bands at 120 bp and 170 bp. Samples which show this two band profile are considered to have pure GAA repeats. An example of an additional band (red box) suggests the presence of GAA repeat interruptions within one or both of the GAA repeat expansions. (B) Chromatographs of alleles amplified by TP-PCR, using an ABI 3730xl DNA analyser and GeneMapper Software version 5.0. Each blue peak is representative of one GAA repeat. Sample 17786 has two additional bands at approximately 300 and 400 bp in the *Mbo*II digestion profile. This does not match the gap (interruption) present in the corresponding chromatograph, which represents approximately five GAA's (15 bp). Sample 18204 has a pure *Mbo*II digestion profile, however the chromatograph shows an interrupted sequence of approximately five GAA's (15 bp). Sample 65331, has an additional band at approximately 600 bp in the *Mbo*II digestion profile, which again, does not recapitulate the approximate nine GAA gap (27 bp) observed in the corresponding chromatograph.

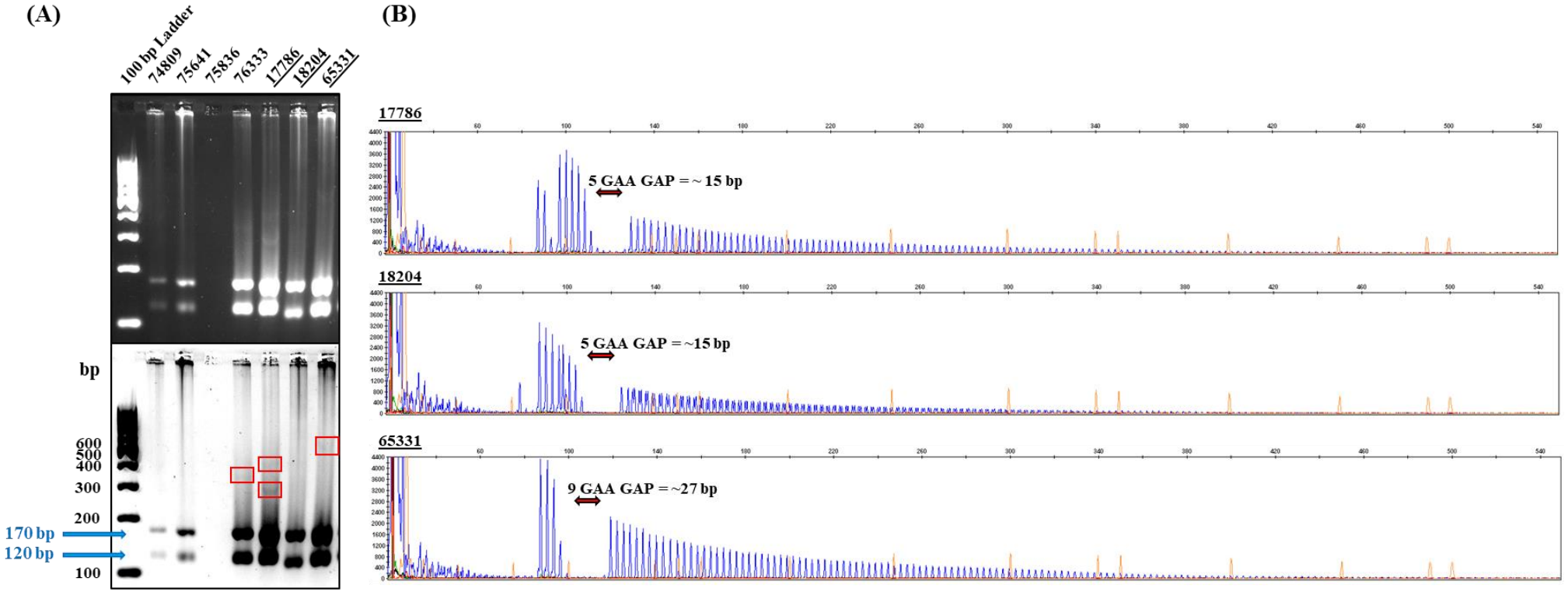


Figure 3.4. Combined *Mbo*II and TP-PCR analysis for three of the FRDA samples

3.2.4 GAA Repeat Size and Purity Influences FRDA Age at Onset

Age at onset data and GAA repeat lengths were available for 203 samples (*Supplementary Data Table 1*). Linear regression and correlation analysis of GAA1 and GAA2 repeat size versus age at onset were performed using GraphPad Prism (version 8.4.2). The analysis of all samples demonstrated the inverse relationship between GAA1 repeat size and age at onset ($R = -0.5814$ and $R^2 = 0.3380$). This highlights that approximately 33.8% of the age at onset variation is determined by the size of the GAA1 repeat. In contrast and as expected, the inverse relationship between the GAA2 repeat size and age at onset was much weaker, with approximately 7.8% of the age at onset variation determined by GAA2 ($R = -0.2806$ and $R^2 = 0.07873$) (Figure 3.5 A and B). Age at onset correlations based on the purity profile of the repeat determined by *Mbo*II digestion and TP-PCR was analysed using the GAA1 repeat size due to the significant inverse relationship with age at onset (Figure 3.5 C, D, E, and F). The correlation between GAA1 repeat size and age at onset for samples with pure *Mbo*II digestion profiles revealed a stronger inverse relationship when compared to the total sample collection ($R = -0.5902$ and $R^2 = 0.3483$). However, this was not observed for samples with pure TP-PCR profiles ($R = -0.2122$ and $R^2 = 0.04504$). The correlation between GAA1 repeat size and age at onset for samples with interrupted *Mbo*II digestion profiles identified a weaker inverse relationship when compared to the total sample collection and to the pure *Mbo*II digestion profiles ($R = -0.4095$ and $R^2 = 0.1679$). Interrupted TP-PCR profiles displayed a stronger correlation ($R = -0.5990$ and $R^2 = 0.3588$) when compared to the total sample collection and in comparison to the pure *Mbo*II digestion GAA1 repeat size correlations. With the exception of the pure TP-PCR profiles ($p = 0.0871$), there was a significant relationship between GAA1 and GAA2 repeat sizes with age at onset ($p = < 0.0001$), and between GAA1 repeat size with pure *Mbo*II digestion ($p = < 0.0001$), interrupted *Mbo*II digestion ($p = 0.0419$), and interrupted TP-PCR ($p = < 0.0001$) profiles.

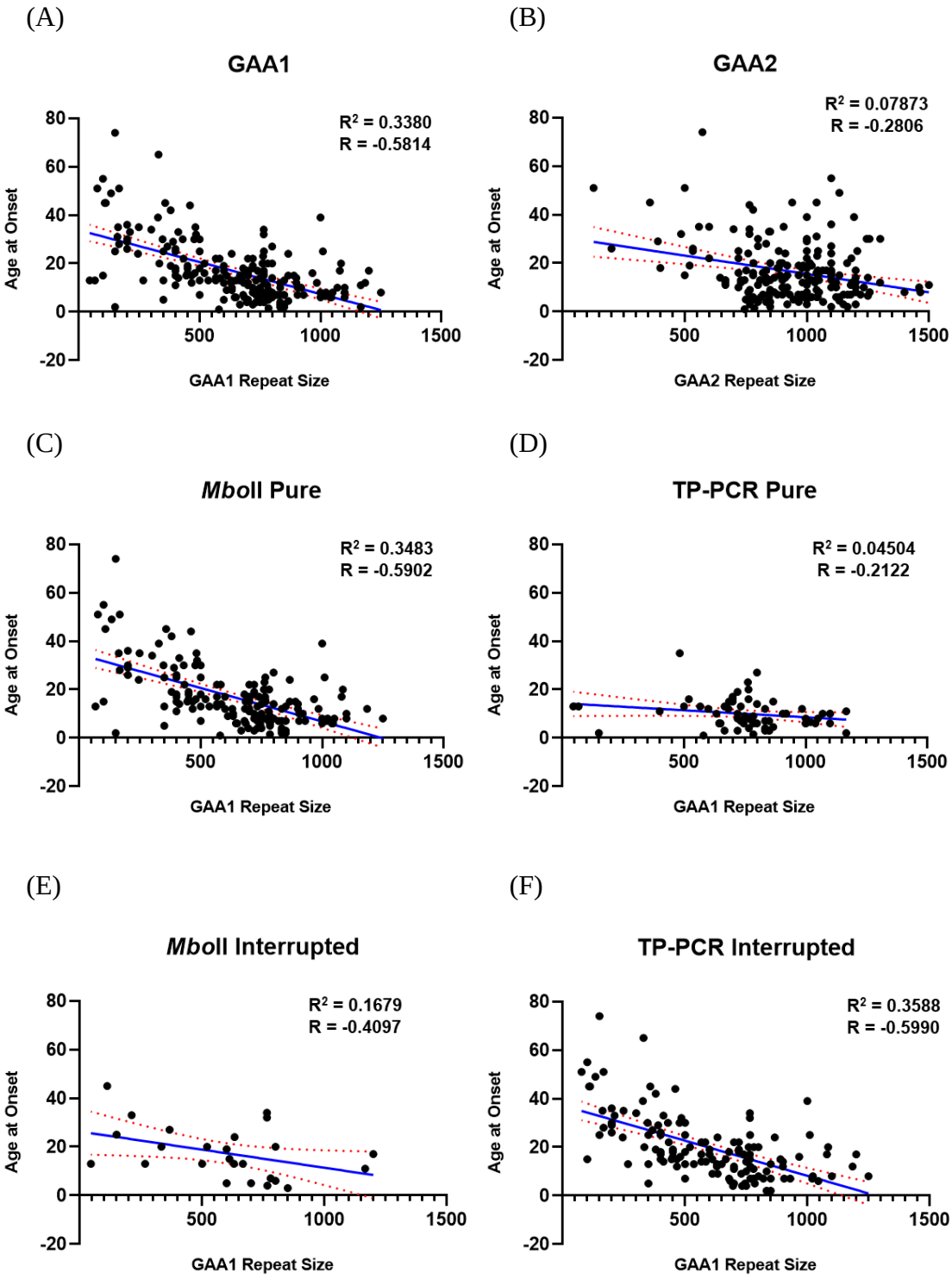


Figure 3.5. Linear regression analysis of GAA repeat size and sequence purity with age at onset

GAA1 repeat size is correlated to age at onset for all samples (A), GAA2 repeat size is correlated to age at onset for all samples (B), GAA1 repeat size is correlated to age at onset for pure *Mbo*II digestion (C) and TP-PCR (D) profiles, and for interrupted *Mbo*II digestion (E) and TP-PCR profiles (F). A significant relationship with age at onset was determined for all purity profiles ($p = \leq 0.0419$), with the exception of TP-PCR pure ($p = 0.0871$).

3.3 Discussion

3.3.1 TP-PCR and *Mbo*II Digestion Determine the GAA Repeat Purity

In relation to expanding on the work by Holloway *et al.*, 2011, who reported GAA interruptions in one of four FRDA patients, our study with a total of 236 FRDA patient and seven carrier samples, to our knowledge, represents the largest cohort of FRDA patients in which the frequency of GAA repeat interruptions has been investigated. We used two methods to identify sequence alterations; TP-PCR analysis of the 3' end of the expansion, which encompasses approximately 100 GAAs, and long-range PCR with *Mbo*II restriction enzyme digestion, which is sensitive to non-$(GAAGA)_n$ interruptions that are approximately ≥ 50 bp. The results of this study revealed that interruptions determined by TP-PCR, and located in the first 30 GAAs at the 3' end of the repeat, were common and that internal repeat interruptions determined by *Mbo*II were rare.

Interruptions were identified within the first 30 GAAs in 66% of our FRDA patient cohort, which is in agreement with previous literature (Sakamoto et al., 2001). DNA sequencing of approximately 200 GAA repeats (the length at which sequencing was technically possible) in 11 expanded FRDA alleles revealed that 45.5% carried interruptions within the last 10 to 15 GAAs at the 3' end of the expansion (Sakamoto et al., 2001). Therefore, our results suggest that with a larger FRDA patient cohort, the location of 3' GAA repeat interruptions can vary, with positions identified further along the expansion than previously reported. However, the larger FRDA cohort size did not drastically increase the frequency percentage of the 3' sequence interruptions. Previous reports have determined the commonality of these small sequence interruptions, however, Sakamoto et al., 2001 could not conclude on the relevance of 3' interruptions, but suggested that sequence variants could play a role in rare patients with atypical FRDA phenotypes (Al-Mahdawi et al., 2018; Sakamoto et al., 2001). Therefore, this highlights the need to develop a method which can give a greater consensus of the entire repeat composition that will subsequently enable a more comprehensive analysis into the relationship between GAA repeat interruptions and FRDA disease progression.

Similarly to TP-PCR, the *Mbo*II digestion technique does not identify the sequence at the base pair level, it instead identifies interruptions by atypical restriction enzyme digestion profiles. The PCR and *Mbo*II digestion-based assay from this exploratory study indicates that 87% of the FRDA samples in this cohort carry primarily pure GAA repeat expansions throughout the length of the repeat. The remaining 13% of FRDA patients had atypical

*Mbo*II-digestion profiles, suggestive of internal GAA repeat interruptions. Subsequent to identifying the purity status of the GAA repeats by *Mbo*II digestion and TP-PCR, the potential effect of repeat interruption profiles upon GAA1 size and age at onset was examined. In agreement with previous reports, our analysis of GAA1 size and age at onset indicated that up to 33.8% of the variation in age at onset in our FRDA cohort is determined by the GAA1 repeat length (Filla et al., 1996; Reetz et al., 2015). This study further identified that in addition to GAA1 size, the purity of the expansion is also an important factor when considering age at onset. Stronger correlations compared to the total sample collection were observed for the pure *Mbo*II digestion and interrupted TP-PCR subset of FRDA patients. Approximately 34.8% and 35.9% of age at onset variation in the pure *Mbo*II and interrupted TP-PCR subset was determined by GAA1 repeat size, respectively. Without the base pair sequence composition it is difficult to understand the exact influence of GAA repeat interruptions on FRDA phenotype. Although somewhat contradictory, these results highlight that sequence purity has the potential to influence age at onset and that further investigation is needed to resolve the GAA repeat sequence.

In contrast to TP-PCR, the *Mbo*II digestion method includes the entire expansion and is not limited to the last approximate 100 GAA repeats at the 3' end of the repeat. Nonetheless, it is important to note the limitations of the *Mbo*II restriction enzyme digestion method. This technique does not identify the sequence at the base pair level and it will only detect approximately 20 bp added to either of the flanking regions, or > 50 bp of internal interruptions within the GAA repeat expansion. *Mbo*II digestion analysis will not detect smaller interruptions less than 50 bp or $(GAAGA)_n$ interruptions as such sequences will be cut by the enzyme. Overall, deciphering the base by base sequence configuration of the GAA repeat expansion is needed to enable further in-depth conclusions regarding the disease modifying effects of interruptions in FRDA. The evolution of TGS and its contribution to determining the sequence of complex genetic regions, suggests that these techniques will soon be applicable to FRDA. This does not however, diminish the value of the information gained through the previous techniques of PCR amplification and/or restriction enzyme digestion and partial sequencing of the GAA repeat as they have revealed alterations within the expansion. This develops the curiosity and creates the need to translate the current TGS techniques to FRDA, which will ultimately determine the exact sequence configuration. Once this is determined, it has the potential to improve the genetic counselling of patients and increase our understanding of the genotype-phenotype relationship.

Chapter 4. Determining the *HTT* Sequence Configuration in Huntington's Disease Patients

4.1 Introduction

In relation to the phenotypic variance observed within HD patients with similarly sized pathogenic alleles, this report set out to determine the sequence of the CAG repeat and flanking CCG repeat regions in a cohort of HD patients with $(CAG)_{41}$ (Table 4.1). These patients presented with a 35.5-year age gap from the earliest to latest age at onset, which recapitulates the phenotypic variation reported for this CAG repeat length (Wexler et al., 2004). One potential source of individual phenotypic variation could be due to internal CAG repeat sequence alterations, such as nonsynonymous repeat interruptions, which have been previously identified as disease modifiers in other CAG repeat diseases (Fratta et al., 2014; Menon et al., 2013).

Table 4.1. HD patient blood samples with $(CAG)_{41}$

HD patient number	$(CAG)_n$	Age at onset		
		Actual	Estimated (Langbehn)	Residual
1	41	40.5	57.5	-17
2	41	42.5	57.5	-15
3	41	48.5	57.5	-9
4	41	50.5	57.5	-7
5	41	50.5	57.5	-7
6	41	51.5	57.5	-6
7	41	56.5	57.5	-1
8	41	56.5	57.5	-1
9	41	55.5	57.5	-2
10	41	58.5	57.5	1
11	41	55.5	57.5	-2
12	41	59.5	57.5	2
13	41	59.5	57.5	2
14	41	60.5	57.5	3
15	41	63	57.5	5.5
16	41	63.5	57.5	6
17	41	63.5	57.5	6
18	41	67.5	57.5	10
19	41	67.5	57.5	10
20	41	73.5	57.5	16
21	41	X	57.5	X
22	41	38	57.5	-19.5
23	41	P	57.5	X
24	41	P	57.5	X
25	41	58	57.5	0.5
26	41	X	57.5	X
27	41	61.5	57.5	4
28	41	54	57.5	-3.5
29	41	61	57.5	3.5
30	41	62	57.5	4.5
31	41	62	57.5	4.5
32	41	66	57.5	8.5
33	41	68	57.5	10.5

$(CAG)_n$: CAG repeat of length n (n = number); (Langbehn): estimated mean age at onset according to Langbehn *et al*., 2004; Residual age at onset: age at onset minus estimated mean age at onset (Langbehn); X: no information available; P: premanifest.

Using an optimised clone sequencing approach, we wanted to identify any sequence alterations that differed from the canonical sequence of the CAG repeat and flanking CCG repeat regions (Figure 4.1). During this work, Illumina developed the MiSeq System and an amplicon-sequencing protocol for the *HTT* exon 1 trinucleotide repeat was established (Ciosi et al., 2018). In collaboration with Prof Darren G Monckton, our cohort of HD patients with $(CAG)_{41}$ were sequenced by Illumina MiSeq, allowing a high-throughput quantification of the number of CAGs and CCGs, as well as the presence of sequence alterations. Third-generation sequencing platforms were subsequently developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). To complement the clone sequencing and Illumina MiSeq work, HD patient DNA was sent to PacBio and ONT for third-generation sequencing. Applying first-, second- and third-generation sequencing technologies to our cohort of HD patients additionally allowed us to investigate their efficiency in sizing the CAG repeat as well as elucidating the base pair configuration of the CAG repeat and flanking CCG repeat regions.

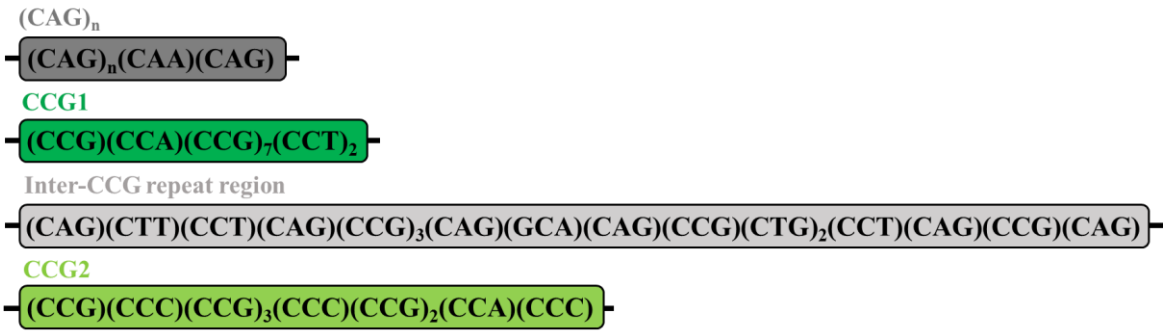


Figure 4.1. Canonical sequence of the CAG and CCG repeat regions of interest

The *HTT* sequence was obtained from the National Centre for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/nuccore/NC_000004.12?report=fasta&from=3074681&to=3243960). $(CAG)_n$: CAG repeat of length n, including the CAA trinucleotide if present; CCG1: CCG repeat region 1; Inter-CCG repeat region: sequence between CCG1 and CCG2; CCG2: CCG repeat region 2.

4.2 Results

4.2.1 Clone Sequencing of *HTT* in the $(CAG)_{41}$ HD Patient Cohort

Clone sequencing was performed on the DNA extracted from the blood of 33 HD patients with $(CAG)_{41}$. In order to investigate possible CAG repeat and flanking CCG repeat sequence alterations in the pathogenic allele only, the two alleles were separated by agarose gel electrophoresis and the expanded allele was excised and purified for subsequent clone sequencing. Clone sequencing was successful for 24 HD patients (Table 4.2) who presented with a 33-year age gap from the earliest to latest age at onset.

Table 4.2. Summary of clones obtained per HD patient

HD patient number	$(CAG)_n$		Number of clones	
	Wild type	Expanded	Wild type	Expanded
1	19	41		1
2	22	41	1	6
3	18	41		13
4	21	41		5
5	22	41	1	9
6	19	41		5
7	26	41		5
8	20	41	1	6
9	16	41		6
10	21	41	2	9
11	18	41		6
13	18	41		7
14	18	41	1	4
15	23	41	3	8
17	18	41	4	16
18	19	41	3	5
20	19	41		7
23	21	41		6
24	18	41		9
25	18	41		6
28	15	41		10
29	16	41		5
30	18	41	1	5
31	18	41	4	4

$(CAG)_n$: CAG repeat tract of length n as determined by fragment analysis.

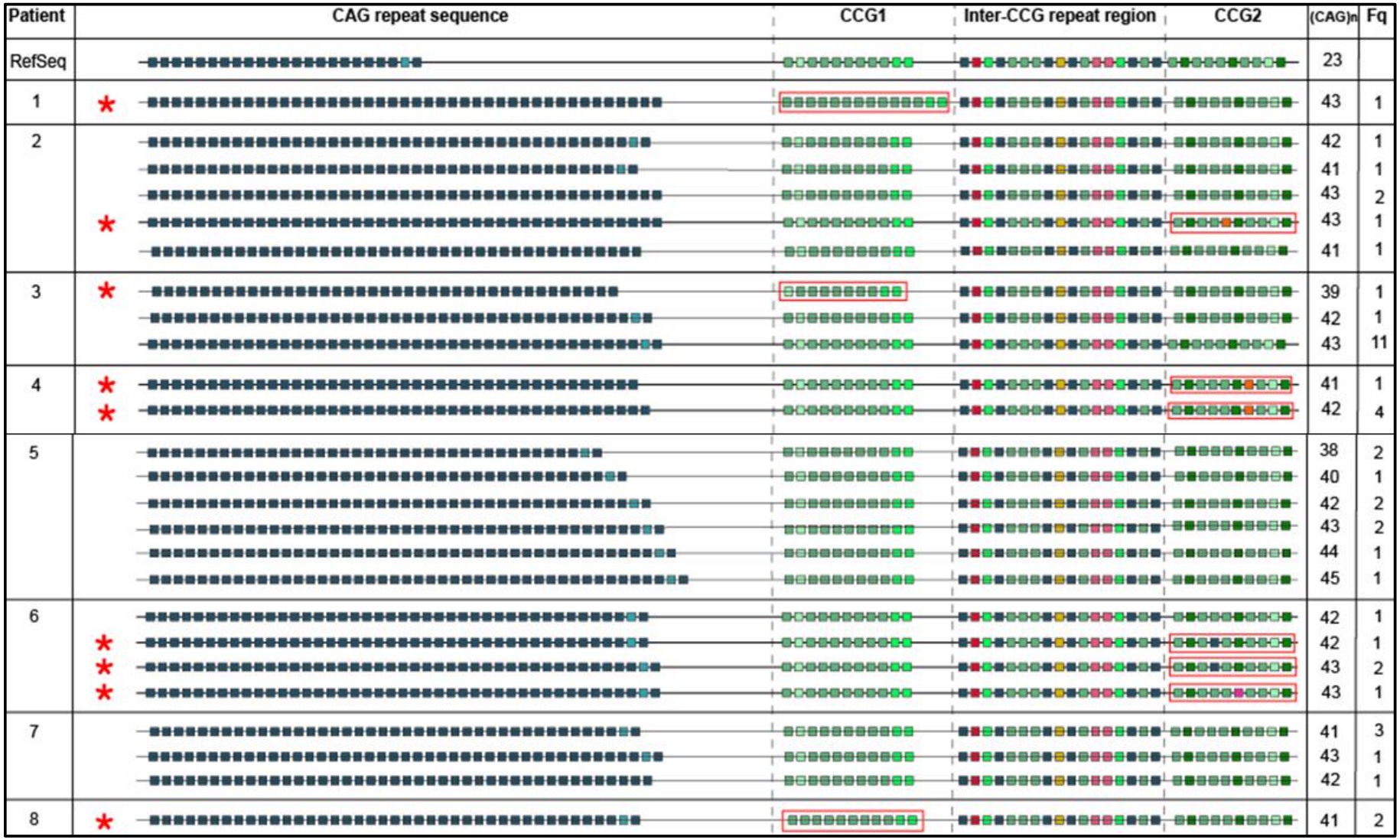
A total of 163 clones containing the *HTT* sequence were obtained, with all of the sequence configurations depicted in Figure 4.2. The specific sequence configurations resolved per patient can be seen in Figure 4.3, and a summary of the text sequence data is detailed in *Supplementary Data Table 2*. Only clones with a minimum of 36 CAGs were selected as expanded alleles. This was to incorporate alleles in the reduced penetrance range where HD pathogenesis has previously been reported to arise (Rubinsztein et al., 1996). No clones obtained from this cohort of HD patients contained any nonsynonymous sequence interruptions in the CAG repeat region. The penultimate synonymous CAA interruption was present in the majority of HD patients with the exception of all clones from patients 1 and 4. An additional CAA interruption was found in one clone each from HD patients 17, 20, 30 and 31, located within the CAG tract after the first 25, 37, 32, and 16 CAG trinucleotides, respectively. HD patients 2, 3, 7, and 28 contain a mixed population of clones with and without the penultimate CAA interruption. HD patient 1 contains $(CCG)_{12}$, which is the largest CCG1 tract in this cohort. The shortest was identified as $(CCG)_6$ in HD patients 9, 13, 17, 23, 24, 25, and 28. An A:G substitution resulting in the absence of the CCA trinucleotide prior to the CCG1 tract is present in HD patients 1 and 8, and a G:A substitution is present in HD patients 9 and 17, resulting in an additional CCA trinucleotide either before or interrupting the CCG tract, respectively. Another alteration where the first CCG trinucleotide is lost, $(CCA)(CCG)_7(CCT)_2$, was identified in one clone from HD patient 3. A G:A substitution was identified in the inter-CCG repeat region of one clone in HD patient 23. The most notable sequence alterations identified within the CCG2 repeat region were substitutions of C:A and C:T. This resulted in the following codons; ACG, CAG, CTC, CTG and TCG, which can be seen to various degrees in HD patients 2, 4, 6, 9, 11, 13, 17, 23 and 25.

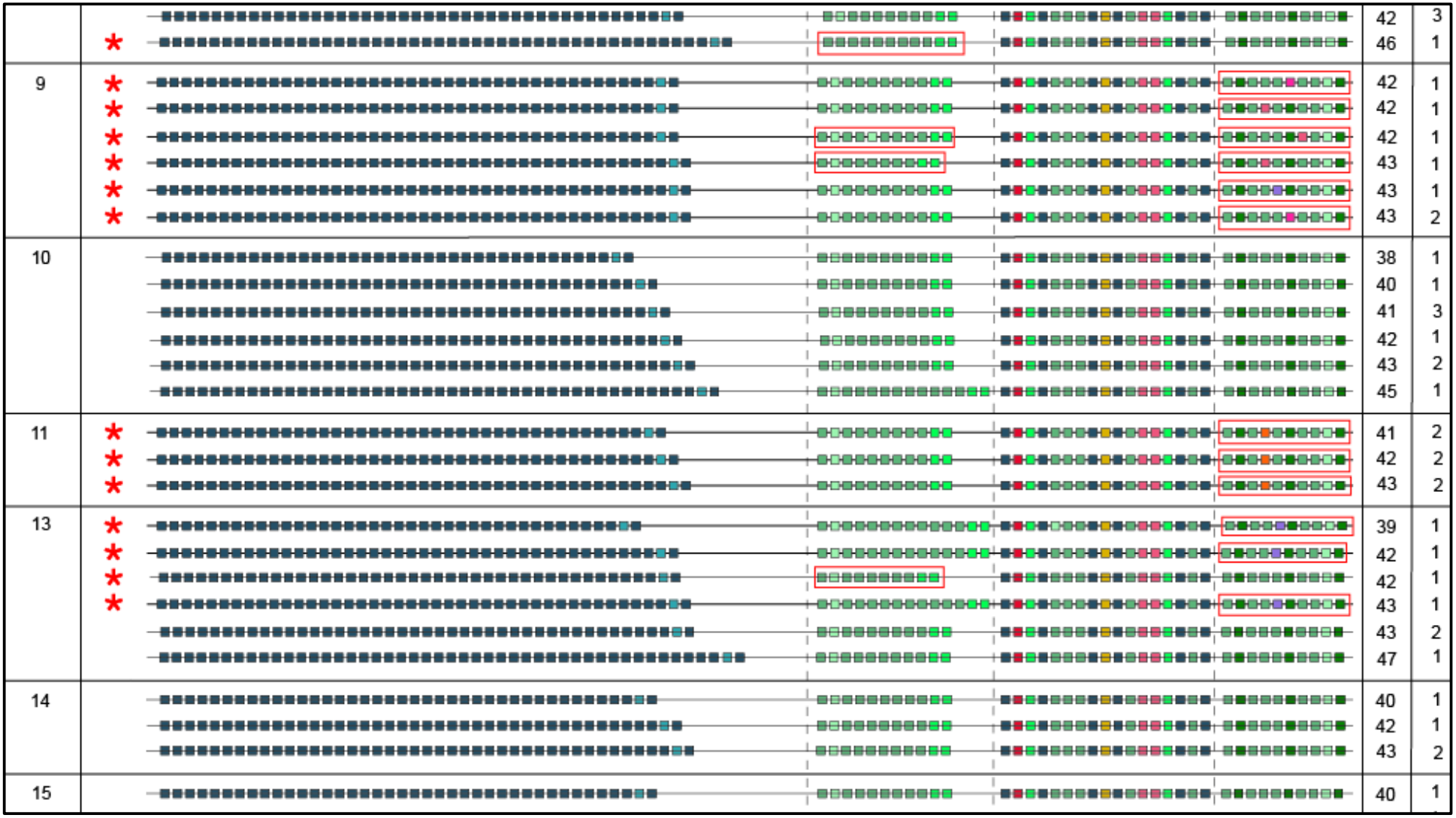
Figure 4.2. Sequencing figure legend

This figure displays all the sequences obtained through clone sequencing, which have been segmented into the various regions of interest. Each codon is colour coded with the single letter amino acid abbreviation above. Q: glutamine; P: proline; L: leucine; A: alanine; T: threonine; S: serine; RefSeq: reference sequence; CCG1: CCG repeat region 1; CCG2: CCG repeat region 2; Inter-CCG repeat region: region in between CCG1 and CCG2.

Figure 4.3. DNA sequences from successfully cloned HD patient blood samples (overleaf)

Each square represents a codon and is colour-coded for visual ease. The frequency (Fq) represents the population of clones obtained per patient and quantifies the presence of each sequence. RefSeq: *HTT* reference sequence according to NCBI (https://www.ncbi.nlm.nih.gov/nuccore/NG_009378.1?from=5001&to=174286&report=fasta); CAG repeat sequence: codon composition including CAA trinucleotides; CCG1: codon composition of CCG repeat region 1; CCG2: codon composition of CCG repeat region 2; Inter-CCG repeat region: codon composition of the region in-between CCG1 and CCG2; $(CAG)_n$: CAG repeat tract of size n including CAA if present; *: HD patients with previously unreported sequence alterations; red rectangle: location of previously unreported sequence alterations.


Patient
CAG repeat sequence
CCG1
Inter-CCG repeat region
CCG2
(CAG)n
Fq
RefSeq
23
1
43
1
2
42
1
41
1
43
2
43
1
41
1
3
39
1
42
1
43
11
4
41
1
42
4
5
38
2
40
1
42
2
43
2
44
1
45
1
6
42
1
42
1
43
2
43
1
7
41
3
43
1
42
1
8
41
2


42 3
* 46 1
9
* 42 1
* 42 1
* 42 1
* 43 1
* 43 1
* 43 2
10
38 1
40 1
41 3
42 1
43 2
45 1
11
* 41 2
* 42 2
* 43 2
13
* 39 1
* 42 1
* 42 1
* 43 1
43 2
47 1
14
40 1
42 1
43 2
15
40 1

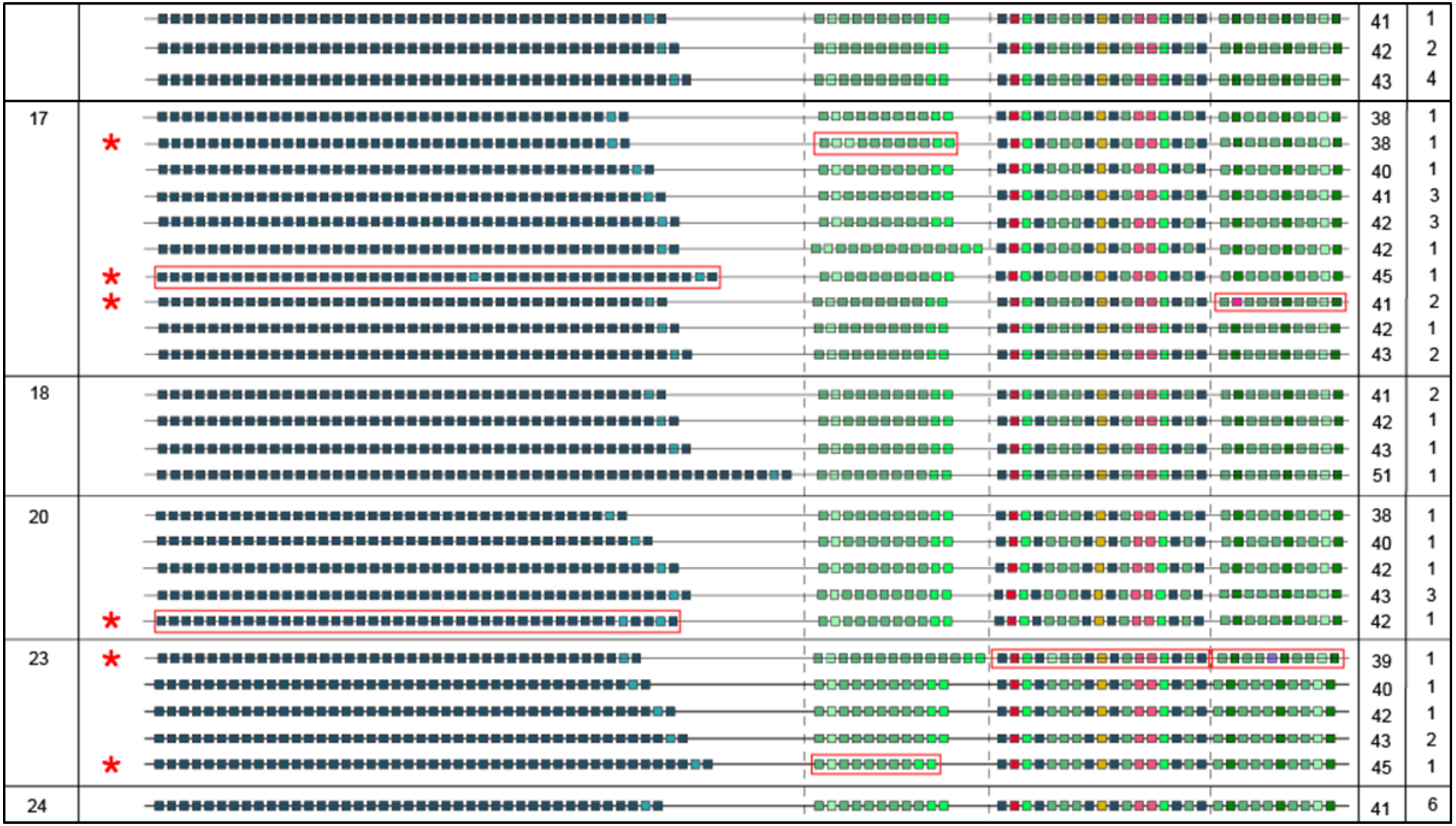

41 1
42 2
43 4
17
* 38 1
* 38 1
40 1
41 3
42 3
42 1
* 45 1
* 41 2
42 1
43 2
18
41 2
42 1
43 1
51 1
20
38 1
40 1
42 1
43 3
* 42 1
23
* 39 1
40 1
42 1
43 2
* 45 1
24
41 6

Figure 4.3. DNA sequences from successfully cloned HD patients

4.2.2. Illumina MiSeq Sequencing of the $(CAG)_{41}$ HD Patient Blood Samples

In order to validate the sequence configurations obtained from clone sequencing, Illumina MiSeq sequencing was performed on the DNA extracted from the blood of 30 HD patients, which yielded results for 29 HD patients (Table 4.3). HD patient 16 was unsuccessfully sequenced and therefore excluded. Illumina MiSeq sequencing revealed six HD patients (patients 1, 2, 4, 15, 22, 28) who did not have $(CAG)_{41}$, however, of these HD patients, their $(CAG)_n$ was within the ±2 CAG sensitivity bracket reported by fragment analysis. Illumina MiSeq sequencing highlighted four HD patients (patients 1, 2, 4, 22) with atypical expanded allele sequence configurations and one HD patient (patient 21) with an atypical wild-type sequence configuration. HD patients 1, 2, 4, and 22 have pure CAG repeats and presented with age at onsets 17, 15, 7, and 19.5 years earlier than their estimated mean age at onset based on having $(CAG)_{41}$, respectively (Table 4.1). Although to a lesser degree, presenting with an earlier age at onset is still evident when the estimated mean age at onset is based on the $(CAG)_n$ determined by Illumina MiSeq with HD patients 1, 2, 4 and 22 onsetting at 7, 5, 2 and 9.5 years earlier than their estimated mean age at onset, respectively.

Table 4.3. Illumina MiSeq sequencing results for HD patient blood samples

HD patient number	$(CAG)_n$	Age at onset	Allele1 (CAG)	Allele1 (CAACAG)	Allele1 (CCGCCA)	Allele1 (CCG)	Allele1 (CCT)	Allele2 (CAG)	Allele2 (CAACAG)	Allele2 (CCGCCA)	Allele2 (CCG)	Allele2 (CCT)	Wild-type allele (allele 1)	Expanded allele (allele 2)
1	41	40.5	17	1	1	7	2	43	0	0	12	2	17_1_1_7_2	**43_0*_0*_12_2**
2	41	42.5	21	1	1	7	2	43	0	1	7	2	21_1_1_7_2	**43_0_1_7_2**
3	41	48.5	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
4	41	50.5	21	1	1	7	2	42	0	1	7	2	21_1_1_7_2	**42_0_1_7_2**
6	41	51.5	17	1	1	10	2	41	1	1	7	2	17_1_1_10_2	41_1_1_7_2
7	41	56.5	25	1	1	7	2	41	1	1	7	2	25_1_1_7_2	41_1_1_7_2
8	41	56.5	19	1	0	9	2	41	1	1	7	2	19_1_0_9_2	41_1_1_7_2
9	41	55.5	15	1	1	10	2	41	1	1	7	2	15_1_1_10_2	41_1_1_7_2
10	41	58.5	20	1	1	10	2	40/41	1	1	7	2	20_1_1_10_2	41_1_1_7_2
11	41	55.5	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
13	41	59.5	15	1	1	7	2	41	1	1	7	2	15_1_1_7_2	41_1_1_7_2
14	41	60.5	17	1	1	10	2	41	1	1	7	2	17_1_1_10_2	41_1_1_7_2
15	41	63	22	1	1	7	2	40	1	1	7	2	22_1_1_7_2	40_1_1_7_2
17	41	63.5	17	1	1	10	2	41	1	1	7	2	17_1_1_10_2	41_1_1_7_2
18	41	67.5	18	1	1	7	2	41	1	1	7	2	18_1_1_7_2	41_1_1_7_2
19	41	67.5	21	1	1	7	2	41	1	1	10	2	21_1_1_7_2	41_1_1_10_2
20	41	73.5	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
21	41	X	16	2	1	7	3	41	1	1	7	2	16_2_1_7_3	41_1_1_7_2
22	41	38	21	1	1	7	2	43	0	0	12	2	21_1_1_7_2	**43_0*_0*_12_2**
23	41	X	20	1	1	7	2	41	1	1	7	2	20_1_1_7_2	41_1_1_7_2
24	41	X	17	1	1	10	2	41	1	1	7	2	17_1_1_10_2	41_1_1_7_2
25	41	58	17	1	1	9	2	41	1	1	7	2	17_1_1_9_2	41_1_1_7_2
26	41	X	18	1	1	7	2	41	1	1	7	2	18_1_1_7_2	41_1_1_7_2
27	41	61.5	20	1	1	7	2	41	1	1	7	2	20_1_1_7_2	41_1_1_7_2
28	41	54	15	1	1	10	2	40	1	1	7	2	15_1_1_10_2	40_1_1_7_2
29	41	61	15	1	1	7	2	41	1	1	7	2	15_1_1_7_2	41_1_1_7_2
30	41	62	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
31	41	62	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
32	41	66	15	1	1	10	2	41	1	1	7	2	15_1_1_10_2	41_1_1_7_2

$(CAG)_n$: CAG repeat tract of size n; X: information unavailable; dark red bold text: atypical sequences; *: missing both (CAA)(CAG) and (CCG)(CCA) sequences; 40/41: 61 reads displayed $(CAG)_{40}$, 62 reads displayed $(CAG)_{41}$; orange fill: allele sizes determined manually, sequencing data can confidently call the allele structure, however, there are an insufficient amount of reads to call the $(CAG)_n$ with a 1 CAG accuracy.

The confidence of each genotype reported per patient was calculated by ScaleHD (Table 4.4). Each allele originates with 100% confidence, which subsequently reduces if certain data characteristics (penalties) are encountered during the genotyping process (https://scalehd.readthedocs.io/en/latest/Definitions.html) (Chapter 2, Table 2.7). Such penalties include low peak thresholds, rare characteristics (homozygous haplotypes), atypical alleles, and total read count. Scores of 60% and above are considered to give reliable genotypes, whereas values below 60% require manual inspection. All of the expanded alleles in the HD patients had confidence scores of above 60%, with the exception of HD patient 10, which had the lowest confidence at 38%. Manual inspection of this sample confirmed the genotype and identified that it had a low read count, one of the penalties that deducts confidence (*inspection performed by Dr Marc Ciosi).* Deciphering the modal allele for HD patient 10 is more difficult as 62 reads aligned to 41 CAGs and 61 reads aligned to 40 CAGs. However, as the modal allele corresponds generally to the repeat length before which there is a significant decrease in reads, it is more likely that the modal allele contains 41 CAGs.

Table 4.4. Illumina MiSeq sequencing confidence results for HD patient blood samples

HD patient number	Confidence (%)		HD patient number	Confidence (%)	
	Wild type	Expanded		Wild type	Expanded
1	100	81	19	100	78
2	100	72	20	100	78
3	100	69	21	100	84
4	100	72	22	100	100
6	100	78	23	100	69
7	100	69	24	100	78
8	100	83	25	100	78
9	100	63	26	100	69
10	91	38	27	100	69
11	100	69	28	100	63
13	100	69	29	100	69
14	100	63	30	100	69
15	100	69	31	100	69
17	100	78	32	100	78
18	100	69			

4.2.3 PacBio SMRT Sequencing of Five HD Patient Blood Samples

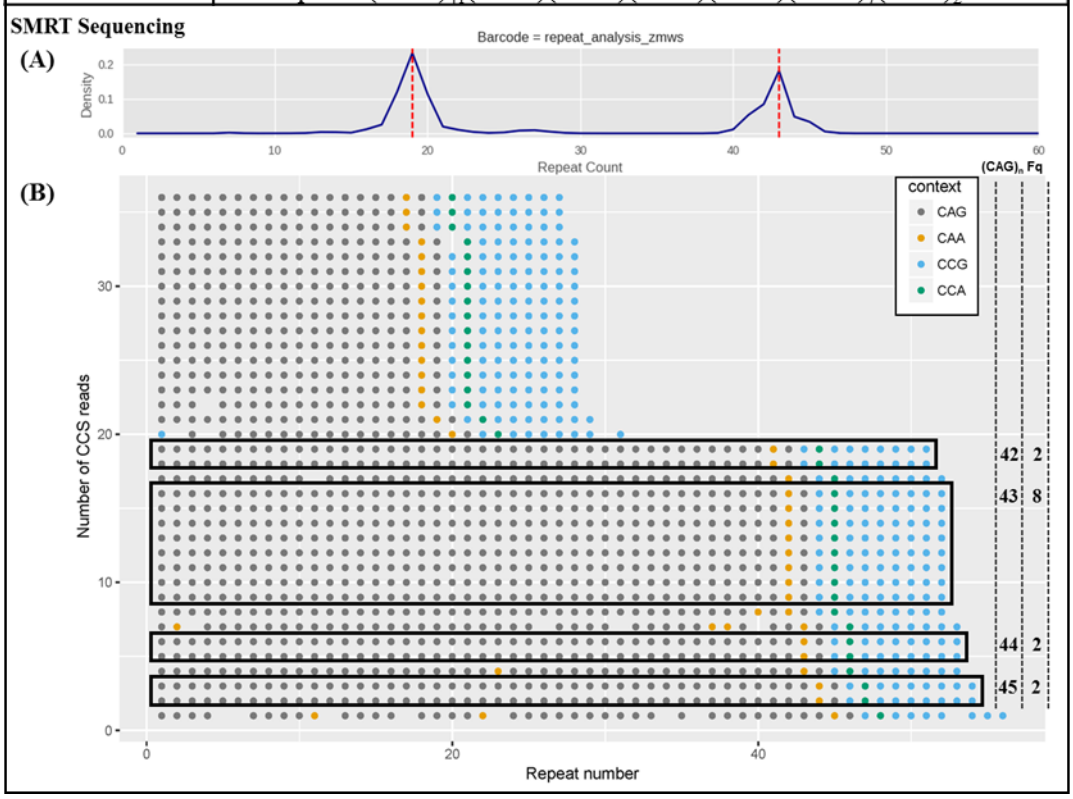
HD patients 3, 5, 8, 13, and 24 were prioritised for third-generation sequencing to explore how useful PacBio's SMRT sequencing is for determining the CAG and CCG repeat sequence configuration in relation to the Illumina MiSeq sequences. HD patients 3, 5, 13 and 24 were SMRT sequenced on the RSII instrument (Figure 4.4). HD patient 8 was sequenced on the latest platform, the Sequel System. Repeat analysis was unavailable for HD patient 5 as there was an insufficient amount of reads. The SMRT sequencing repeat count results in Figure 4.4 (A) determined the modal (red dotted line) $(CAG)_n$ at approximately 43, 43, and 42 CAGs in HD patients 3, 13, and 24, respectively, which is +2, +2 and +1 CAGs greater than that determined by Illumina MiSeq at 41 CAGs, respectively. An additional population of 42 and 43 CAGs were identified in HD patients 13 and 24, respectively, and the wild-type allele was not resolved in HD patient 24. With the exception of HD patient 3, SMRT sequencing determined the expanded allele sequence configuration to be invariant at $(CAG)_n(CAA)(CAG)(CCG)(CCA)(CCG)_7$, with no sequence alterations identified (Figure 4.4 (B)). HD patient 3 presented with 2 sequences that contain an additional CAA interruption to the penultimate CAA trinucleotide. The additional CAA trinucleotides are positioned variability within the CAG repeat with one CAA located at the previously reported position of 2 trinucleotides before the penultimate CAA and the other CAA is positioned after the first 22 CAGs.

The remaining sample, HD patient 8, was subsequently analysed on the Sequel System. The DNA quality results determined that it was highly fragmented (Figure 4.5). HD patient 8 had the smallest average fragment size of 4,873 bp compared to PacBio's internal control, sample M94, which had the largest average fragment size of 50,239 bp, 10-fold that of HD patient 8. Therefore, due to insufficient quality and quantity of DNA for the SMRT library preparation, no result was obtained for HD patient 8.

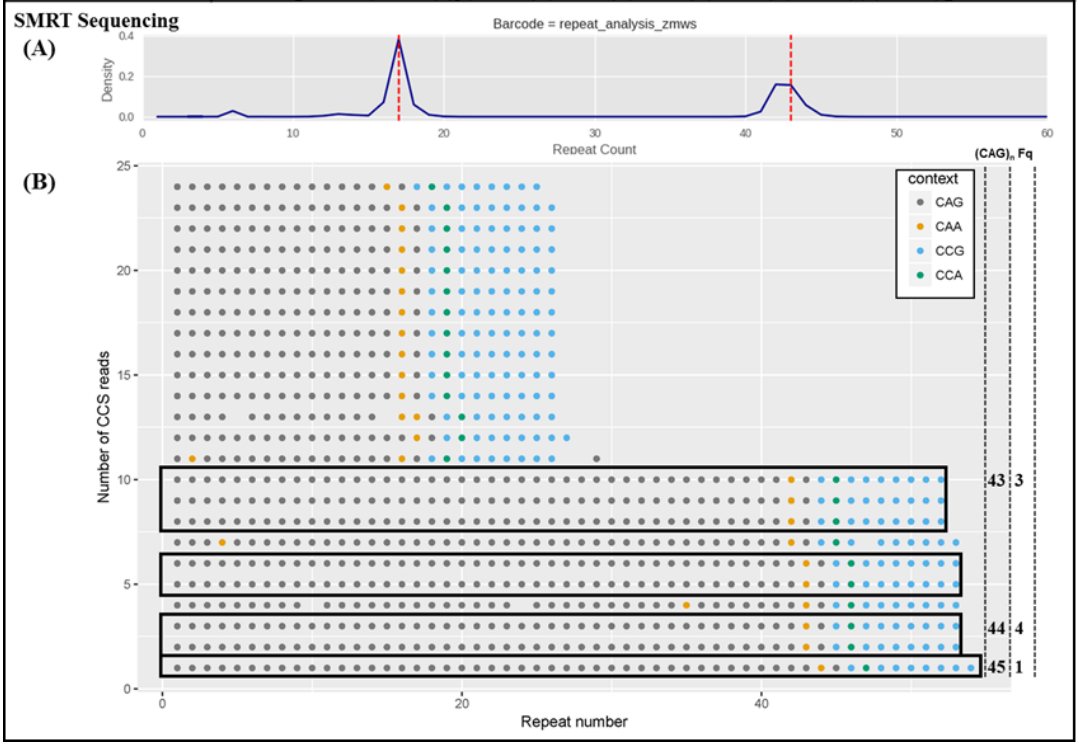
Figure 4.4. Sequence configuration determined by Illumina MiSeq and SMRT sequencing (overleaf)

MiSeq EA: expanded allele sequence determined by Illumina MiSeq. PacBio SMRT sequencing repeat analysis is shown as repeat count (A) and repeat number (B). Black boxes: highlight the sequence configuration of intact sequences; Gaps: trinucleotides which could not be read accurately due to an indistinguishable signal-to-noise ratio between incorporated and unincorporated bases; $(CAG)_n$: CAG repeat number including CAA trinucleotides; Fq: frequency.

HD Patient 3	MiSeq EA: $(CAG)_{41}(CAA)(CAG)(CCG)(CCA)(CCG)_7(CCT)_2$



HD Patient 13	MiSeq EA: $(CAG)_{41}(CAA)(CAG)(CCG)(CCA)(CCG)_7(CCT)_2$



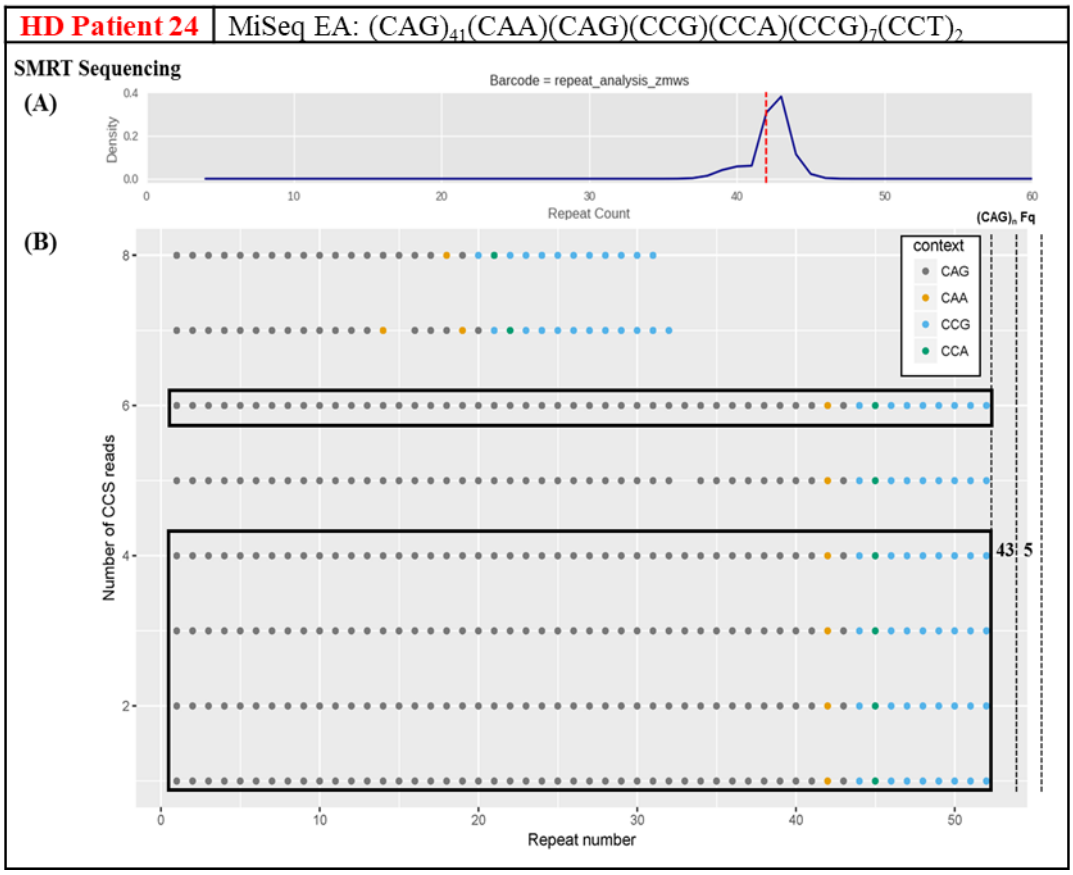
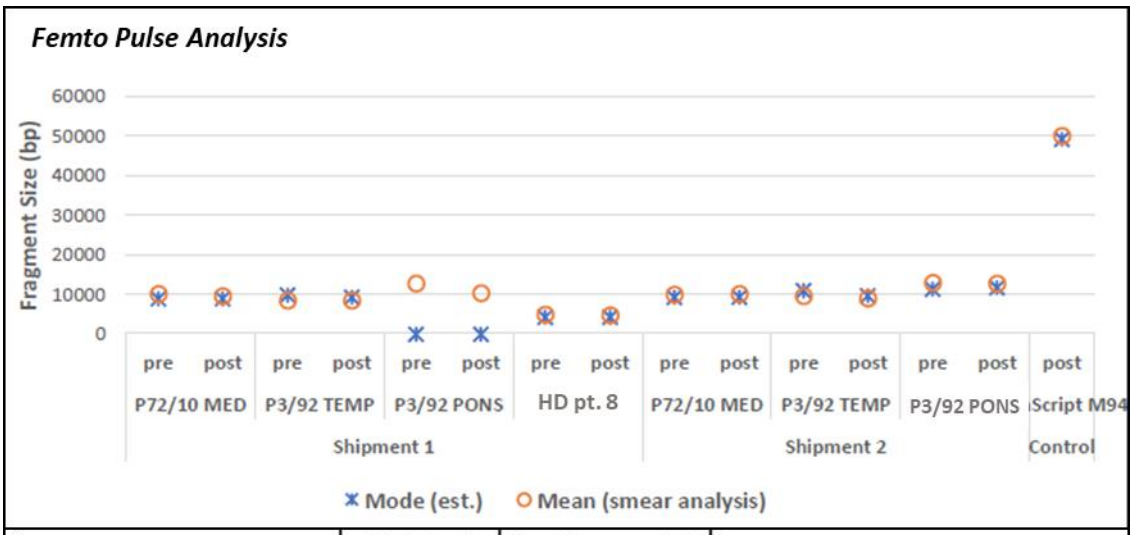


Figure 4.4. Sequence configuration determined by Illumina MiSeq and SMRT sequencing



Fragment size

Shipment	Sample ID	AMPure State	Mode (est.) Value	Mean (smear analysis) Value
Shipment 1	P72/10 MED	pre	8907	10235
		post	8907	9721
	P3/92 TEMP	pre	9757	8551
		post	9198	8556
	P3/92 PONS	pre	low RFU	12824
		post	low RFU	10380
	HD pt. 8	pre	4290	4873
		post	4217	4829
Shipment 2	P72/10 MED	pre	9316	10058
		post	9218	10243
	P3/92 TEMP	pre	10856	9551
		post	9609	9125
	P3/92 PONS	pre	11360	12959
		post	11762	12762
Control	GenScript M94	post	49154	50239

Figure 4.5. FEMTO Pulse analysis of HD patient samples

The FEMTO Pulse quantifies, qualifies and sizes DNA samples using a pulse-field power supply with an optical detection platform yielding highly sensitive detection of DNA down to the femtogram range. HD pt. 8: HD patient 8, this sample was not sent in the second shipment due to lack of availability. The fragment size for HD patient 8 was 4,873 bp compared to PacBio's internal control, GenScript M94, with a fragment size of 50,239 bp. P72/10 MED, P3/92 TEMP and PONS: HD patient *post-mortem* brain samples sequenced for somatic instability analysis, which is discussed in Chapter 6, section 6.2.2.

4.2.4 Nanopore Sequencing of the HD Patient Blood Samples

DNA from 26 HD patients with concentrations greater than 70 ng/µL were sent to Kings College London for Nanopore sequencing by Dr Graham Taylor. The raw data obtained from the Nanopore sequencing was extremely noisy (data not shown). In collaboration with David Murphy (*UCL Queen Square Genomics)*, the output files were run through a bioinformatics pipeline based on RepeatHMM to extract the CAG repeat sizes only (Chapter 2, section 2.12) (Liu et al., 2017). The Nanopore $(CAG)_n$ sizing data is presented in graph format, which can be seen in *Supplementary Data Figure 1* and summarised in Table 4.5 alongside the CAG repeat sizing results from fragment analysis and Illumina MiSeq. In comparing the Nanopore sequencing CAG repeat sizes of the expanded allele to those determined by fragment analysis, the $(CAG)_n$ differs by ± 1 CAG, which is within the sensitivity bracket reported by fragment analysis. Similarly, the expanded allele CAG repeat size determined by Nanopore sequencing differs by ± 1 CAG compared to the Illumina MiSeq CAG repeat sizing results.

In order to determine the fidelity of the CAG repeat sizing results from the Nanopore sequencing data, the total number of reads from 0 to approximately 70 CAGs were calculated per sample in which the modal wild-type and expanded allele read counts were determined as a percentage of total reads (Table 4.6). The threshold of 70 CAGs was set due to the lack of reads thereafter. HD patient 1 presents with the lowest percentage of read counts for the expanded allele with 2.6% of total reads mapping to 42 CAGs. HD patient 24 presents with the highest percentage of read counts for the expanded allele with 9.4% of total reads mapping to 42 CAGs. The total read count per sample can be found in *Supplementary Data Figure 1*.

Table 4.5. $(CAG)_n$ sizing by fragment analysis, Illumina MiSeq and Nanopore sequencing

HD patient number	Fragment analysis		Illumina MiSeq		Nanopore sequencing	
	WT	EA	WT	EA	WT	EA
1	19	41	17	43	19	42
2	22	41	21	43	22	42
3	18	41	17	41	19	42
4	21	41	21	42	23	41
6	19	41	17	41	19	42
7	26	41	25	41	26	42
9	16	41	15	41	17	41
11	18	41	17	41	19	42
13	18	41	15	41	17	41
14	18	41	17	41	19	41
15	23	41	22	40	24	41
17	18	41	17	41	19	41
18	19	41	18	41	20	42
19	22	41	21	41	22	42
20	19	41	18	41	20	42
21	19	41	16	41	20	41
23	21	41	20	41	22	42
24	18	41	17	41	19	42
25	18	41	17	41	19	41
26	18	41	18	41	20	41
27	21	41	20	41	22	41
28	15	41	15	40	-	-
29	16	41	15	41	17	42
30	18	41	17	41	19	41
31	18	41	17	41	19	41
32	16	41	15	41	17	42

WT: wild-type allele CAG repeat size; EA: expanded allele CAG repeat size; - : CAG repeat size undetermined.

Table 4.6. Nanopore sequencing repeat count percentages for wild-type and expanded alleles of HD patient blood samples

HD patient number	Read count (%)		HD patient number	Read count (%)	
	Wild type	Expanded		Wild type	Expanded
1	23.4	2.6	19	13.8	5.4
2	13.6	6	20	11.3	7.9
3	16.6	6.3	21	15.4	6.6
4	13.2	7.3	23	14.4	6.2
6	11.6	7.6	24	10.2	9.4
7	11.8	6.8	25	16.1	6
9	12.4	7.8	26	15.8	6
11	15.7	7.3	27	13.7	6.7
13	16.3	6.5	29	18.2	6
14	11.4	7.9	30	14.2	7
15	12.1	7.6	31	16.5	5.8
17	11.6	8	32	11.4	8.4
18	15.7	5.8			

4.2.5 Using $(CAG)_n$ Sizing Results from Fragment Analysis, Illumina MiSeq and Nanopore Sequencing as Predictors of HD Patient Age at Onset

To investigate which CAG sizing method best predicted the HD patients age at onset in relation to actual age at onset, their mean age at onset was estimated based on the Langbehn model with the CAG sizes determined by fragment analysis, Illumina MiSeq and Nanopore sequencing (Table 4.7) (Langbehn et al., 2004). Only HD patients with age at onset information and CAG sizing results from all three methods were analysed. The age at onset predicted by fragment analysis, Illumina MiSeq and Nanopore sequencing was subtracted from the HD patients' actual age at onset to calculate the residual age at onset, which is graphed in Figure 4.6. There is no statistical difference in residual age at onset between the three CAG repeat sizing methods ($p = 0.1807$). However, the slope of the line from Illumina MiSeq is closer to zero, which suggests that Illumina MiSeq more accurately predicts the actual age at onset. The R^2 values determined that Nanopore sequencing has the highest variance between the data points to the trendline ($R^2 = 0.3915$), compared to Illumina MiSeq ($R^2 = 0.4477$) and fragment analysis ($R^2 = 0.5162$).

Table 4.7. Estimated mean age at onset determined by fragment analysis, Illumina MiSeq and Nanopore sequencing $(CAG)_n$ sizing results

HD patient number	Actual age at onset	Estimated mean age at onset		
		Fragment analysis	MiSeq	Nanopore
1	40.5	57.5	47.5	52.5
2	42.5	57.5	47.5	52.5
3	48.5	57.5	57.5	52.5
4	50.5	57.5	52.5	52.5
6	51.5	57.5	57.5	52.5
7	56.5	57.5	57.5	52.5
9	55.5	57.5	57.5	52.5
11	55.5	57.5	57.5	57.5
13	59.5	57.5	57.5	52.5
14	60.5	57.5	57.5	52.5
15	63	57.5	60	57.5
17	63.5	57.5	57.5	47.5
18	67.5	57.5	57.5	52.5
19	67.5	57.5	57.5	52.5
20	73.5	57.5	57.5	52.5
25	58	57.5	57.5	52.5
27	61.5	57.5	57.5	52.5
29	61	57.5	57.5	57.5
30	62	57.5	57.5	57.5
31	62	57.5	57.5	52.5
32	66	57.5	57.5	52.5

Estimated mean age at onset in years is calculated based on the fragment analysis, Illumina MiSeq, and Nanopore sequencing $(CAG)_n$ sizing results (Langbehn et al., 2004).

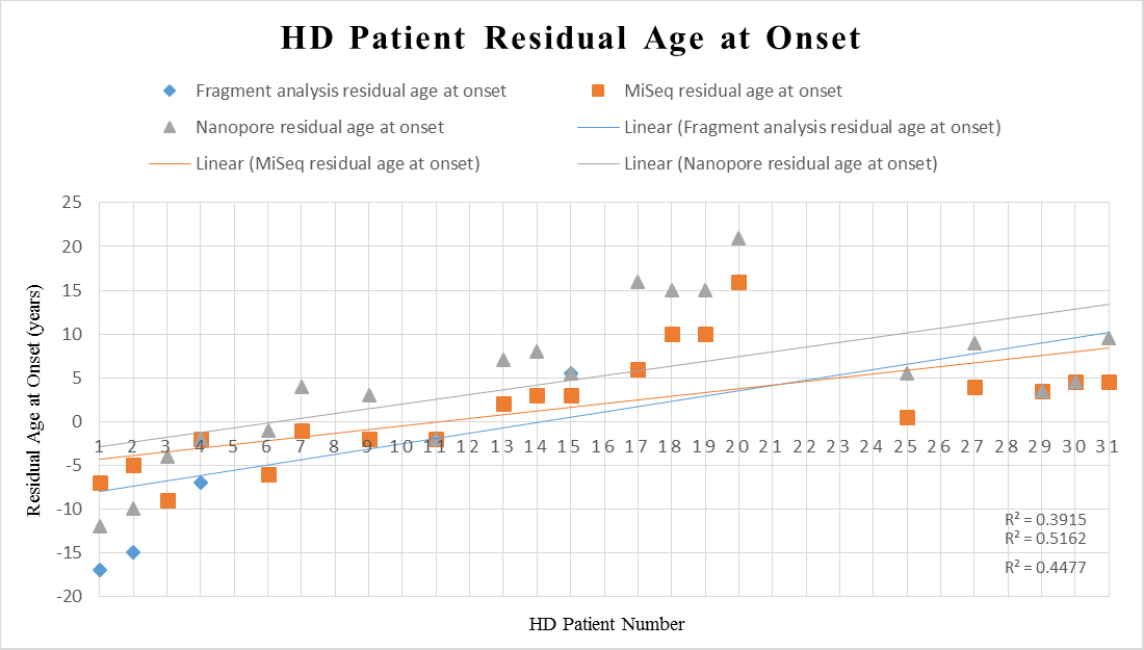


Figure 4.6. HD patient's residual age at onset

There was no statistical difference between the residual age at onset calculated from the $(CAG)_n$ determined by fragment analysis, Illumina MiSeq and Nanopore sequencing, ANOVA was carried out in Excel ($p = 0.1807$).

4.3 Discussion

4.3.1 Sequence Alterations Identified by Clone Sequencing and Potential Sources of Experimentally Induced Artefacts

It is challenging to form a solid conclusion on the faithfulness of the sequence alterations identified in the HD patients as this work is limited by both patient and clone number, and by the methodology. In the sequences determined by clone sequencing, an additional CAA trinucleotide is variably positioned throughout the CAG repeat in one clone each from HD patients 17, 20, 30 and 31. However, previous and recent research has described the additional CAA interruption to be consistently located at the fourth last trinucleotide of the CAG repeat; $(CAG)_n$(**CAA**)(CAG)(CAA)(CAG) (Pêcheux et al. 1995; Wright et al. 2019; Lee et al. 2019). HD patients 2, 3, 7 and 28 contain at least one clone that lacks the penultimate CAA trinucleotide, resulting in pure CAG tracts. In relation to the total number of clones obtained, patient 2 contained the highest percentage of pure CAG repeat tracts (66%) with $(CAG)_{43}$, which was subsequently confirmed by Illumina MiSeq. Illumina MiSeq does not recapitulate the mixed CAA-interrupted CAG repeat sequences or the pure CAG sequences reported by clone sequencing in the above HD patients. Loss of the first CCG trinucleotide in CCG1 was identified in one clone only from HD patient 3 who contained a pure tract of 39 CAGs. This sequence variation has recently been reported in both wild-type and pathogenic HD alleles, as well as in non-HD controls, which gives greater confidence in this finding (GEM-HD, 2019). Yet, Illumina MiSeq does not confirm this sequence and as it is present in only one clone, it suggests that this and the intra-CAG repeat alterations in this cohort are artefacts introduced through clone sequencing.

Online tools which record human genetic variants are one such method to filter out potential errors and can confirm the fidelity of the sequence alterations identified within this cohort of HD patients. In the data sets that are openly accessible, there is one variant described at the third last trinucleotide in CCG1 which results in an inframe insertion due to a G:T substitution and can be seen in two clones from HD patient 24 and one clone in HD patient 25 (http://www.ensembl.org/Homosapiens/Location/View?db=core;g=ENSG00000197386;r=4:3074943-3075098). This substitution has also been reported as a synonymous insertion as CCT additionally codes for proline and current literature has described a short CCG1 containing $(CCG)_6$ and its association with $(CCT)_3$, which further supports this finding in HD patients 24 and 25 (Pêcheux et al. 1995; Lee et al. 2019). In the Exome Aggregation Consortium (ExAC) database, three variants have been identified

in the inter-CCG repeat region out of approximately 9,000 alleles sequenced. These variants include the following substitutions, which result in synonymous proline-coding codons; G:T, G:A, and C:T (http://exac.broadinstitute.org/variant/4-3076714-G-T, http://exac.broadinstitute.org/variant/43076714-G-A, http://exac.broadinstitute.org/variant/43076733-C-T). The G:A substitution is reported in one clone from HD patient 23, which is located at the fifth trinucleotide in the inter-CCG repeat region. However, there is no variant described in the 122 bp from the CCG1 repeat to the end of HTT exon 1 in the 1000 genomes project data (http://www.ensembl.org/Homosapiens/Location/View?-db=core;g=ENSG00000197386;r=4:30749433075098), which suggests that the CCG2 sequence alterations described in this report are artefacts that have been introduced experimentally.

In clone sequencing, PCR effectively purifies a target DNA sequence away from the rest of the genome for insertion into self-replicating bacterial cells to generate identical copies of the target sequence. Although steps have been taken to minimise the introduction of artefacts including the use of a high-fidelity proofreading polymerase for PCR, the PCR-blunt vector with a low background of non-recombinants and recombination-deficient Stbl3 *E.coli*, it is not possible to completely eliminate experimental artefacts. Base substitutions are the predominant errors associated with DNA polymerases in PCR reactions, with G:A and C:T transitions acting as the largest class of mutations for proofreading polymerases (Potapov and Ong, 2017). The substitutions reported in this study by clone sequencing are mostly transitions of A:G, G:A, G:T, C:T and C:A with only one transversion identified, C:G. This suggests that these base substitutions are errors which have been introduced during the PCR reaction and cannot be held accountable for the phenotypic variation reported in this cohort of HD patients (Castillo-Lizardo et al., 2014).

4.3.2 Methods Determining CAG Repeat Length; Fragment Analysis, Illumina MiSeq and Nanopore Sequencing

For HD patients, an agonising question for them is “when will the disease onset?”, if it hasn’t already. The CAG repeat length defines HD development and is the primary indicator for age at disease onset and severity, which necessitates the accuracy of CAG repeat sizing (Bates et al., 2015). This study used fragment analysis, Illumina MiSeq and Nanopore sequencing to size the CAG repeat in our cohort of HD patients. Traditional sizing of the CAG repeat is performed by PCR based fragment analysis in which the number of CAGs is estimated from the PCR product by capillary electrophoresis.

Fragment analysis sized the HD patient cohort presented here with 41 CAGs, with a sensitivity bracket of ±2 CAGs. Illumina MiSeq revealed six out of 29 HD patients that had sizes other than 41 CAGs, but within the sensitivity threshold set by fragment analysis. Additionally, although Nanopore sequencing sized these patients within the sensitivity bracket, it reported 13 out of 25 HD patients with sizes greater than 41 CAGs.

To identify which sizing method best predicts age at onset in relation to each patient's actual age at onset, the residual age at onset was plotted for each patient. Statistically, there was no significant difference between the sizing methods. However, based on the slope of the line in which a value of zero describes the actual age at onset for the HD patients, Illumina MiSeq depicted the best fit (m=0.3091) compared to fragment analysis (m=0.4737) and Nanopore sequencing (m=0.5386). This suggests that predicting the mean age at onset is more accurate when estimating from CAG sizes determined by Illumina MiSeq. The confidence scores reported by ScaleHD give further assurance in the faithfulness of the Illumina MiSeq data. Only HD patient 10 had confidence scores below the threshold of 60% for the expanded allele in which subsequent manual inspection confirmed the accuracy of the genotype and the factor hindering the confidence score, reduced read count due to poor PCR efficiency (*Dr Marc Ciosi*, *personal communication*). The remaining HD patients had confidence scores of 100% for the wild-type allele and > 62% for the expanded allele, which reinforces the correctness of the genotyping results.

The most common definition of read accuracy in Nanopore sequencing is the percentage of bases in a segment of a read that match with a reference relative to the length of the read segment minus the reference alignment. Additionally, the read accuracy is dependent on the alignment algorithm performance, in which different alignment tools can result in different reported accuracies (Rang et al., 2018). As our data set was solely used to size the CAG repeat, the largest number of reads that mapped to the wild-type and expanded CAG repeat was calculated as a percentage against the total sum of reads which mapped from 0 to 70 CAGs. In all of the HD patients sequenced by Nanopore, the wild-type allele had the highest percentage of read counts compared to the expanded allele, which ranged from 0.8% to 20.8% greater than that of the expanded allele. The tools for aligning Nanopore sequencing long-read data have not been thoroughly evaluated and it has been reported that additional optimisations are needed to improve structural variation detection accuracy and sensitivity, which could account for why the wild-type allele is better resolved than the expanded allele (Zhou et al., 2019).

Additionally, Nanopore sequencing on the MinION is associated with low read accuracy in comparison to short-read technologies and has an error rate of approximately 5-20%, which is significantly greater than the 0.1% associated with Illumina next generation sequencing (Pfeiffer et al., 2018; Sedlazeck et al., 2018). There are two sources of errors which can arise from Nanopore sequencing; sequencing errors due to a low signal-to-noise ratio and errors in the translation of the raw electric current into a DNA sequence due to incorrect interpretation in the analysis (Rang et al., 2018). The low read accuracy, which interferes with the detection analysis of single nucleotide variations and thus requires high-coverage sequencing, and the high percentage of error rates suggests that further optimisation is needed for this technology to accurately size and decipher complex regions of DNA.

4.3.3 Illumina MiSeq Genotyping-by-Sequencing and PacBio SMRT Sequencing

The main impediment in fragment analysis is the lack of sensitivity and the bias towards sizing CAGs that conform to the canonical sequence of $(CAG)_n(CAA)(CAG)(CCG)(CCA)(CCG)_7(CCT)_2$. Sequences that deviate from this are often not assessed due to incompatibility with the amplification protocol or mis-sized. If a sample appears homozygous for a wild-type allele, additional testing is often required to ensure that an expanded allele was not amplified during the PCR reaction, which includes testing for heterozygosity in the CCG repeat by amplifying over the CAG repeat and flanking CCG repeat regions (Jama et al., 2013). Similarly, fragment analysis does not determine the base pair sequence configuration and thus cannot inform on the presence of any sequence alterations that may be influencing the phenotype. Illumina MiSeq of *HTT* exon 1 quantifies the number of CAGs, CCGs, and identifies atypical sequence variations. The major benefit of this approach is that it can reveal *HTT* variants that have been missed by fragment analysis and accurately determine the CAG repeat length, which in turn has the potential to improve diagnostics.

Illumina MiSeq revealed atypical sequences in HD patients 1, 2, 4 and 22, who were previously sized with 41 CAGs. The sequences of these four patients did not contain the penultimate CAA trinucleotide and HD patients 1 and 22 had pure CCG1 repeat regions. Based on their $(CAG)_n$ determined by Illumina MiSeq, HD patients 1, 2, 4 and 22 have ages at onsets that are -7, -5, -2 and -9.5 years earlier than their estimated mean age at onset. This is consistent with current reports detailing the correlation between the absence of the CAA interruption and an earlier age at onset (Lee et al. 2019). The CCA

trinucleotide in the flanking CCG1 sequence is absent from HD patients 1 and 22 resulting in a sequence containing a pure run of 12 CCGs. Both patients present with the earliest age at onsets (-7 and -9.5 years, respectively). In combination with a pure CAG repeat tract, a pure CCG repeat has most recently been associated with an onset-hastening modifier effect in HD, which highlights the influence of downstream sequences on onset-determining properties (GEM-HD, 2019). Therefore, Illumina MiSeq identified sequence alterations that can be held accountable for some of the phenotypic variance observed in these 4 patients, which could not be determined by fragment analysis (Lee et al. 2019; Wright et al. 2019).

For the HD patient blood-derived samples sequenced by PacBio's SMRT sequencing, two yielded enough on-target molecules to give convincing results; 39 molecules for HD patient 13 and 48 molecules for HD patient 3. HD patient 24 yielded 10 on-target molecules, which is sufficient for the repeat analysis tool but insufficient to give full confidence in the result. HD patient 8 was analysed on the Sequel System, which yielded no results, as the sample did not survive library preparation. PacBio SMRT sequencing has an error rate of 10-15% with indels being most prominent (Sedlazeck et al., 2018). In considering all sequences outputted from SMRT sequencing, additional CAA interruptions were identified dotted throughout the CAG repeat in both wild-type and expanded alleles. However, current literature reports that the additional CAA interruption is invariably located two trinucleotides prior to the penultimate CAA (GEM-HD, 2019). This suggests that the additional CAA interruptions identified by SMRT sequencing could be artefacts. The PacBio SMRT sequencing data depicts the base pair configuration, yet due to the limited number of samples and reads, these results instead highlight the potential to use SMRT sequencing in the future with further optimisation.

Chapter 5. Investigating a Panel of DNA Repair Pathway SNPs as Potential Phenotypic Modifiers

5.1 Background

The CAG repeat length only accounts for up to 60% of the variation observed in age at motor onset in HD patients carrying expansions of 40 to 55 CAGs, with the remaining percentage attributable to heritable factors, such as genetic variants (Bates et al., 2015; Wexler et al., 2004). In search of these genetic variations, a genome-wide association study (GWAS) was carried out on just over 4,000 HD patients using their residual age at onset, which is the difference between their estimated mean age at onset based on their CAG repeat size and their actual age at onset (GEM-HD, 2015). The GWAS identified two loci with three significant modifier signals. The first genome-wide locus associated with modifying HD age at onset was identified on chromosome 15 (*FAN1*) and had two independent effects, which accelerate or delay onset by 6.1 years and 1.4 years, respectively. The second locus was identified on chromosome 8 (*PMS2*), which was associated with an earlier age at onset by 1.6 years (GEM-HD, 2015). Pathway analysis of the significant modifier signals highlighted DNA maintenance and mitochondrial regulation as the two prominent modifying processes (GEM-HD, 2015). This work was extended to over 9,000 HD patients, which replicated the previous DNA repair modifier loci and identified additional DNA repair pathway genes with modifying effects on age at onset (GEM-HD, 2019). Most recently, DNA repair-associated loci which have disease onset determining properties have been identified on chromosome 2 (*PMS1*), chromosome 3 (*MLH1*), chromosome 5 (*MSH3*/*DHFR*), chromosome 7 (*PMS2*), chromosome 15 (*FAN1*) and chromosome 19 (*LIG1*) (GEM-HD, 2019). To examine if DNA repair modifier loci were influencing the age at onset in our HD cohort, the patient blood samples with approximately $(CAG)_{41}$ and the six HD *post-mortem* brains were genotyped on the NeuroChip array against a customised panel of 23 SNPs all associated with DNA repair pathway genes (Table 5.1). Each gene was prioritised based on recent studies of associations between age at onset and DNA repair genetic variants in CAG repeat-expanded diseases and the most significant SNPs for each gene were selected (Bettencourt et al., 2016; GEM-HD, 2015, 2019; Moss et al., 2017).

Table 5.1. DNA repair pathway SNP panel and corresponding proxy SNPs

SNP ID	Chr:position (bp) (GRCh38)	Gene
rs1037699	8:102238702	*RRM2B*
rs1037700	8:102238547	*RRM2B*
rs114136100	15:30905773	*FAN1*
rs115109737	5:80806625	*MSH3*
rs12531179*	7:5989056	*PMS2*
rs1382539[x]	5:80656335	*MSH3*
rs146353869	15:30834198	*HERC2P10*
rs150393409	15:30910758	*FAN1*
rs16869352	8:102293805	*UBR5*
rs175080*	14:75047125	*MLH3*
rs1799977	3:37012077	*MLH1*
rs1800937	2:47798625	*MSH6*
rs1805323*	7:5987311	*PMS2*
rs20579*	19:48165573	*LIG1*
rs3512*	15:30942802	*FAN1*
rs3735721	8:102205467	*RRM2B*
rs4150407*	2:127292055	*ERCC3*
rs5742933*	2:189784590	*PMS1*
rs5893603	8:102238612	*RRM2B*
rs6151792*	5:80761142	*MSH3*
rs61752302	8:102298925	*UBR5*
rs71636247	5:80823157	*MSH3*
rs72734283	14:75028356	*MLH3*
Reference SNP ID	**Proxy SNP ID (r^2 value)**	**Proxy Gene**
rs12531179*	rs852151 (0.811)	*EIF2AK1*
rs175080*	rs175084 (1.0)	*MLH3*
rs1805323*	rs12534423 (1.0)	*PMS2*
rs20579*	rs3730872 (0.732)	*LIG1*
rs3512*	rs11293 (1.0)	*FAN1*
rs4150407*	rs1566822 (1.0)	*ERCC3*
rs5742933*	rs3791767 (1.0)	*ORMDL1*
rs6151792*	rs6151816 (0.806)	*MSH3*

Proxy SNPs were identified in high linkage disequilibrium (LD) with the most significant SNP (r^2 > 0.7) using LDlink (https://analysistools.nci.nih.gov/LDlink/) in the British population. Chr: chromosome; dark red and bold: not available on the NeuroChip array and/or proxy SNP r^2 < 0.7; *: proxy SNPs required; [x]: rs1382539 was analysed instead of rs557874766 as the genotyping data of rs1382539 are of higher quality, however the two are in high LD and tag the same association signal (Moss et al., 2017). Chr: chromosome; bp: base pair; GRCh38: Genome Reference Consortium Human Build 38.

5.2 Results

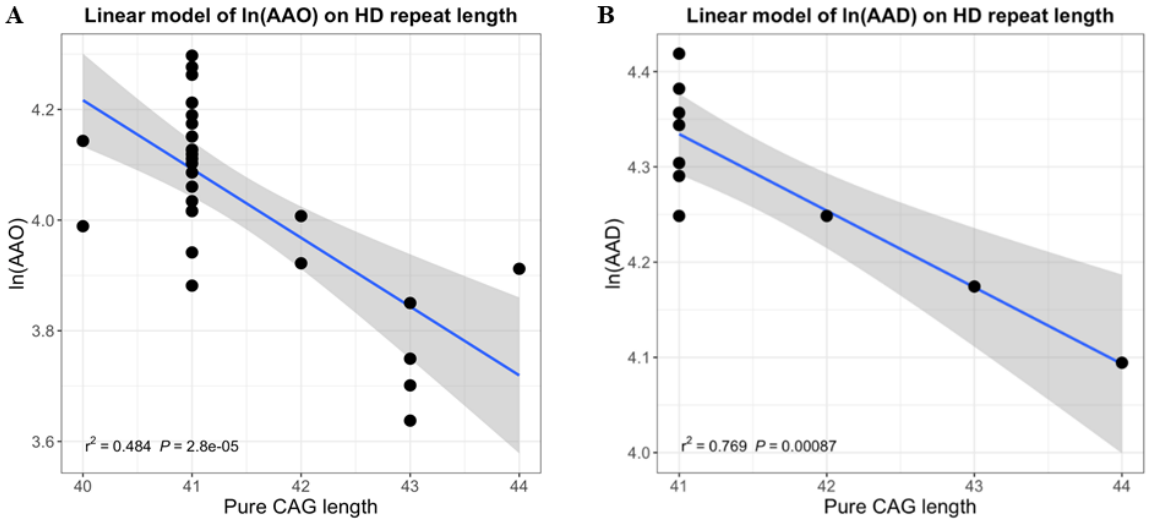
5.2.1 Cohort Descriptive

The 39 sample cohort consisted of 27 HD patient blood samples, six HD *post-mortem* brains and six control *post-mortem* brains. DNA was extracted from the blood of all samples with the exception of three HD *post-mortem* brains in which the DNA was extracted from cerebellar tissue. If the blood of the corresponding HD *post-mortem* brain was unavailable, the cerebellum was chosen based on somatic mosaicism results, as its instability profile is closest to that of blood (Chapter 6, Figure 6.1). Age at onset, which was defined by motor onset, was available for 23 HD patient blood samples and six HD *post-mortem* brains, and age at death for four HD patient blood samples, six HD *post-mortem* brains and six control *post-mortem* brains (**Error! Reference source not found.**2). The following analysis included only the HD samples with age at onset and/or age at death available. Linear regression analysis was performed with the natural logarithm (ln) of age at onset and age at death on expanded pure CAG repeat length as determined by Illumina MiSeq with the corresponding regression parameters (Figure 5.1). The regression parameters were used to calculate an expected age at onset value based on CAG repeat length, which was then subtracted from the HD patients' actual age at onset to give a residual age at onset. In this cohort, CAG repeat length accounted for 48.4% of variability in age at onset and 95.2% of variability in age at death, with each CAG advancing age at onset and age at death by 0.88 years ($p = 2.8^{e-05}$) and 0.93 years ($p = 8.67^{e-04}$), respectively. Similar regression analysis parameters were achieved in Bettencourt *et al*., 2016 with a much larger cohort of 445 HD patients (Bettencourt et al., 2016).

Table 5.2. Sample cohort

Condition	**n**	**CAG** mean ± sd (range)	**Age at onset** mean ± sd (range)	**Age at death** mean ± sd (range)
HD	33	41.33 ± 0.89 (40-44)	57.88 ± 9.21 (38-73.5)	73 ± 7.01 (60-83)
Control	6	-	-	81.00 ± 13.04 (56-93)

n: number; sd: standard deviation; CAG: pure CAG length of the expanded allele as determined by Illumina MiSeq.



Condition	n	Intercept	B	R^2	p
Age at onset	29	9.1903403	-0.1243404	0.484	2.82^{e-05}
Age at death	10	7.6328570	-0.0804494	0.769	8.67^{e-04}
HD (Bettencourt et al., 2016)	445	6.119939	−0.052966		$< 2^{e-16}$

Figure 5.1. Linear regression and parameters of ln(age at onset) and ln(age at death) on expanded pure CAG repeat length

AAO: age at onset; AAD: age at death; B: regression parameter, compares the relationship of the independent variable (CAG length) to the dependent variable (age at onset/death), value from 0 to 1 or 0 to -1 depending on the direction of the relationship; R^2: indicates the percentage of the variance in the dependent variable that the independent variable explains; Bettencourt et al., 2016: Results of fitting a linear regression ln(age at onset) = A + B*(CAG)n; p: the significance of the regression parameter (B) indexing the effect of repeat length.

5.2.2 NeuroChip Genotyping

All samples were successfully genotyped on the NeuroChip array for the customised panel of DNA repair pathway SNPs. Hardy-Weinberg equilibrium (HWE) was calculated using the Hardy-Weinberg package (Graffelman, 2020) and Bonferroni corrected to protect against population stratification effects (Table 5.3). All SNPs had a Hardy-Weinberg p value of > 0.001 and thus, none were removed for further analysis. Additionally, none of the SNPs significantly deviate from HWE, which is displayed through ternary plots for which the HWE condition defines a parabola (Figure 5.2).

Table 5.3. Genotypes, allele frequencies, and HWE for the directly sequenced SNPs

SNPs sequenced	Gene	proxy SNP of interest	LD (r^2)	HD Genotype	Controls Genotype	MAF (HD)	MAF (CTRLs)	HWE p (HD)	HWE p (CTRLs)	HWE p corr (HD)	HWE p corr (CTRLs)
rs1037699	RRM2B			36/2/0	6/0/0	0.026	0.000	1.000	1.000	1.000	1.000
rs1037700	RRM2B			36/2/0	6/0/0	0.026	0.000	1.000	1.000	1.000	1.000
rs11293	FAN1	rs3512	1.000	18/18/2	3/3/0	0.289	0.250	0.459	1.000	1.000	1.000
rs114136100	FAN1			36/2/0	6/0/0	0.026	0.000	1.000	1.000	1.000	1.000
rs115109737	MSH3			31/6/1	6/0/0	0.105	0.000	0.336	1.000	1.000	1.000
rs12534423	PMS2	rs1805323	1.000	34/4/0	6/0/0	0.053	0.000	1.000	1.000	1.000	1.000
rs1382539	MSH3			17/19/2	1/4/1	0.303	0.500	0.443	1.000	1.000	1.000
rs146353869	HERC2P10			37/1/0	6/0/0	0.013	0.000	1.000	1.000	1.000	1.000
rs150393409	FAN1			36/2/0	6/0/0	0.026	0.000	1.000	1.000	1.000	1.000
rs1566822	ERCC3	rs4150407	1.000	11/23/4	4/1/1	0.408	0.250	0.184	0.273	1.000	1.000
rs16869352	UBR5			36/2/0	6/0/0	0.026	0.000	1.000	1.000	1.000	1.000
rs175084	MLH3	rs175080	1.000	13/15/10	3/2/1	0.461	0.333	0.206	1.000	1.000	1.000
rs1799977	MLH1			18/18/2	4/1/1	0.289	0.250	0.459	0.273	1.000	1.000
rs3730872	LIG1	rs20579	0.732	26/12/0	6/0/0	0.158	0.000	0.563	1.000	1.000	1.000
rs3735721	RRM2B			36/2/0	6/0/0	0.026	0.000	1.000	1.000	1.000	1.000
rs3791767	ORMDL1	rs5742933	1.000	28/9/1	4/2/0	0.145	0.167	0.570	1.000	1.000	1.000
rs5893603	RRM2B			36/2/0	6/0/0	0.026	0.000	1.000	1.000	1.000	1.000
rs6151816	MSH3	rs6151792	0.806	28/9/1	5/1/0	0.145	0.083	0.570	1.000	1.000	1.000
rs61752302	UBR5			38/0/0	6/0/0	0.000	0.000	1.000	1.000	1.000	1.000
rs71636247	MSH3			32/6/0	6/0/0	0.079	0.000	1.000	1.000	1.000	1.000
rs72734283	MLH3			25/13/0	4/2/0	0.171	0.167	0.564	1.000	1.000	1.000
rs852151	EIF2AK1	rs12531179	0.811	27/11/0	3/3/0	0.145	0.250	1.000	1.000	1.000	1.000

The proxy SNP of interest, for which the sequenced SNP was selected is given, along with its LD with the sequenced SNP. CTRLs: controls; corr: Bonferroni corrected. LD: linkage disequilibrium; MAF: minor allele frequency; HWE: Hardy-Weinberg equilibrium; CTRLs: controls; corr: Bonferroni corrected.

(A) (B)

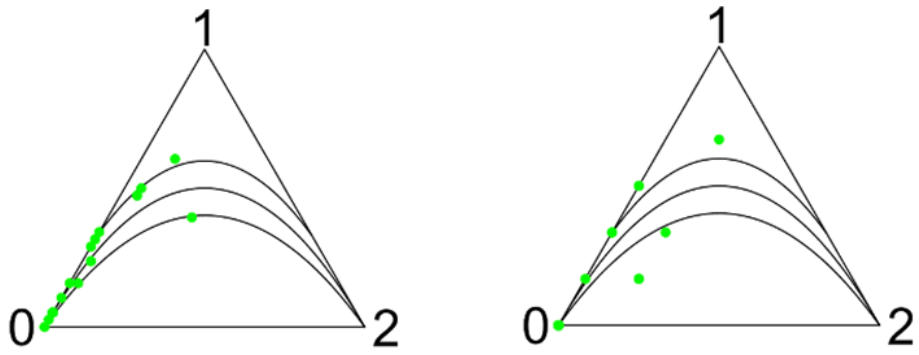


Figure 5.2. Ternary plots of HWE of the SNP genotypes

Plots for HD samples (A) and controls (B) are shown. Markers are coloured according to a chi-square test for HWE (red points are significant, green points are not significant). HWE parabola and acceptance region for a chi-square test are shown. 0: homozygous for wild type; 1: heterozygous for variant; 2: homozygous for variant.

5.2.3 SNP Associations with HD Age at Onset and Age at Death

The linear regression models were used to calculate estimated age at onset for each patient, based on their CAG repeat length, which was subsequently subtracted from their actual age at onset to give a residual value. The association of the SNPs with age at onset was tested by performing a linear regression of the residual values with the number of minor alleles. Additionally, the regression parameters from Bettencourt *et al*., 2016 were used in a parallel analysis due to the low power of this study. The association with age at onset across all 22 SNPs was tested by combining the association p values for each SNP using the following two methods; Brown's method (Brown et al., 1975), which combines non-independent, one-sided tests of significance, and the harmonic mean method (Wilson et al., 2019), which is based on Bayesian model averaging (Table 5.4). A similar analysis was performed with age at death, however, only the regression parameters from this report were used. One-sided p values were used for association in the same direction as that previously observed in genetic association studies (Ciosi et al., 2019; GEM-HD, 2019, 2015; Moss et al., 2017). The overall directionality of the associations was assessed by comparing the significance to that obtained from a similar analysis using two-sided p values. The p values resulting from the analysis were Bonferroni corrected. After Bonferroni correction, significant associations were observed in HD patients when the regression parameters of Bettencourt *at al*., 2016 were used (Brown p =1.99^{e-03} and harmonic mean p = 0.001). The undirected two-sided p values did not reach significance, which indicates the effect direction across these SNPs was concordant with previous genetic association studies (Ciosi et al., 2019; GEM-HD, 2019, 2015; Moss et al., 2017). In contrast, there was no significance of combined associated p values for all SNPs with age at death using the regression parameters in this report, which is due from the comparator studies using age at onset only.

Table 5.3. Combined p values for association of all 22 selected SNPs with HD age at onset and age at death

SNPs included	Condition	Brown's one-sided p	Brown's one-sided p (Bettencourt parameters)	Brown's two-sided p	Brown's two-sided p (Bettencourt parameters)	Harmonic one-sided p	Harmonic one-sided p (Bettencourt parameters)	Harmonic two-sided p	Harmonic two-sided p (Bettencourt parameters)
All SNPs	Age at onset	1.000	**1.99^{e-03}**	0.610	0.807	1.000	**0.001**	0.600	0.529
	Age at death	0.874		0.614		1.000		0.221	

Significant associations (p < 0.05) are in red and bold.

Associations of individual SNPs were subsequently tested with residual age at onset and age at death. Before Bonferroni correction, there were significant associations with earlier age at onset for *FAN1* (rs114136100 p = 0.048, rs150393409 p = 0.0048), *PMS2* (rs12534423 p = 0.024) and *ERCC3* (rs1566822 p = 0.016) in HD. However, none survived correction for comparison of all SNPs, most likely due to the small power of this study (Table 5.5). Before correction, there were significant associations with age at death for the following SNPs in *MSH3* (rs115109737 p = 0.018, rs6151816 p = 0.008, rs71636247 p = 0.018) and *LIG1* (rs3730872 p = 0.018), yet none survived correction for comparison of all SNPs (Table 5.6).

Table 4.5. Association of each SNP with age at onset

SNP	Gene	B	B (Bettencourt parameters)	Same direction as GWAS	p	p (Bettencourt parameters)	p bonf	p bonf (Bettencourt parameters)
rs1037699	RRM2B	0.085	0.113	No	0.500	0.432	1.000	1.000
rs1037700	RRM2B	0.085	0.113	No	0.500	0.432	1.000	1.000
rs11293	FAN1	0.010	0.008	Yes	0.799	0.844	1.000	1.000
rs114136100	FAN1	-0.152	-0.199	Yes	0.087	**0.048**	1.000	0.999
rs115109737	MSH3	0.037	0.087	Yes	0.543	0.206	1.000	1.000
rs12534423	PMS2	-0.144	-0.092	No	**0.024**	0.224	0.510	1.000
rs1382539	MSH3	0.005	0.010	No	0.902	0.815	1.000	1.000
rs146353869	HERC2P10	-0.079	-0.051	Yes	0.532	0.726	1.000	1.000
rs150393409	FAN1	-0.152	-0.199	Yes	0.087	**0.048**	1.000	0.999
rs1566822	ERCC3	0.084	0.074	Yes	**0.016**	0.072	0.339	1.000
rs16869352	UBR5	0.085	0.113	No	0.500	0.432	1.000	1.000
rs175084	MLH3	0.016	-0.003	No	0.574	0.940	1.000	1.000
rs1799977	MLH1	0.015	0.031	Yes	0.708	0.502	1.000	1.000
rs3730872	LIG1	-0.021	0.004	No	0.684	0.945	1.000	1.000
rs3735721	RRM2B	0.085	0.113	No	0.500	0.432	1.000	1.000
rs3791767	ORMDL1	-0.080	-0.098	Yes	0.100	0.077	1.000	1.000
rs5893603	RRM2B	0.085	0.113	Yes	0.500	0.432	1.000	1.000
rs6151816	MSH3	-0.082	-0.082	No	0.103	0.158	1.000	1.000
rs61752302	UBR5							
rs71636247	MSH3	0.111	0.142	No	0.090	0.055	1.000	1.000
rs72734283	MLH3	0.071	0.071	Yes	0.162	0.220	1.000	1.000
rs852151	EIF2AK1	-0.029	-0.047	No	0.587	0.441	1.000	1.000

Parallel analyses were run using parameters from this cohort or from Bettencourt et al., 2016. Significant associations ($p < 0.05$) are in red and bold. B: regression coefficient; p: two-sided p-value for association; bonf: Bonferroni correction.

Table 5.6. Association of each SNP with age at death

SNP	Gene	B	Same direction as GWAS*	p	p bonf
rs1037699	RRM2B	0.010	No	0.849	1.000
rs1037700	RRM2B	0.010	No	0.849	1.000
rs11293	FAN1	0.001	Yes	0.984	1.000
rs114136100	FAN1				
rs115109737	MSH3	-0.081	No	**0.018**	0.313
rs12534423	PMS2				
rs1382539	MSH3	0.037	No	0.244	1.000
rs146353869	HERC2P10				
rs150393409	FAN1				
rs1566822	ERCC3	-0.044	Yes	0.059	1.000
rs16869352	UBR5	0.010	No	0.849	1.000
rs175084	MLH3	0.011	No	0.588	1.000
rs1799977	MLH1	-0.015	Yes	0.534	1.000
rs3730872	LIG1	-0.081	Yes	**0.018**	0.313
rs3735721	RRM2B	0.010	No	0.849	1.000
rs3791767	ORMDL1	0.015	No	0.719	1.000
rs5893603	RRM2B	0.010	Yes	0.849	1.000
rs6151816	MSH3	-0.076	No	**0.008**	0.138
rs61752302	UBR5				
rs71636247	MSH3	-0.081	Yes	**0.018**	0.313
rs72734283	MLH3	0.009	Yes	0.788	1.000
rs852151	EIF2AK1	0.001	Yes	0.986	1.000

Significant associations ($p < 0.05$) are in red and bold. B: regression coefficient; *: same direction as age at onset GWAS; p: two-sided p-value for association; bonf: Bonferroni correction.

5.2.4 Polygenic Risk Scores

A polygenic age at onset and age at death score was derived to visualise the combined effect of the selected SNPs on residual age at onset and age at death, respectively (negative scores correspond to earlier onset) (Figure 5.3). For the polygenic age at onset score, this was defined by the sum of the number of minor alleles at each locus weighted by their effect size in this study using the Bettencourt *et al.*, 2016 regression parameters. Negative polygenic risk scores were associated with earlier age at onset in HD ($p = 4.9^{e-03}$), accounting for approximately 26% of variability in residual age at onset. Similarly, the polygenic age at death score was defined as the sum of the number of minor alleles at each locus weighted by their effect size in this study using the regression parameters defined for age at death. Although not significant ($p = 0.063$), negative polygenic risk scores were associated with earlier age at death, and accounted for approximately 37% of variability in residual age at death.

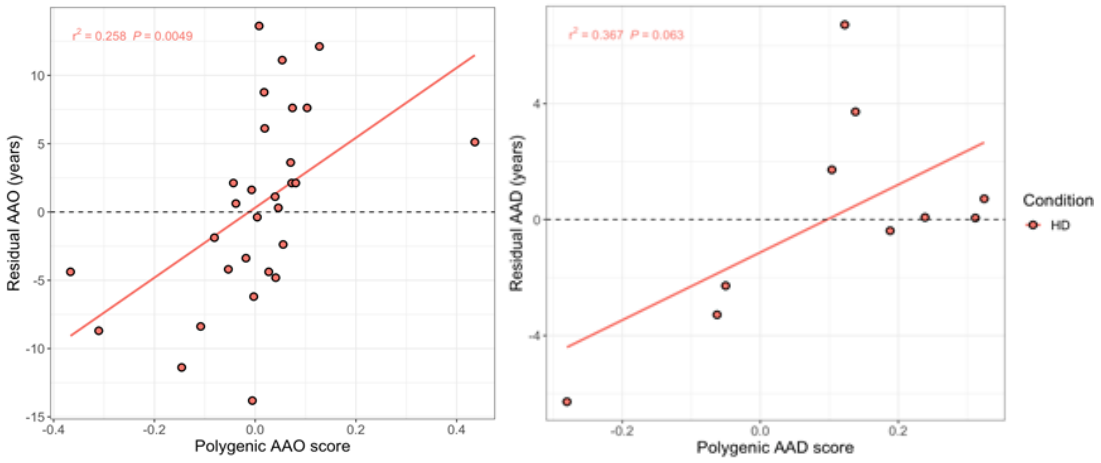


Figure 3.3. Polygenic scores for age at onset and age at death

AAO: age at onset; AAD: age at death.

Using somatic instability values obtained from ScaleHD for the HD samples with age at onset data, the polygenic age at onset score was plotted against the relative rate of somatic instability. The relative rate of somatic instability was determined by calculating the proportion of common variants relative to the mode of the progenitor allele using the following equation; n + (1 to 40 CAGs)/n, in which "n" is the CAG repeat size of the progenitor allele (Chapter 6, section 6.2.2). The polygenic age at onset score was inversely associated with somatic instability (p = 0.035). However, for individual tissues, this association was only significant for blood (p = 0.018), which showed that a more negative polygenic age at onset score was associated with less expansion. This is in comparison to the HD *post-mortem* brain regions, which indicate a trend for greater expansion associated with a more negative polygenic age at onset score (Figure 5.4). There was no association between the polygenic age at death score and somatic instability (p = 0.15).

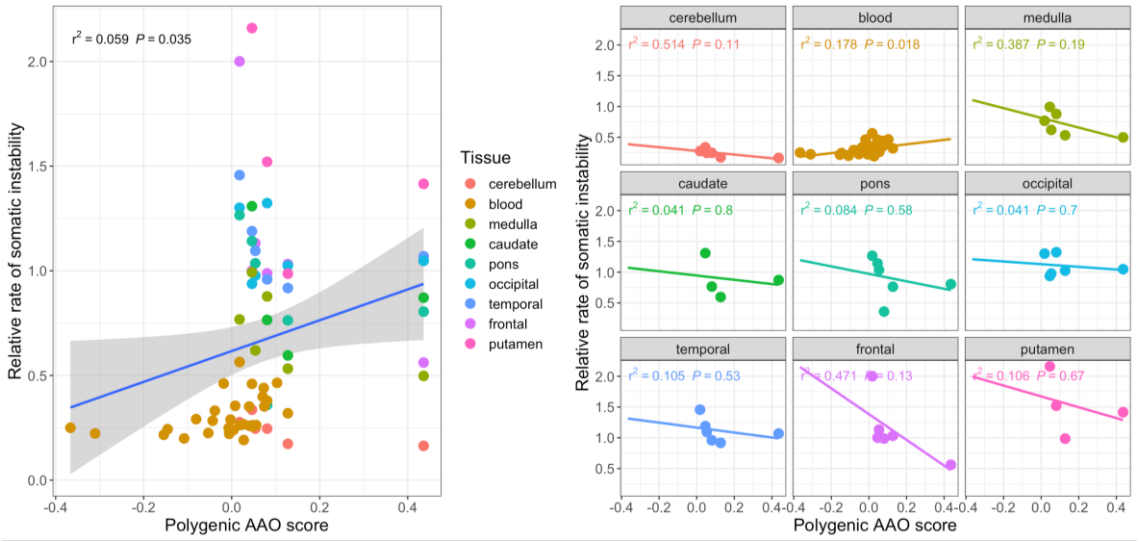


Figure 4.4. Effect of polygenic age at onset score on the relative rate of somatic instability

AAO: age at onset.

5.3 Discussion

5.3.1 Age at Onset Variability

The CAG repeat length is the dominant predictor of HD age at onset and in the cohort presented in this report, 48.4% of the variability in age at onset was attributed to the CAG repeat length, with each CAG significantly advancing age at onset by 0.88 years. Additionally, the CAG repeat length accounted for 95.2% of variability in age at death, with each CAG advancing age at death by 0.93 years, however, the limited availability of age at death information should be taken into consideration. The CAG repeat length accounts for approximately 40% to 60% of variability in HD age at onset, which is represented in our cohort and highlights the residual variability of which a large portion is thought to be heritable (Djoussé et al., 2003). Therefore, our cohort was genotyped against the most implicated SNPs from the previous GWA studies that were identified as modifiers of the HD phenotype (Bettencourt et al., 2016; GEM-HD, 2019, 2015).

Assessing the overall effect of all 22 SNPs on the residual age at onset in our HD patient cohort revealed significant associations after Bonferroni correction when the Bettencourt *et al*., 2016 regression parameters were used. The analysis determined a lack of significance of undirected two-sided p values, which indicates that the effect direction of these SNPs mirrors that of the previous GWAS studies, thus replicating the previous GWAS results in an independent sample (Bettencourt et al., 2016; GEM-HD, 2019, 2015). However, there was no significant association with age at death. This is because the previous GWAS studies did not include patient age at death information, and therefore, the Bettencourt *et al*., 2016 regression parameters were not used.

A subsequent analysis in our HD patient cohort to examine the association of individual SNPs with residual age at onset and age at death did not reveal any significant associations after Bonferroni correction. This is most likely due to the low power of the study, which is 15-fold less than that of the HD cohort in Bettencourt *et al*., 2016. Similarly, although the GWAS analysis performed by Bettencourt *et al*., 2016 in the polyglutamine diseases including HD, SCA1, SCA2, SCA3, SCA6, SCA7, and SCA17 identified modifier loci in two DNA repair genes, *FAN1* (rs3512) and *PMS2* (rs1805323), it did not replicate the most significant SNP in *FAN1* (rs146353869) identified by GEM-HD, 2015 (Bettencourt et al., 2016). This was concluded to be due to the reduced power of the study. Therefore, as the cohort presented here is but a fraction of the HD cohort analysed in Bettencourt *et al*., 2016, it is not surprising that the most significant modifying SNPs are not replicated

in this report. However, before Bonferroni correction, there were significant associations with earlier age at onset for SNPs in *FAN1* (rs114136100 and rs150393409), *PMS2* (rs12534423, proxy SNP of rs1805323) and *ERCC3* (rs1566822, proxy SNP of 4150407). Additionally, significant associations with age at death were identified before Bonferroni correction for the following SNPs in *MSH3* (rs115109737, rs6151816, proxy SNP of rs6151792, and rs71636247), and *LIG1* (rs3730872, proxy SNP of rs20579). Of these SNPs identified prior to Bonferroni correction, it is interesting that rs114136100 and rs150393409 are the 25th and 1st ranked top coding SNPs from the analysis of GEM-HD, 2019, 2015 and Moss *et al*., 2017, respectively ($p = 1.976191^{e-23}$, $p = 1.754193^{e-28}$).

The significant association with earlier age at onset for the rs150393409 SNP in *FAN1* was determined to have an onset hastening effect of 5.2 years in the latest GWAS (GEM-HD, 2019). This SNP was reported as the most significant coding SNP in GEM-HD, 2019, 2015. Additionally, a recent report has highlighted the protective effect of *FAN1* on CAG repeat instability in HD as reduced function results in hastened onset and increased expression leads to delayed onset (Goold et al., 2018). Specifically, the lowering of *FAN1* expression in mammalian cells and HD-patient derived induced pluripotent stem cells increased CAG repeat expansions (Goold et al., 2018). In addition, the inactivation of *Fan1* in a mouse model of Fragile X syndrome induced the somatic expansion of the CGG repeat, which suggests that the impact of *FAN1* variation can extend to other trinucleotide diseases (Zhao and Usdin, 2018).

PMS2 encodes the PMS2 protein, which is a subunit of the MutLα complex (MLH1-PMS2) involved in mismatch repair and has previously been identified as a genetic enhancer of the CTG expansion in a mouse model of DM1 (Gomes-Pereira, 2004). Somatic expansion of the CTG repeat was reduced in *Pms2* null DM1 mice. Additionally, the previous and recent GWAS in human HD and polyglutamine SCA patients have identified the function of the mismatch repair components, MSH3, MLH1, MLH3 and PMS2, as genetic modifiers of age at onset (Bettencourt et al., 2016; GEM-HD, 2019). *ERCC3* (ERCC excision repair 3) is translated into the XPB protein, which is an essential subunit of the transcription factor IIH (TFIIH) complex. The two major functions of the TFIIH complex are gene transcription and repairing damaged DNA by nucleotide excision repair (Oh et al., 2007). *FAN1*, *PMS2* and *ERCC3* all play a role in DNA repair, which reinforces that genetic variation in DNA repair pathways plays a role in modifying the HD phenotype.

5.3.2 Polygenic Risk Scores and Somatic Instability

A polygenic risk score was derived to examine the combined effect of the SNPs on residual age at onset and age at death in our HD cohort. The polygenic risk score is a number based on variation in multiple genetic loci and their associated weights (regression analysis), which reflects the individuals inherited susceptibility to these SNPs. The polygenic risk scores in this study were calculated by adding risk alleles weighted by their effect size in this study and the regression parameters derived from Bettencourt *et al*., 2016 for the age at onset analysis and the regression parameters from this report for the age at death analysis. Overall, negative polygenic scores were associated with an earlier age at onset and age at death in the HD cohort presented here. This is in concordance with previous studies which reported a positive correlation between the residual age at onset and increasing polygenic age at onset score that accounted for a small percentage of variance in the residual age at onset (Bettencourt et al., 2016).

In examining the relationship between somatic instability and the polygenic age at onset score, the effect was shown to differ between tissues. A significant association was determined between a more negative polygenic age at onset score and decreased somatic instability in the blood. In contrast, greater somatic instability in the brainstem, putamen and frontal lobe is associated with a more negative polygenic risk score (Chapter 6, Figure 6.2). Although non-significant, it is interesting that the brain regions, which are primarily affected and have the greatest somatic instability, are associated with a negative polygenic age at onset score, which in turn is associated with an earlier age at onset. We show that the combined effect of inheriting the minor alleles associated with the 22 selected SNPs in this study is associated with an earlier age at onset that is linked to increased somatic instability in the most affected brain regions. This reinforces previous and recent studies supporting the role of somatic instability in modifying disease onset and progression (Flower et al., 2019; Kennedy, 2003; Shelbourne et al., 2007; Swami et al., 2009). The results presented here further cement the relationship between DNA repair genes and somatic instability as modifiers of HD.

Chapter 6. Somatic Mosaicism in Huntington's Disease *Post-mortem* Brains

6.1 Background

The CAG repeat length defines HD development and is the primary determinant for age at disease onset and severity (Andrew et al., 1993). However, the CAG repeat is somatically unstable, progressively increasing in size over time. This phenomenon has been reported to be exacerbated in the brain regions specifically vulnerable to HD pathogenesis, namely the striatum and the cerebral cortex (Telenius et al., 1994; Kennedy, 2003). Due to the tissue-specificity of somatic mosaicism seen in HD, it raises the hypothesis that repeat instability itself contributes to HD pathogenesis. Previous work by Swami et al., 2009 supports this; forty-eight HD *post-mortem* brains were divided into early and late onset groups to examine the relative contribution of somatic mosaicism to HD pathogenesis (Swami et al., 2009). Small pool PCR (SP-PCR) of the frontal cortex determined that early onset *post-mortem* brains contained a greater average maximum expansion (mean 42 CAGs) compared to late onset (mean 29 CAGs). Somatic instability was additionally determined using skewness, a measurement of the degree of symmetry of a distribution. The results indicated that early onset *post-mortem* brains displayed a greater right skewness, highlighting the bias towards further expansion. The degree of skewness was also shown to have a negative association with the residual onset age, substantiating the link between greater somatic expansion and earlier disease onset (Swami et al., 2009). To further understand the relative contribution of small and large CAG repeat length changes to HD pathogenesis, the somatic mosaicism profile in six HD *post-mortem* brains of varying inherited CAG length and phenotype were analysed by Illumina MiSeq sequencing and SP-PCR, respectively (Table 6.1). These two methods contrastingly quantify somatic mosaicism. Illumina MiSeq quantifies small changes in repeat size by calculating the proportion of common variants (1 to 40 CAGs greater than the mode) relative to the mode of the progenitor allele, whereas SP-PCR quantifies the mutational load by depicting the presence of large CAG repeat sizes in single molecules of DNA.

Table 6.1. HD patient *post-mortem* brains

HD *Post-mortem* brain	PMI (hours)	$(CAG)_n$	Disease duration (years)	Age at onset		
				Actual	Estimated (Langbehn)	Residual
P40.97	48	41	12	65	57.5	7.5
P2.03	74	43	18	47	47.5	-0.5
P72.10	37	42	15	55	52.5	2.5
P3.92	10	41	3	71	57.5	13.5
P7.96	48	41	< 1	72	57.5	14.5
P28.98	96	44	9	50	42.5	7.5

PMI: *post-mortem* index; $(CAG)_n$: CAG repeat of length n; (Langbehn): estimated mean age at onset according to Langbehn *et al.*, 2004; Residual age at onset: actual age at onset minus estimated mean age at onset (Langbehn).

6.2 Results

6.2.1 HD *Post-mortem* Brain Macroscopy and Microscopy Reports

To encapsulate brain regions with a range of CAG repeat instability levels, the HD *post-mortem* brains were divided into the following regions; frontal lobe, temporal lobe, occipital lobe, putamen, caudate nucleus, cerebellum, pons, and medulla. The macroscopy and microscopy reports from the HD *post-mortem* brains were obtained to investigate the environment and state of the tissue (Table 6.2 and Table 6.3). The macroscopic and histological appearance of the *post-mortem* brains were all consistent with the clinical diagnosis of HD. However, in some patients an additional diagnosis was present, including pathological aging (P2.03), frontotemporal contusions with subarachnoid haemorrhage (P72.10), and grade IV astrocytoma (P3.92). With the exception of P3.92 and P7.96, all patients had disease durations from 9 to 18 years. Patients P3.92 and P7.96 had an extreme late onset in their seventies with a disease duration of 3 years and < 1 year, respectively. Clinical notes reveal a previous history of hypomanic schizoaffective disorder in patient P3.92 and senile onset chorea in patient P7.96. *Post-mortem* brain P3.92 was sent to Prof Vonsattel, as although the striatum appeared atrophic the brain was distorted due to the diffuse infiltration of the striatum by a malignant astrocytoma, who confirmed the HD diagnosis based on brain morphology. Only macroscopic appearances were available for P7.96 due to the freeze-thaw artefact, yet striatal atrophy consistent with HD was evident.

Table 6.2. Summary of the macroscopy reports for the HD *post-mortem* brains

Macroscopy Summary

HD *post-mortem* Brain	Caudate nucleus	Putamen	Globus pallidus	Cerebral cortex	Cerebellum	Brain stem
P40.97	Highly atrophic	Atrophic	Atrophic	Moderate dilation of lateral ventricle	Normal	Artefact in lower brain stem
P2.03*	Reduced in bulk, atrophic		Slightly discoloured	Mild dilation of right frontal horn	Normal	Normal
P72.10†	Reduced in bulk	Normal bulk	Normal	Swollen left cerebral hemisphere, extensive subarachnoid haemorrhage of frontal, temporal and parietal lobes	Subarachnoid haemorrhage covering the vermis	
P3.92	Atrophic, although difficult to tell due to distortion from the tumour	Tumour impinges	Tumour impinges	Tumour infiltration within the temporal pole showing haemorrhage, necrosis and cystic degeneration, compressed right lateral ventricle	Normal	
P7.96‡	Atrophic	Atrophic		Atrophic, dilation of lateral ventricle	Atrophic	Atrophic
P28.98	Atrophic	Atrophic	Normal	Moderate dilation of the lateral ventricle	Normal	Normal

*: reported as Vonsattel grade 3; †: This HD *post-mortem* brain had a frontal and temporo-parietal subarachnoid haemorrhage, extensive fronto-basal (orbital surface) contusions, milder involvement of the temporal lobe, mild reduction in bulk of the caudate nucleus and minimal pallor of the locus coeruleus; ‡: *Post-mortem* brain received by the UCL Queen Square Brain Bank already cut mid-sagittally with both halves frozen resulting in severe *post-mortem* freeze artefact.

Table 6.3. Summary of the microscopy reports for the HD *post-mortem* brains

Microscopy Summary

HD *post-mortem* brain	Caudate nucleus	Putamen	Globus pallidus	Cerebral cortex	Cerebellum	Brain stem
P40.97	Neuronal loss, astrocytosis	Nerve cells remain, moderate gliosis	Nerve cells remain, moderate gliosis	No definite neuronal depletion, increased gliosis (frontal).	Slight depletion of Purkinje cells	Midbrain and pons are normal
P2.03*	Atrophic, neuronal loss, spongiosis, gliosis and presence of intranuclear inclusions	Presence of intranuclear inclusions			Moderate Purkinje cell loss	
P72.10†	Reduced in bulk, astrogliosis, presence of intranuclear inclusions	Presence of intranuclear inclusions		subarachnoid haemorrhage, contusions, presence of intranuclear inclusions	Moderate Purkinje cell loss	
P3.92	Increased cellularity and gliosis	Increased cellularity, increased gliosis		Architecture of temporal lobe is destroyed by Grade IV malignant astrocytoma	Slight depletion of Purkinje cells	Slight increased gliosis in the medulla
P7.96‡	Artefact	Artefact	Artefact	Artefact	Artefact	Artefact
P28.98	Severely atrophic, reactive astrocytes	Severely atrophic, reactive astrocytes	Reduced striato-pallidal fibres	Mild atrophy and gliosis, minor involvement	Normal	Normal

*: reported as Vonsattel grade 3; †: This HD *post-mortem* brain had a frontal and temporo-parietal subarachnoid haemorrhage, extensive fronto-basal (orbital surface) contusions, milder involvement of the temporal lobe, mild reduction in bulk of the caudate nucleus and minimal pallor of the locus coeruleus; ‡: *Post-mortem* brain received by the UCL Queen Square Brain Bank already cut mid-sagittally with both halves frozen resulting in severe *post-mortem* freeze artefact.

6.2.2 The Somatic Mosaicism Profile in the HD *Post-mortem* Brains

To investigate the profile of somatic mosaicism regionally in the HD *post-mortem* brains, we initially performed clone sequencing on DNA extracted from each brain region, which yielded 156 clones containing the *HTT* sequence. *Supplementary Data Figure 2* depicts the sequence configurations resolved per region in each HD *post-mortem* brain. A summary of the text sequence data can be found in *Supplementary Data Table 3*. However, as clone sequencing was not successful for all available regions per HD *post-mortem* brain and due to the variation in number of clones obtained per region, it was not possible to perform somatic mosaicism analysis.

DNA extracted from the HD *post-mortem* brain regions and from their corresponding blood samples were sequenced by Illumina MiSeq, which informed on the size and sequence of the CAG and CCG repeats as well as determining the profile of somatic mosaicism. Illumina MiSeq was successful for all samples (Table 6.4). No atypical sequences were identified and the modal $(CAG)_n$ of each *post-mortem* brain was consistent between all brain regions and the blood, which is comparable to the sizing results from previous fragment analysis (Chapter 2, section 2.2.2, Table 2.3). The confidence of each genotype reported per patient was calculated by ScaleHD (Table 6.5). Each allele originates with 100% confidence, which subsequently reduces if certain data characteristics (penalties) are encountered during the genotyping process (https://scalehd.readthedocs.io/en/latest/Definitions.html) (Chapter 2, section 2.10, Table 2.7). Such penalties include low peak thresholds, rare characteristics (homozygous haplotypes), atypical alleles, and total read count. Scores of 60% and above are considered to give reliable genotypes, whereas values below 60% require manual inspection. All of the expanded alleles in the HD *post-mortem* brains had confidence scores of above 60%, with the exception of the cerebellum in P40.97, P2.03 and P3.92, which had the lowest confidence at 59%, 34% and 58%, respectively. Manual inspection of these samples confirmed the genotype (*inspection performed by Dr Marc Ciosi)*.

Table 6.4. Illumina MiSeq sequencing results of the HD *post-mortem* brain and blood samples

HD post-mortem brain	Allele1 (CAG)	Allele1 (CAACAG)	Allele1 (CCGCCA)	Allele1 (CCG)	Allele1 (CCT)	Allele2 (CAG)	Allele2 (CAACAG)	Allele2 (CCGCCA)	Allele2 (CCG)	Allele2 (CCT)	Wild-type allele (allele 1)	Expanded allele (allele 2)
P40.97 Frontal Lobe	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P40.97 Temporal Lobe	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P40.97 Occipital Lobe	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P40.97 Putamen	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P40.97 Caudate Nucleus	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P40.97 Cerebellum	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P40.97 Pons	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P40.97 Medulla	18	1	1	10	2	41	1	1	7	2	18_1_1_10_2	41_1_1_7_2
P2.03 Frontal Lobe	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Temporal Lobe	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Occipital Lobe	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Putamen	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Caudate Nucleus	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Cerebellum	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Pons	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Medulla	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P2.03 Blood	17	1	1	10	2	43	1	1	7	2	17_1_1_10_2	43_1_1_7_2
P72.10 Frontal Lobe	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Temporal Lobe	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Occipital Lobe	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Putamen	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Caudate Nucleus	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Cerebellum	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Pons	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Medulla	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P72.10 Blood	17	1	1	7	2	42	1	1	7	2	17_1_1_7_2	42_1_1_7_2
P3.92 Frontal Lobe	19	1	1	10	2	41	1	1	7	2	19_1_1_10_2	41_1_1_7_2
P3.92 Temporal Lobe	19	1	1	10	2	41	1	1	7	2	19_1_1_10_2	41_1_1_7_2
P3.92 Occipital Lobe	19	1	1	10	2	41	1	1	7	2	19_1_1_10_2	41_1_1_7_2
P3.92 Cerebellum	19	1	1	10	2	41	1	1	7	2	19_1_1_10_2	41_1_1_7_2
P3.92 Pons	19	1	1	10	2	41	1	1	7	2	19_1_1_10_2	41_1_1_7_2
P3.92 Medulla	19	1	1	10	2	41	1	1	7	2	19_1_1_10_2	41_1_1_7_2
P7.96 Frontal Lobe	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Temporal Lobe	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Occipital Lobe	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Putamen	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Caudate Nucleus	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Cerebellum	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Pons	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Medulla	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P7.96 Blood	17	1	1	7	2	41	1	1	7	2	17_1_1_7_2	41_1_1_7_2
P28.98 Frontal Lobe	17	1	1	7	2	44	1	1	7	2	17_1_1_7_2	44_1_1_7_2
P28.98 Temporal Lobe	17	1	1	7	2	44	1	1	7	2	17_1_1_7_2	44_1_1_7_2
P28.98 Occipital Lobe	17	1	1	7	2	44	1	1	7	2	17_1_1_7_2	44_1_1_7_2
P28.98 Cerebellum	17	1	1	7	2	44	1	1	7	2	17_1_1_7_2	44_1_1_7_2
P28.98 Pons	17	1	1	7	2	44	1	1	7	2	17_1_1_7_2	44_1_1_7_2
P28.98 Medulla	17	1	1	7	2	44	1	1	7	2	17_1_1_7_2	44_1_1_7_2
P28.98 Blood	17	1	1	7	2	44	1	1	7	2	17_1_1_7_2	44_1_1_7_2

Table 6.5. Illumina MiSeq sequencing confidence results for HD *post-mortem* brains

***Post-mortem* brain**	**Confidence (%)**		***Post-mortem* brain**	**Confidence (%)**	
	WT	**EA**		**WT**	**EA**
P40.97 FRNT	100	78	P72.10 MED	100	69
P40.97 TEMP	100	63	P72.10 BLOOD	100	69
P40.97 OCCIP	100	100	P3.92 FRNT	100	100
P40.97 PUT	100	83	P3.92 TEMP	100	100
P40.97 CNUC	100	63	P3.92 OCCIP	100	63
P40.97 CBM	100	59	P3.92 CBM	100	58
P40.97 PONS	100	63	P3.92 PONS	100	78
P40.97 MED	100	63	P3.92 MED	100	63
P2.03 FRNT	100	78	P7.96 FRNT	100	69
P2.03 TEMP	100	100	P7.96 TEMP	100	60
P2.03 OCCIP	100	100	P7.96 OCCIP	100	69
P2.03 PUT	100	83	P7.96 PUT	100	65
P2.03 CNUC	100	63	P7.96 CNUC	100	69
P2.03 CBM	100	34	P7.96 CBM	100	65
P2.03 PONS	100	100	P7.96 PONS	100	69
P2.03 MED	100	63	P7.96 MED	100	69
P2.03 BLOOD	100	74	P7.96 BLOOD	100	69
P72.10 FRNT	100	64	P28.98 FRNT	100	64
P72.10 TEMP	100	65	P28.98 TEMP	100	64
P72.10 OCCIP	100	64	P28.98 OCCIP	100	100
P72.10 PUT	100	65	P28.98 CBM	100	65
P72.10 CNUC	100	69	P28.98 PONS	100	69
P72.10 CBM	100	65	P28.98 MED	100	69
P72.10 PONS	100	69	P28.98 BLOOD	100	69

WT: wild-type allele CAG repeat size; EA: expanded allele CAG repeat size; - : CAG repeat size undetermined; FRNT: frontal lobe; TEMP: temporal lobe; OCCIP: occipital lobe; PUT: putamen; CNUC: caudate nucleus; CBM: cerebellum; PONS: pons; MED: medulla.

Illumina MiSeq data analysis outputted repeat count distributions (*Supplementary Data Figure 3*) in which the quantification of somatic mosaicism in each tissue was determined by calculating the proportion of common variants relative to the mode of the progenitor allele using the following equation; n + (1 to 40 CAGs)/n, in which "n" is the CAG repeat size of the progenitor allele. The striatum in *post-mortem* brains P3.92 and P28.98 was not available and thus, are excluded from the below figures. With the exception of P7.96, the results indicate that the putamen contains the largest proportion of common variants, with the cerebellum, followed by the blood, containing the least (Figure 6.1). P7.96 has the greatest proportion of common variants in the frontal and occipital lobe, followed closely by the putamen. Due to *post-mortem* freezing of this brain, microscopic analysis could not be performed and thus we cannot tell the extent of neurodegeneration, which could inform on the somatic mosaicism profile. In all four HD *post-mortem* brains, the putamen contains the largest proportion of common variants compared to the caudate nucleus, which could be due to the caudate nucleus being the primary site of HD neuropathogenesis. When the tissues are grouped together structurally; blood, brainstem (pons and medulla), cerebellum, cerebral cortex (frontal, temporal, and occipital lobe) and striatum (caudate nucleus and putamen), the results reveal that the striatum contains the largest proportion of common variants. In contrast, the cerebellum contains the lowest proportion of common variants (Figure 6.2).

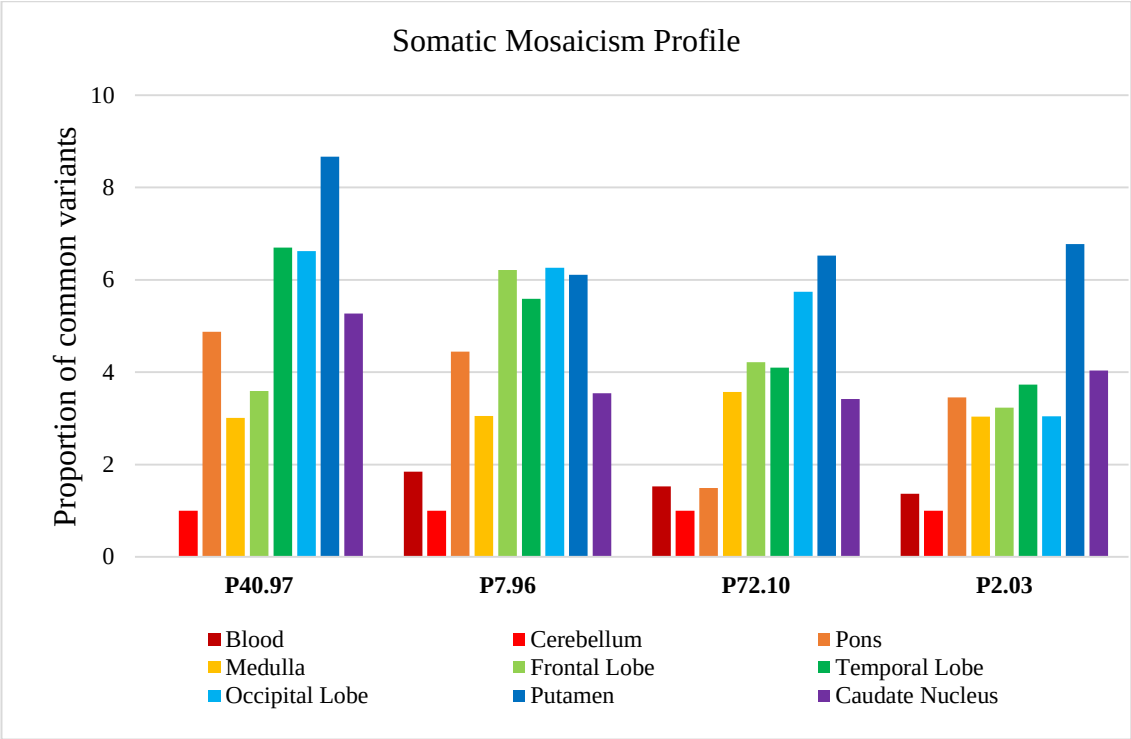


Figure 6.1. Somatic mosaicism profile in HD *post-mortem* brains and corresponding blood

The proportion of common variants is relative to the cerebellum, the most stable region, which is set at 1. The blood was unavailable for P40.97.

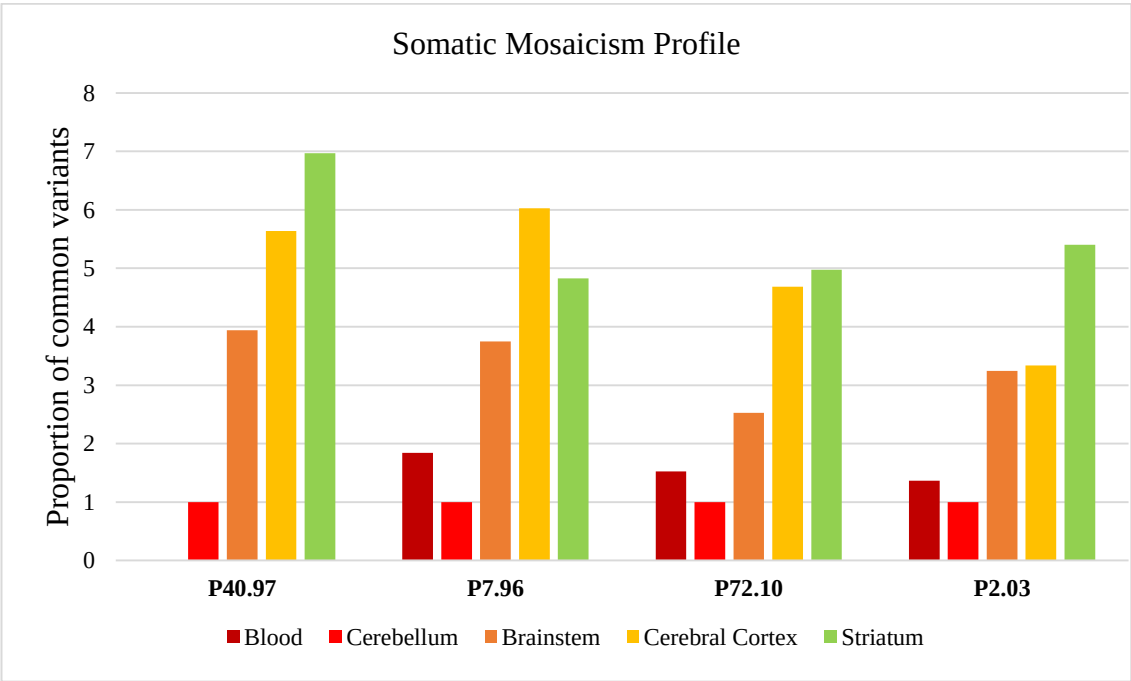


Figure 6.2. Somatic mosaicism profile in HD *post-mortem* brain structures and corresponding blood

The proportion of common variants is relative to the cerebellum, the most stable region, which is set at 1. The blood was unavailable for P40.97.

6.2.3 SP-PCR Analysis of HD *Post-mortem* Brain and Corresponding Blood Samples

To examine the mutational load between the CNS and non-CNS tissues, SP-PCR analysis was performed on all HD *post-mortem* brain regions and corresponding blood samples. To calculate the number of diploid genomes in 1, 0.25, and 0.05 ng/µL of DNA, which were the serial dilutions used in the SP-PCR, the following values were used; Avogadro's number (6×10^{23}), base pair mass (660 g) and the number of bases in a haploid genome (3×10^{9}). The molecular weight of a diploid genome was calculated as follows; (660 g x 3×10^{9})*2 = 4×10^{12} g. Therefore, there are 6×10^{23} diploid genomes in 4×10^{12} g and the number of diploid genomes in 1, 0.25, and 0.05 ng/µL of DNA is 1.5×10^{2}, 37.5, and 7.5, respectively. Repeat length variation was evident regionally within each *post-mortem* brain and in the corresponding blood samples, which is summarised in Table 6.6 and displayed in Figure 6.3. In P40.97, the repeat within the occipital lobe and the putamen is extremely unstable, which is in contrast to both the caudate nucleus and the cerebellum, where the repeat appears most stable. Repeat instability is also evident, but to a lesser degree, in the frontal and temporal lobe, pons and medulla, with the repeat reaching approximately 465 CAGs in the pons alone. In P2.03, the putamen, caudate nucleus and medulla appear to be more stable than the frontal, temporal and occipital lobe, which display the greatest level of repeat instability. In the pons, a band is present at 465 CAGs, and has a smear above which extends up to 620 CAGs. The frontal, temporal and occipital lobes present with bands up to 310 CAGs compared to the cerebellum and the blood, having mirroring patterns of repeat stability, with repeat sizes below 62 CAGs. P72.10 exhibits an alternative profile in that the putamen and caudate nucleus display the greatest repeat instability, followed by the temporal and occipital lobe, compared to the other brain regions. The cerebellum and the blood appear most stable. P7.96 displays a mosaic profile with the highest degree of instability present in the frontal, temporal and occipital lobe and the putamen. A lesser degree of instability is observed in the caudate nucleus, pons and medulla. In contrast, the cerebellum and blood are stable, even with the expansion reaching up to 256 CAGs. The profile of mosaicism in P3.92 revealed an unstable pattern within the frontal, temporal and occipital lobe, and in the pons. The cerebellum revealed the highest degree of repeat stability, which was followed by the medulla. In P28.98, extreme instability is observed in the occipital lobe, followed by the temporal and frontal lobe compared to all other regions. The cerebellum and blood display the most stable profiles.

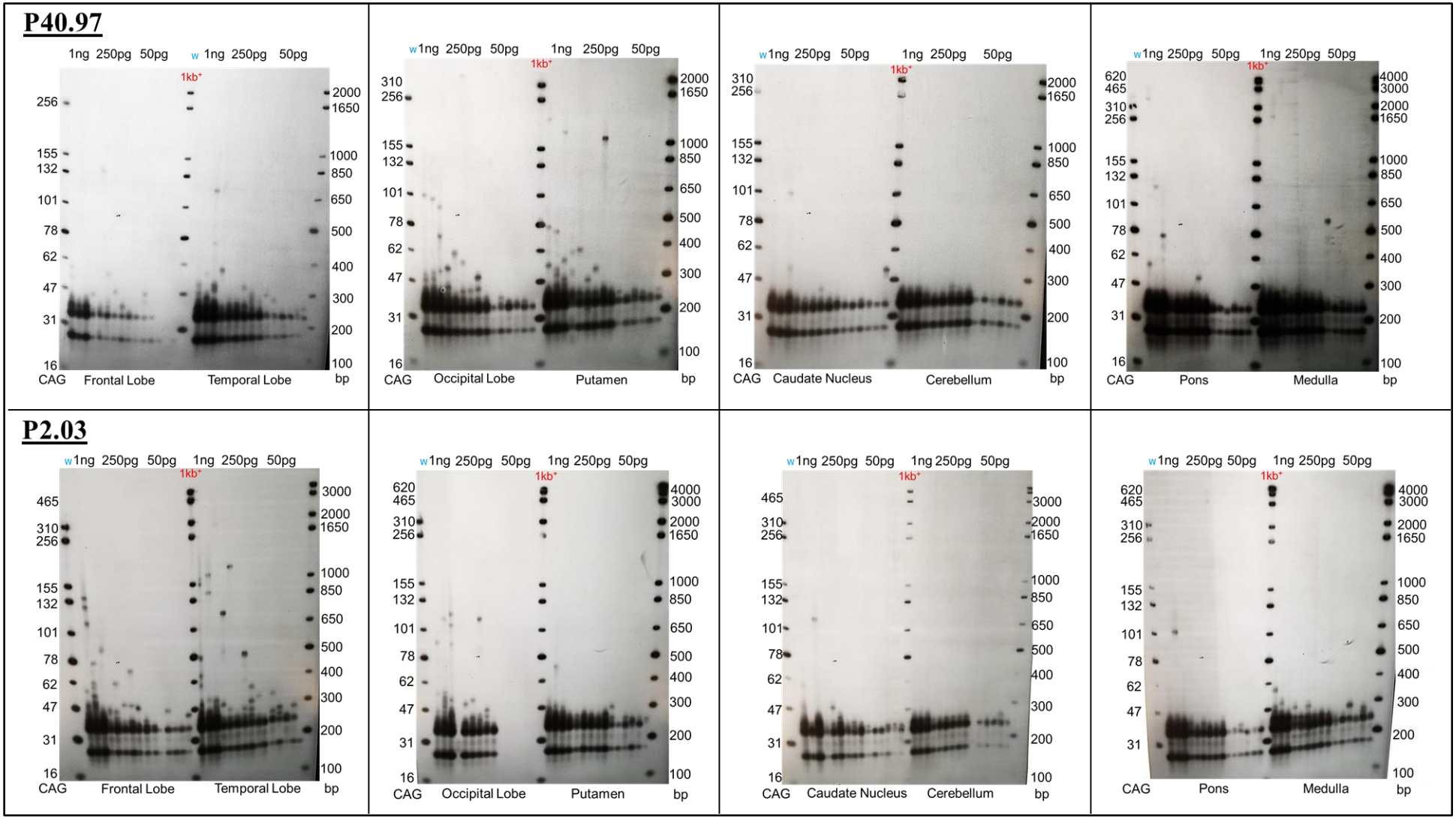
Table 6.6. Summary of SP-PCR determined somatic instability

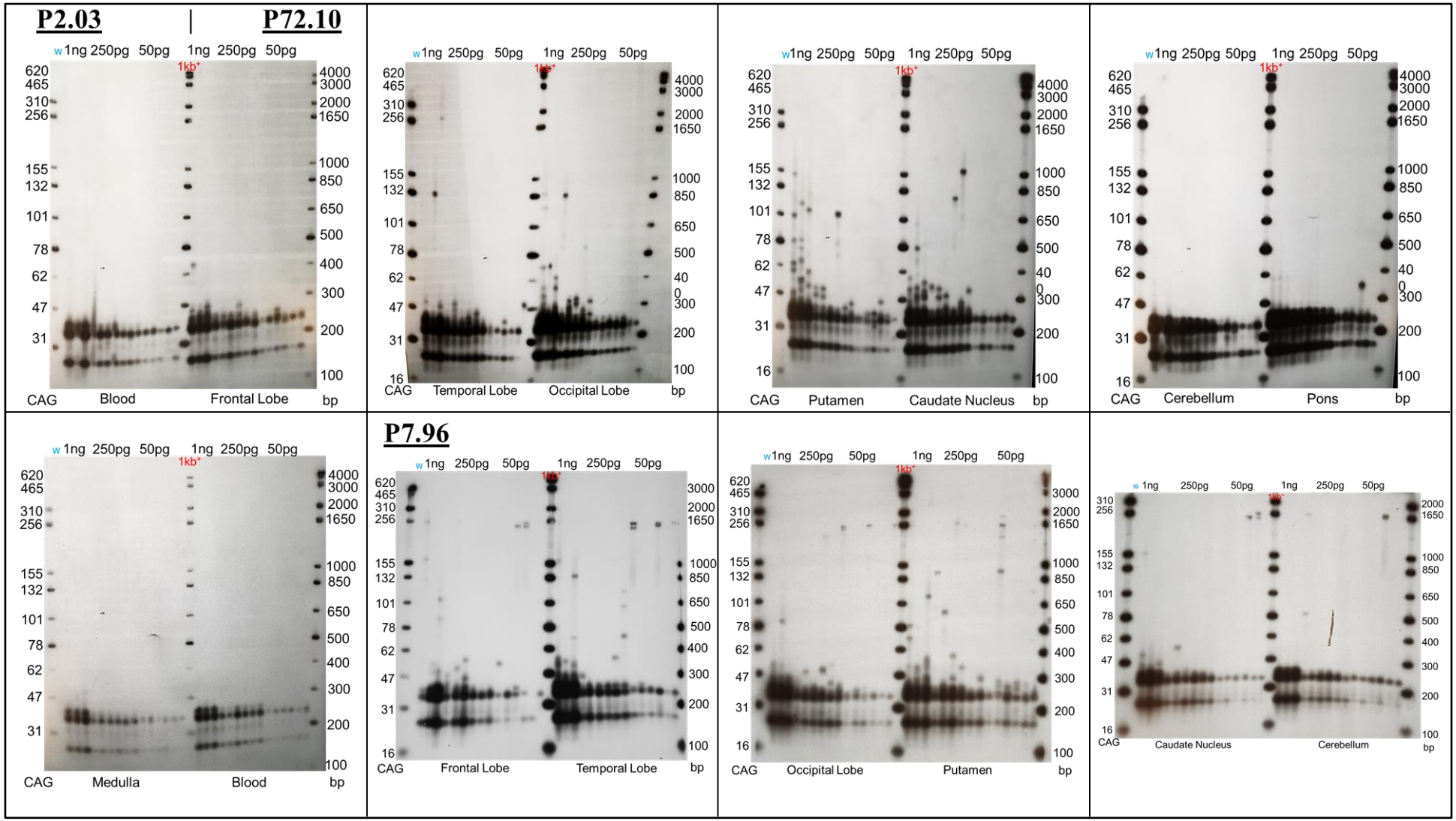
	P40.97	P2.03	P72.10	P7.96	P3.92	P28.98
Frontal Lobe	++	+++	++	+++	+++	+++
Temporal Lobe	++	+++	++	+++	+++	+++
Occipital Lobe	+++	+++	++	+++	+++	+++
Putamen	+++	++	+++	+++	-	-
Caudate Nucleus	+	++	+++	++	-	-
Cerebellum	+	+	+	+	+	+
Pons	++	++	++	++	+++	++
Medulla	++	++	+	++	+	++
Blood	-	+	+	+	-	+

+++: greatest instability; ++: moderate instability; +: stable; -: sample unavailable.

Figure 6.3. SP-PCR analysis of HD *post-mortem* brain and blood samples (overleaf)

The $(CAG)_n$ is labelled on the left and the base pair (bp) on the right according to the 1 kb^+ ladder. The first 3 lanes represent 1 ng of DNA, followed by 6 lanes each of 250 pg and 50 pg of DNA. Water (w) is used as a control to demonstrate that no PCR products are detected in the absence of a DNA template. SP-PCR analysis was conducted once, P3.92 medulla was replicated and confirmed reproducibility (*Supplementary Data Figure 4)*.


P40.97
1ng 250pg 50pg w 1ng 250pg 50pg
1kb+
256
155
132
101
78
62
47
31
16
2000
1650
1000
850
650
500
400
300
200
100
CAG Frontal Lobe Temporal Lobe bp
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
310
256
155
132
101
78
62
47
31
16
2000
1650
1000
850
650
500
400
300
200
100
CAG Occipital Lobe Putamen bp
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
310
256
155
132
101
78
62
47
31
16
2000
1650
1000
850
650
500
400
300
200
100
CAG Caudate Nucleus Cerebellum bp
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
620
465
310
256
155
132
101
78
62
47
31
16
4000
3000
2000
1650
1000
850
650
500
400
300
200
100
CAG Pons Medulla bp
P2.03
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
465
310
256
155
132
101
78
62
47
31
16
3000
2000
1650
1000
850
650
500
400
300
200
100
CAG Frontal Lobe Temporal Lobe bp
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
620
465
310
256
155
132
101
78
62
47
31
16
4000
3000
2000
1650
1000
850
650
500
400
300
200
100
CAG Occipital Lobe Putamen bp
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
465
310
256
155
132
101
78
62
47
31
16
3000
2000
1650
1000
850
650
500
400
300
200
100
CAG Caudate Nucleus Cerebellum bp
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
620
465
310
256
155
132
101
78
62
47
31
16
4000
3000
2000
1650
1000
850
650
500
400
300
200
100
CAG Pons Medulla bp


P2.03
P72.10
w 1ng 250pg 50pg 1ng 250pg 50pg
1kb+
620
465
310
256
155
132
101
78
62
47
31
16
4000
3000
2000
1650
1000
850
650
500
400
300
200
100
40
0
CAG
Blood
Frontal Lobe
bp
Temporal Lobe
Occipital Lobe
Putamen
Caudate Nucleus
Cerebellum
Pons
Medulla
P7.96
Frontal Lobe
Temporal Lobe
Occipital Lobe
Caudate Nucleus

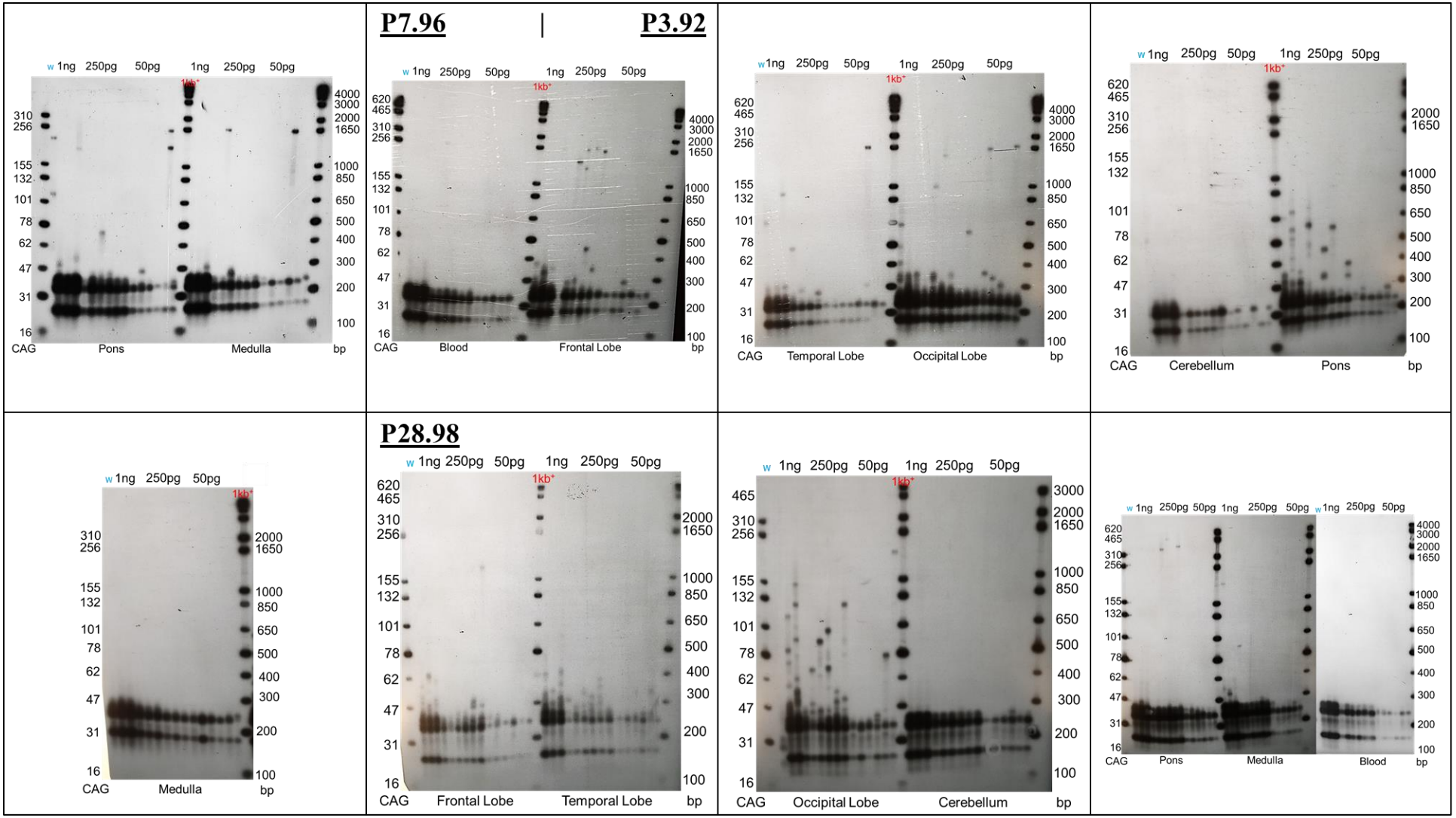


Figure 6.3. SP-PCR analysis of HD *post-mortem* brains and corresponding blood samples

6.2.4 PacBio SMRT Sequencing of HD *Post-mortem* Brains

The HD *post-mortem* brain samples containing the highest number of clones previously obtained from clone sequencing; P72.10 occipital lobe and medulla, P3.92 temporal lobe, cerebellum and pons, were prioritised for SMRT sequencing. P72.10 occipital lobe and P3.92 cerebellum were primarily sequenced on the RSII instrument. Repeat analysis was unavailable for P3.92 cerebellum, as there was an insufficient amount of reads. The SMRT sequencing repeat count results for P72.10 occipital lobe determined the modal (red dotted line) $(CAG)_n$ at approximately 46 CAGs, which is +4 CAGs greater than that determined by Illumina MiSeq at 42 CAGs (Figure 6.4 (A)). Additionally, the wild-type allele for P72.10 occipital lobe was not resolved, only the pathogenic allele was captured with multiple populations of 43, 44 and 50 CAGs. No sequence alterations were evident in the CAG or CCG repeats (Figure 6.4 (B)). The remaining samples, P72.10 medulla, P3.92 temporal lobe and pons, were subsequently sequenced on the Sequel System. No result was obtained as the samples did not survive the library preparation step. The DNA quality results determined that it was highly fragmented, to a varying degree, in all samples (Chapter 4, section 4.2.3, Figure 4.5).

Figure 6.4. Sequence configuration determined by Illumina MiSeq and SMRT sequencing (overleaf)

PacBio SMRT sequencing repeat analysis is shown as repeat count (A) and repeat number (B). MiSeq EA: expanded allele sequence determined by Illumina MiSeq; $(CAG)_n$: CAG repeat number including CAA trinucleotides; Fq: frequency.

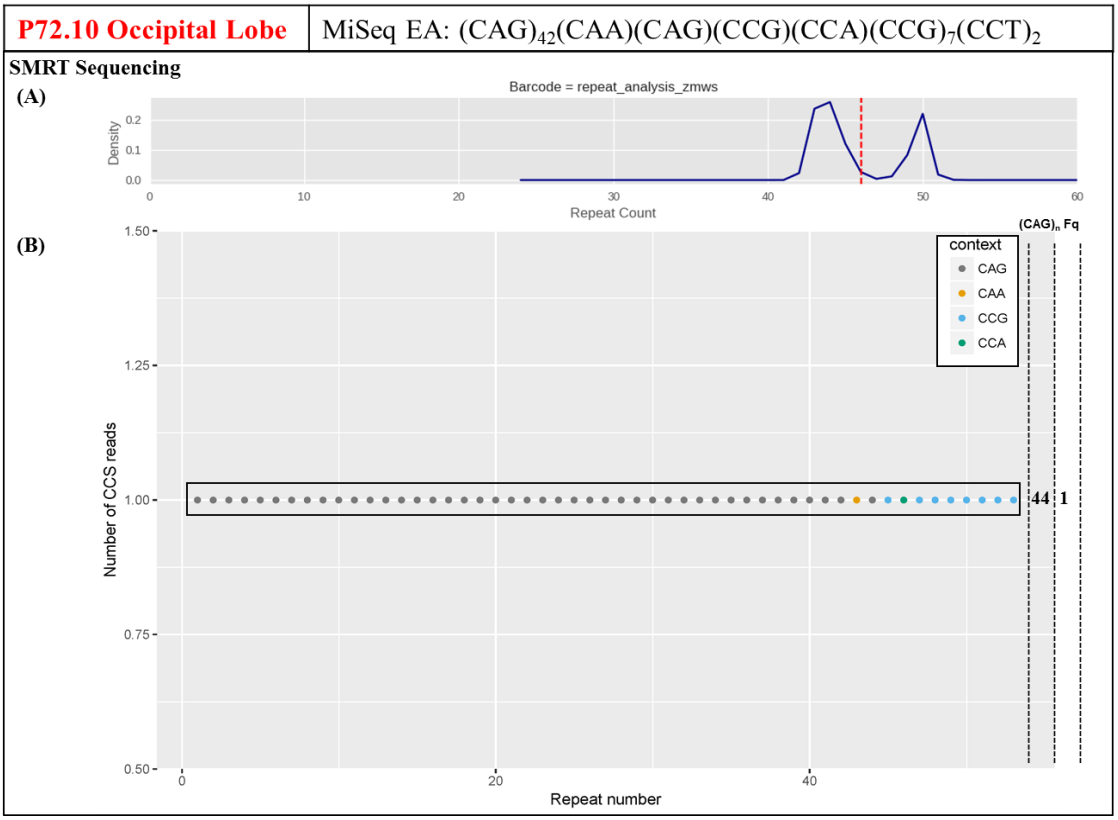


Figure 6.4. Sequence configuration determined by Illumina MiSeq and SMRT sequencing

6.2.5 Nanopore Sequencing of the HD *Post-mortem* Brains

DNA from two HD *post-mortem* brains was sent to Kings College London for Nanopore sequencing by Dr Graham Taylor. The raw data obtained from the Nanopore sequencing was extremely noisy (data not shown). In collaboration with David Murphy (*UCL Queen Square Genomics)*, the output files were run through a bioinformatics pipeline based on RepeatHMM to extract the CAG repeat sizes (Chapter 2, section 2.12) (Liu et al., 2017). The Nanopore $(CAG)_n$ sizing data is presented in graph format, which can be viewed in *Supplementary Data Figure 5* and summarised in Table 6.7 alongside the CAG repeat sizing results from fragment analysis and Illumina MiSeq. In comparing the Nanopore sequencing CAG repeat sizing results of the expanded allele in the HD *post-mortem* brains to those determined by fragment analysis, the $(CAG)_n$ differs by ± 2 CAGs, which is within the sensitivity bracket reported by fragment analysis. Similarly, the expanded allele CAG repeat size determined by Nanopore sequencing differs by ± 1 CAG compared to the Illumina MiSeq CAG repeat sizing results. Control *post-mortem* brains were analysed to ensure that Nanopore sequencing resolved the correct alleles and could distinguish between similarly sized CAG repeats. There is a size discrepancy of ± 4 CAGs in comparison to fragment analysis, which is outside of the sensitivity bracket.

In order to determine the fidelity of the CAG repeat sizing results from the Nanopore sequencing data, the total number of reads from 0 to approximately 80 CAGs were calculated per sample in which the modal wild-type and expanded allele read counts were determined as a percentage of total reads (Table 6.8). The threshold of 80 CAGs was set due to the lack of reads thereafter. P72.10 medulla presents with the lowest percentage of read counts for the expanded allele with 4.9% of total reads mapping to 42 CAGs. P3.92 temporal lobe presents with the highest percentage of read counts for the expanded allele with 7.7% of total reads mapping to 41 CAGs. Read count percentages for the wild-type alleles were higher in the control *post-mortem* brains compared to the HD *post-mortem* brains. Somatic mosaicism analysis was attempted in P72.10 using the same equation to calculate the proportion of common variants, n + (1 to 40 CAGs)/n, in which "n" is the CAG repeat size of the progenitor allele. However, due to inconsistent read counts between regions, somatic mosaicism analysis was unsuccessful.

Table 6.7. $(CAG)_n$ sizing by Nanopore sequencing, fragment analysis, and Illumina MiSeq

***Post-mortem* Brain**	**Nanopore Sequencing**		**Fragment Analysis**		**Illumina MiSeq**	
HD	**WT**	**EA**	**WT**	**EA**	**WT**	**EA**
P72.10 FRNT	19	42	18	42	17	42
P72.10 TEMP	19	43	18	44	17	42
P72.10 OCCIP	19	42	18	44	17	42
P72.10 PUT	19	43	18	42	17	42
P72.10 CNUC	19	42	18	42	17	42
P72.10 CBM	19	42	18	42	17	42
P72.10 PONS	19	42	18	42	17	42
P72.10 MED	19	43	18	42	17	42
P72.10 BLOOD	19	42	18	42	17	42
P3.92 FRNT	21	42	20	41	19	41
P3.92 TEMP	21	42	20	41	19	41
P3.92 OCCIP	21	42	20	41	19	41
P3.92 CBM	-	-	20	41	19	41
P3.92 PONS	21	42	20	41	19	41
P3.92 MED	21	41	20	41	19	41
Control	**WT**	**WT**	**WT**	**WT**	**WT**	**WT**
P82.10 FRNT	18	19	18	22	-	-
P82.10 TEMP	22	23	18	22	-	-
P82.10 OCCIP	19	22	18	22	-	-
P82.10 CBM	18	19	18	22	-	-
P82.10 PONS	18	19	18	22	-	-

WT: wild-type allele CAG repeat size; EA: expanded allele CAG repeat size; - : CAG repeat size undetermined; FRNT: frontal lobe; TEMP: temporal lobe; OCCIP: occipital lobe; PUT: putamen; CNUC: caudate nucleus; CBM: cerebellum; PONS: pons; MED: medulla.

Table 6.8. Nanopore sequencing repeat count percentages for wild-type and expanded alleles of HD and control *post-mortem* brains

***Post-mortem* brain**	**Read count (%)**		***Post-mortem* brain**	**Read count (%)**	
HD	**WT**	**EA**	**HD**	**WT**	**EA**
P72.10 FRNT	14.4	6.1	P3.92 TEMP	9.3	7.7
P72.10 TEMP	16.2	5	P3.92 OCCIP	11.2	6.7
P72.10 OCCIP	13.1	6.6	P3.92 PONS	10.1	7
P72.10 PUT	15.7	5.6	P3.92 MED	12.2	6.6
P72.10 CNUC	16.4	5.6	**Control**	**WT**	**WT**
P72.10 CBM	15.8	6.4	P82.10 FRNT	13.7	17.6
P72.10 PONS	15.2	7.1	P82.10 TEMP	19.7	19.5
P72.10 MED	17.9	4.9	P82.10 OCCIP	13.5	15.1
P72.10 BLOOD	16.2	6.2	P82.10 CBM	17.4	22.8
P3.92 FRNT	11.7	6.4	P82.10 PONS	13.5	18

WT: wild-type allele CAG repeat size; EA: expanded allele CAG repeat size; FRNT: frontal lobe; TEMP: temporal lobe; OCCIP: occipital lobe; PUT: putamen; CNUC: caudate nucleus; CBM: cerebellum; PONS: pons; MED: medulla.

6.3 Discussion

6.3.1 Somatic Mosaicism Profiles Determined by Illumina MiSeq and SP-PCR

Previous studies imply that the profile of somatic mosaicism in HD initially correlated with the specific neuropathogenesis, with the highest levels of CAG repeat instability observed in the striatum and cortex (Telenius et al., 1994). Further investigation in HD *post-mortem* brains including 24 individuals with an early age at onset and 24 individuals with a late age at onset determined that various degrees of somatic instability were observed in the cortex of all *post-mortem* brains with a bias towards further expansion and that early onset individuals tended to have larger repeat expansions compared to late onset individuals (Swami et al., 2009). As the striatum displays little instability in end-stage disease, the frontal cortex was solely examined as it displays relatively high levels of instability. Additionally, experiments in a knock-in HD mouse model, *Hdh*Q^{111}, demonstrated that the CAG repeat length dependent phenotype is significantly delayed in mice that lack somatic instability, which suggests that somatic instability itself contributes to HD pathogenesis (Wheeler et al., 2003). Extreme striatal CAG repeat instability is only seen in HD *post-mortem* brains when the individual has died before symptom onset (Kennedy, 2003). In contrast, *post-mortem* brains that are analysed from HD patients after an average disease duration display extreme instability in the cortical regions and to a

lesser extent in the striatum (Kennedy, 2003). In this study, the somatic mosaicism profile of four HD *post-mortem* brains was determined by Illumina MiSeq and SP-PCR to elucidate the instability level using two methods quantifying both small and large changes in the CAG repeat length.

The application of the bioinformatics pipeline by Dr Marc Ciosi, ScaleHD, to the Illumina MiSeq sequencing data quantifies somatic mosaicism by calculating the proportion of common variants using the following equation; n + (1 to 40 CAGs)/n, in which "n" is the CAG repeat size of the progenitor allele. This method revealed that, with the exception of P.7.96, the greatest proportion of common variants in the HD *post-mortem* brains P40.97, P72.10, and P2.03, and corresponding blood samples, was in the putamen, whereas the cerebellum presented as most stable. When the brain regions were grouped together structurally, the cerebellum remained most stable, followed by the blood, brainstem and the cerebral cortex, with the striatum presenting with the highest proportion of common variants. The striatum has previously been examined as one structure when detailing the somatic mosaicism profile of both early- and end-stage HD *post-mortem* brains (Kennedy, 2003). In this study, we looked at both the putamen and caudate nucleus, which revealed that even though the striatum is the most mosaic structure, regionally the caudate nucleus is much less mosaic. One reason for this is the selective neuropathology of HD that is further emphasised in the neuropathological reports of these *post-mortem* brains. Therefore, it is likely that the reduced mosaic profile within the caudate nucleus is due to the high level of cell death in this primarily affected region.

The SP-PCR somatic mosaicism profiles reveal that in three out of the four HD *post-mortem* brains where the striatum is available (P40.97, P2.03, P7.96), the cortical regions display the greatest levels of CAG repeat instability. These results compliment the profile of instability previously determined in an end-stage HD patient with $(CAG)_{87}$ (Kennedy, 2003). Although we did not recapitulate the extreme cortical expansion size of > 700 CAGs, a similar fold change in $(CAG)_n$ relative to the progenitor allele was observed. In P40.97 occipital lobe, P2.03 temporal lobe and all cortical regions of P7.96, we observe $(CAG)_{310}$, $(CAG)_{310}$, and $(CAG)_{256}$, which is approximately 7.5-fold, 7-fold and 6-fold that of the progenitor allele, respectively. Interestingly, the pons and the medulla similarly displayed extreme repeat sizes, with the pons being more unstable than the medulla. In P40.97 there are bands present in the pons and medulla reaching to 465 CAGs and greater than 256 CAGs, respectively, which is 11.3-fold and 6.2-fold greater than the size of the progenitor allele. In P2.03 pons, a band is present at 465 CAGs, and has a smear above

which extends up to 620 CAGs, which represents a fold change of 10.8-fold to 14.4-fold greater than the size of the progenitor allele. P7.96 pons and medulla both contain bands at 256 CAGs, which is a 6.2-fold increase compared to the size of the progenitor allele. P3.92 pons is as unstable as cortical regions and P28.98 pons has a band at 465 CAGs, 10.6-fold greater than the progenitor allele CAG repeat size. As primarily unaffected regions in HD, the pons and medulla have not been examined in depth for their instability profile in previous studies. Aronin et al., 1995, reported the instability of the pons in the *post-mortem* brain of a juvenile HD patient to be due to the heterogeneous expression of mutant huntingtin (Aronin et al., 1995). The results reported here again highlight the instability of the pons, but in multiple HD *post-mortem* brains, which suggests that the somatic mosaicism pattern is tissue specific.

Compared to the cortical regions, repeat instability is lower in the striatum yet within the striatum, repeat instability is greater in the putamen than in the caudate nucleus, which is mirrored by the Illumina MiSeq results. The disparity of instability between the putamen and the caudate nucleus may be due to the caudate nucleus being primarily affected and thus, there are less cells to quantify the $(CAG)_n$ population. This is in keeping with previous literature on the neuropathogenesis of HD. Another hypothesis proposes that a higher mutational load in vulnerable cells may expedite downstream pathological events (Kennedy and Shelbourne, 2000). HD *post-mortem* brain P72.10 displayed an alternative pattern of somatic mosaicism in that the striatum had the highest level of repeat instability compared to the cortical regions. This is the profile expected in early-stage disease, and mirrors the mosaic pattern described in the aforementioned presumed premanifest HD individuals reported in Kennedy *et al*., 2003. Further examination into the clinical notes of P72.10 revealed that this individual stayed active and mobile for their 15-year disease course. This would suggest that the motor region is not as atrophic as the general neuropathology for HD patients with a dominant motor phenotype and therefore, repeat instability is highly visible.

In contrast to SP-PCR which quantifies the presence of extremely large CAG repeats, calculating the proportion of common variants mirrors that of the premanifest HD individuals reported in Kennedy *et al*., 2003, even though the *post-mortem* brains analysed here were from manifest HD patients. This raises the question of whether HD neuropathogenesis is driven by the rarer extreme CAG repeat lengths or the greater population of common variants, i.e., smaller repeat changes reaching only 40 CAGs more than the progenitor allele. As the proportion of common variants is largest in the striatum,

although carried by the putamen, over the cortical regions, this suggests that the most vulnerable cells can survive with the burden of the CAG repeat up to approximately 84 CAGs. However, SP-PCR analysis determined that the largest mutational load is in the cortical regions and not in the striatum. This potentially supports the hypothesis that cells with the largest CAG repeat sizes are primarily lost, hence why the caudate nucleus appears more stable than the putamen. Taken together, these profiles confirm that the somatic mosaicism profile mirrors that of HD neuropathogenesis and further suggests that the most vulnerable cells do not survive when carrying extremely large CAG repeat sizes.

6.3.2 Third Generation Sequencing Attempts in HD *Post-mortem* Brains; PacBio SMRT Sequencing and Nanopore Sequencing

Of the HD *post-mortem* brain regions sent for SMRT sequencing on the RSII system, P72.10 occipital lobe was the only successful sample and yielded 3 reads containing the *HTT* target region. This read count is too low to give confidence in the result. From the only sequence obtained, no interruptions or alterations were identified in P72.10 occipital lobe, which mirrors the configuration determined by Illumina MiSeq. Additionally, somatic mosaicism analysis could not be performed on 3 reads. Of the samples sequenced on the sequel system, no results were obtained. Although this system is built on the previously established SMRT technology and therefore most of the sequencing workflow is unchanged, the analysis method and new chemistry used for the Sequel System is not compatible on the RSII. Another modification is the addition of more restriction enzymes in the initial SMRTbell library preparation, which takes place before the CRISPR/Cas9 enrichment step and aids in target enrichment by reducing genome complexity. The restriction enzymes allow further reduction of the unwanted DNA and therefore prioritises loading of on-target molecules on the sequencing instrument. However, as our DNA was highly fragmented, this suggests that the use of more restriction enzymes was a factor in the samples failing the library preparation step. This further emphasises the requirement of high quality and quantity DNA samples for successful SMRT sequencing.

Two HD *post-mortem* brains were sent for Nanopore sequencing. In contrast to polymerase-mediated DNA synthesis, which is used by all major sequencing technologies, nanopore-based sequencing infers sequences from changes in the ionic current across a membrane when a single DNA molecule passes through a protein nanopore (Bowden et al., 2019). The sizing analysis used in this study is based on RepeatHMM, which has the potential to detect microsatellites from long read sequencing data and incorporates predefined models for well-known TNRs including *HTT*. It uses a

split-and-align strategy to improve alignments, perform error correction and peak calling algorithm based on the Gaussian mixture model to determine repeat counts (Liu et al., 2017). This tool allows the users to specify error parameters of the sequencing experiments. Our data was analysed using the default error correction established by Nanopore, as optimising the script surpassed the scope for this study. However, the read counts per $(CAG)_n$ were inconsistent between samples which hindered any somatic mosaicism analysis being performed. One factor that could be at play is the quality of the input DNA. Library preparation cannot be successfully achieved with sub-sufficient DNA quality, which may account for the samples where repeat counts could not be determined for the expanded allele (P3.92 cerebellum) and where the read counts are low, with one example being 7 reads for 42 CAGs in P72.10 occipital lobe compared to 564 reads for 43 CAGs in 72.10 temporal lobe (*Supplementary Data Figure 5*). This further highlights the importance of having sufficient DNA quality prior to library preparation.

In summary, Illumina MiSeq sequencing was the most efficient technology to not only elucidate the sequence configuration at the base pair level, but also to reveal the somatic instability profile throughout the HD patient *post-mortem* brains and corresponding blood samples. Illumina MiSeq determined the proportion of common variants per region, in which 1 to 40 CAGs greater than the progenitor allele was considered as a common variant. This quantification of somatic mosaicism revealed that the striatum contained the greatest level of instability and the cerebellum presented as stable. These results highlight a new pattern of somatic mosaicism that SP-PCR is unable to determine, and ultimately reinforces the previous hypothesis that the somatic mosaicism profile mirrors that of the specific neuropathology observed in HD. SP-PCR provides a quantitative measure of the extreme CAG repeat mutational load and identified CAG repeat sizes up to 14-fold that of their respective progenitor alleles within the HD *post-mortem* brains. These results reveal that in addition to the instability present in the striatal and cortical regions, the pons also contained high levels of CAG repeat instability. Furthermore, PacBio SMRT sequencing and Nanopore sequencing were able to size the CAG repeat with some confidence, however, instability profiles could not be determined due to low and inconsistent read counts. In order to gain full advantage of the third generation sequencing technologies, it is clear from the above results that optimisation is necessary for routine use in the laboratory.

Chapter 7. Discussion and Future Work

7.1 Thesis Overview

The work presented in this report was undertaken to investigate the role of modifiers in trinucleotide repeat diseases, specifically between patients with similarly sized pathogenic alleles. In brief, the presence and potential role of interruptions as disease modifiers in FRDA and HD were determined by long-range PCR with restriction enzyme digestion, TP-PCR, clone (Sanger) sequencing, and next- and third-generation sequencing technologies. The application of the three sequencing generations to our HD samples, allowed us to define the most suitable and efficient platform for our needs. Next generation sequencing by Illumina MiSeq accurately sized the CAG repeat in our cohort of HD patients with the greatest fidelity while also revealing the sequence composition at the base pair level. Subsequently, this method was used to validate the sequences obtained from clone sequencing and identified HD patients with a loss of interruption, which modified disease phenotype. As the loss of interruption phenotype did not account for all of the phenotypic variability observed, we examined other potential modifiers of HD including DNA repair pathway genes and somatic instability. The influence of DNA repair pathway genes in our HD cohort was investigated by genotyping for the DNA repair SNPs previously implicated as disease modifiers (GEM-HD, 2019). The results showed no significant association between individual SNPs and age at onset, which was concluded to be due to the small power of this study. Finally, the somatic mosaicism pattern in HD patient *post-mortem* brains was established to explore its relationship with the specific neuropathogenesis and the potential role of somatic instability as a disease modifier. Illumina MiSeq and SP-PCR were used to assess the relative contribution of small and large CAG repeat length changes to HD neuropathogenesis, respectively. Although contrasting methods, both showed a tissue specific profile of somatic instability. Illumina MiSeq determined that the cerebellum was the most stable region and that the striatum presented with the highest proportion of common variants. In contrast, SP-PCR determined the profile of somatic mosaicism to be greatest in the cortical regions, with the lowest level of instability recorded in the cerebellum, which is in accordance with the previous literature and suggests that the cells carrying the largest expansions are primarily lost (Kennedy, 2003). Overall, this report identifies CAG repeat sequence interruptions, somatic instability and DNA repair genes as disease modifiers and reinforces the hypothesis that cells carrying the largest CAG repeat lengths are preferentially lost in the regions of specific neuropathogenesis.

7.2 Friedreich's Ataxia

7.2.1 Main Findings

In a large cohort of FRDA patients, long-range PCR with *Mbo*II restriction enzyme digestion, which detects ≥ 50 bp of non-$(GAAGA)_n$ interruptions internally of the GAA repeat sequence, and TP-PCR, which is limited to approximately 100 GAAs at the 3' end of the expansion, were applied to investigate the presence of non-GAA repeat sequences. In contrast to TP-PCR, the *Mbo*II digestion method includes the entire expansion, however, neither method resolves the sequence of the repeat at the base pair level. The main findings reported here revealed that interruptions identified by TP-PCR were common, with 66% of the cohort carrying interruptions in the 3'end of the repeat. In contrast, *Mbo*II digestion analysis identified non-$(GAAGA)_n$ sequence interruptions in 13% of the cohort. Within this cohort, GAA1 repeat size accounts for up to 33.8% of the variation in age at onset. The potential effect of the interruption profile on GAA1 size and age at onset determined that in addition to GAA1 size, stronger correlations were identified for the pure MboII digestion (34.8%) and interrupted TP-PCR (35.9%) subset of patients. These results highlight that sequence purity has the potential to influence age at onset, however, without the base pair sequence composition, it is difficult to draw definitive conclusions on the influence of GAA repeat sequence interruptions on the FRDA phenotype.

7.3 Huntington's Disease

7.3.1 Main Findings

Clone sequencing was performed on a cohort of HD patients with similarly sized pathogenic alleles to determine the presence of potentially disease modifying repeat interruptions. These patients displayed phenotypic variability, evidenced by ages at onset that spanned 33 years from the earliest to latest presentation. The CAG repeat length is the dominant predictor of HD age at onset and in the cohort presented in this report, 48.4% of the variability in age at onset was attributed to the CAG repeat length, with each CAG significantly advancing age at onset by 0.88 years. We hypothesised that the remaining variability in age at onset could be attributed to alterations in the CAG and adjacent CCG repeats. In short and most notably, out of the 163 clones obtained overall for 23 HD patients, none contained any nonsynonymous sequence interruptions and all bar two patients (HD patient 1 and 4) carried the penultimate CAA interruption. Additionally, HD patient 1 had a loss of interruption in their CCG1 repeat tract, which was sequenced as a pure stretch of CCGs. During the clone sequencing of these patients, Dr Marc Ciosi developed a genotyping-by-sequencing protocol for *HTT* exon 1 using Illumina MiSeq (Ciosi et al., 2018). Due to the variation in number of clones and population of repeat sizes determined per patient, Illumina-MiSeq was performed on our sample cohort to validate the sequence configurations obtained and to further elucidate the sequence configuration of the remaining samples. This method highlighted four HD patients (patients 1, 2, 4, 22) with atypical expanded allele sequence configurations. HD patients 2 and 4 carried pure CAG repeat tracts and HD patients 1 and 22 carried pure CAG and CCG1 repeats. Consequently, HD patients 1, 2, 4, and 22 presented with age at onsets 7, 5, 2 and 9.5 years earlier than their estimated mean age at onset based on the modal $(CAG)_n$ determined by Illumina MiSeq, respectively. HD patients 1 and 22 presented with the earliest age at onsets, which highlights that the loss of interruption in both the CAG and CCG1 repeats has a stronger modifying effect than the loss of interruption in just the CAG repeat. In concordance with the current literature, these results highlight the influence of sequence interruptions, in this case, the loss of interruption, as a modifier of disease progression and reinforced what was found by clone sequencing in HD patients 1 and 4 (GEM-HD, 2019; Wright et al., 2019).

To further understand the contribution of CAG repeat length changes to HD pathogenesis, the somatic mosaicism profile in six HD *post-mortem* brains with varying progenitor alleles and phenotypes were analysed by Illumina MiSeq sequencing and SP-PCR.

Illumina MiSeq quantifies small changes in repeat size by calculating the proportion of common variants (1 to 40 CAGs greater than the mode) relative to the mode of the progenitor allele, whereas SP-PCR quantifies the mutational load by depicting the presence of large CAG repeat sizes in single molecules of DNA. In this report, Illumina MiSeq of the HD *post-mortem* brains determined that the putamen contained the largest proportion of common variants compared to the caudate nucleus, which could be due to the caudate nucleus being the primary site of HD neuropathogenesis. This was mirrored by the macroscopy and microscopy reports of the *post-mortem* brains. When the tissues were grouped by structure; blood, brainstem, cerebellum, cerebral cortex and striatum, the results revealed that the striatum contained the largest proportion of common variants, whereas the cerebellum contained the lowest. The SP-PCR somatic mosaicism profiles revealed that in the majority of the HD *post-mortem* brains, the cortical regions display the greatest levels of CAG repeat instability. These results complement the profile of instability previously determined in HD *post-mortem* brains (Kennedy, 2003).

To examine if DNA repair modifier loci were additionally attributable to the remaining variability in age at onset, the HD patient blood samples and the six HD *post-mortem* brains were genotyped on the NeuroChip array against a customised panel of SNPs associated with DNA repair pathway genes. Assessing the overall effect of all SNPs on the residual age at onset in our cohort revealed significant associations with an earlier age at onset, however, the association of each SNP singularly revealed no significant associations. Before Bonferroni correction, there were significant associations with earlier age at onset for *FAN1* (rs114136100, rs150393409), *PMS2* (rs12534423) and *ERCC3* (rs1566822), which implies a modifying role of these DNA repair SNPs in this cohort. Likewise, calculating a polygenic age at onset risk score revealed that negative polygenic risk scores were associated with earlier age at onset, accounting for approximately 26% of variability. The polygenic age at onset risk score was subsequently plotted against the relative rate of somatic instability, which showed a significant inverse association with somatic instability. However, for individual tissues, only a trend was observed for the association of greater somatic instability with a more negative polygenic age at onset score. Although not significant, we show that the combined effect of inheriting the minor alleles associated with all combined SNPs is associated with an earlier age at onset that is potentially linked to increased somatic instability. The results suggest that with a greater sample size, the modifying role of these DNA repair pathway genes could be determined in relation to disease progression and somatic instability.

7.3.2 CAG Repeat Interruptions

Greater than 95% of European ancestry HD chromosomes carry a canonical sequence of the polyglutamine repeat, which includes a penultimate synonymous interruption of the CAA trinucleotide. As CAA codes for glutamine, the length of the CAG repeat is consistently greater by two residues, compared to the length of the uninterrupted CAG repeat tract. Therefore, it has not previously been possible to interpret whether the correlation with age at onset is due to polyglutamine size or the pure CAG repeat size (GEM-HD, 2019). A recent GWAS in HD patients identified a signal on chromosome 4 near *HTT*, with two independent effects on age at onset; an onset-hastening effect of approximately 12.7 years and an onset-delaying effect of approximately 5.7 years (GEM-HD, 2019). It was hypothesised that these modifying effects could be associated with non-canonical sequence variations. Sequencing of the *HTT* polyglutamine repeat determined that the onset-hastening effect was associated with a loss of the CAA trinucleotide and conversely, the onset-delaying effect was associated with an additional CAA trinucleotide. It was further shown that age at onset estimation is more concise when based on the uninterrupted CAG repeat length, which indicates sequence interruptions as modifiers of HD phenotype (GEM-HD, 2019).

To recapitulate the use of clone sequencing in determining the polyglutamine repeat configuration, an additional study identified the loss of the CAA interruption in 16 symptomatic HD patients, which was represented by pure CAG and CCG repeat tracts (Wright et al., 2019). In a large cohort of control individuals with wild-type alleles, the loss of interruption was not observed, which suggests that the variant is more prevalent at longer CAG repeat lengths, possibly due to the greater propensity for expansion (Wright et al., 2019). All HD patients absent for the CAA trinucleotide presented with an earlier age at onset based on their modal $(CAG)_n$ compared to their predicted age at onset (Langbehn et al., 2004; Wright et al., 2019). This highlights the protective effect of the CAA trinucleotide on disease progression as its loss has now been established to advance age at onset. The results presented in this report fit with the current literature by identifying the loss of interruption variant in four HD patients, whom presented with early age at onsets. Overall, this reinforces that repeat interruptions are modifiers HD age at onset and highlights the importance of deducing the sequence composition, which gives more informative age at onset predictions.

7.3.3 Somatic Mosaicism

The somatic mosaicism profile determined in this report by Illumina Miseq and SP-PCR raises the question of whether HD neuropathogenesis is driven by the rarer extreme CAG repeat lengths or the greater population of common variants, i.e., smaller repeat changes reaching up to 40 CAGs more than the progenitor allele. Calculating the proportion of common variants mirrors the somatic mosaicism profile of the premanifest HD individuals reported in Kennedy *et al*., 2003, even though the *post-mortem* brains analysed here were from end stage HD patients. As the proportion of common variants is largest in the striatum (putamen > caudate nucleus) compared to the cortical regions, this suggests that the cells in the most vulnerable regions can survive with the burden of the CAG repeat up to approximately 84 CAGs. However, SP-PCR analysis determined that the largest mutational load is in the cortical regions and not in the striatum. This potentially supports the hypothesis that cells with the largest CAG repeat sizes are primarily lost, hence why the caudate nucleus appears more stable than the putamen. Taken together, these profiles confirm that the somatic mosaicism profile echoes that of HD neuropathogenesis and further suggests that the most vulnerable cells do not survive when carrying extremely large CAG repeat sizes.

This is further supported in a previous report, which hypothesised that tissue and cell type-specific differences in CAG repeat expansion plays a role in HD neurodegeneration (Shelbourne et al., 2007). Investigation of CAG repeat length variation between striatal and cortical neurons in both Vonsattel grade 0 and grade 1 *post-mortem* tissues revealed that the median mutational load did not significantly differ between the two regions. However, the proportion of cortical neurons that had CAG repeat gains of ≥ 20% the inherited mutation length was significantly less in grade 0 compared to grade 1 tissues, where the proportion of striatal and cortical neurons containing longer repeat tracts was similar. This suggests that the cells containing the largest CAG repeats in the striatum are primarily lost (Shelbourne et al., 2007). The pattern of somatic mosaicism in the *post-mortem* human HD brains suggests that the mutational load variability may alter as disease pathology progresses in that CAG repeat length gains are more prominent in the striatum compared to the cortex in low grade HD cases. This distinction becomes less obvious as the disease progresses. As it is generally accepted that there is a correlation between increasing CAG repeat length gains and accelerated pathology, it is therefore conceivable that the results presented in Shelbourne *et al*., 2007 support the idea that

disease progression may be partly driven by extreme neuronal CAG repeat length gains evident in the striatum at grade 0 compared to grade 1 (Shelbourne et al., 2007).

Furthermore, Swami *et al*., 2009 used skewness, a measurement of the degree of symmetry of a distribution, to measure somatic instability in a cohort of HD *post-mortem* samples defined by either early or late onset with mean age at onsets differing by approximately 30 years (Swami et al., 2009). The frontal cortex was targeted for this study as in contrast to the striatum, the frontal cortex displays relatively high levels of somatic instability in end-stage HD *post-mortem* brains (Kennedy, 2003). SP-PCR analysis determined various degrees of somatic mosaicism in the cortical samples with a dominant expansion bias. The results revealed that as the CAG repeat length increases, the age at onset decreases (Swami et al., 2009). Early onset individuals presented with larger somatic expansions than late onset individuals. A marked statistical difference was observed in the magnitude of the average maximum expansion for each group. The early onset individuals had a mean of 42 CAGs compared to the late onset individuals with a mean of 29 CAGs greater than the progenitor allele. These results suggest that further repeat expansions are biased towards longer alleles in individuals with an earlier disease onset (Swami et al., 2009). Early onset individuals exhibited a greater right skewness than late onset individuals, which is based on the assumption that as repeat length changes are bias towards expansion, their distributions are skewed to the right. A negative correlation was determined between skewness and residual age at onset, which suggests an association between greater somatic instability and an earlier onset (Swami et al., 2009). Similarly, regression analysis revealed that skewness was a significant predictor of residual onset, with an increase in right skewness being associated with a lower residual onset (Swami et al., 2009). The results demonstrate that greater somatic instability in the cortex is associated with an earlier age at onset, which highlights the modifying role of somatic mosaicism (Swami et al., 2009).

Most recently, the loss of the CAA trinucleotide interruption was determined to be associated with increased somatic and germline instability (Wright et al., 2019). A CAG expansion ratio was calculated from whole blood samples of HD individuals with the loss of interruption variant and those with canonical sequences. Accordingly, the somatic expansion ratio was associated with increased CAG repeat length and older age. The results revealed that carriers of the loss of interruption variant had an increase in expansions compared to canonical sequence carriers (Wright *et al*., 2019). Furthermore, the absence of the CAA trinucleotide was determined to be associated with increased

germline CAG repeat instability, which was evident from chromatograph traces of CAG sizing PCR products and SP-PCR analysis in sperm samples (Wright et al., 2019). Therefore, in addition to sequence variants, somatic instability has been identified as a modifier of HD phenotype.

7.3.4 DNA repair

Recent GWAS of HD patients have highlighted DNA repair genes as modifiers of age at onset and disease severity (GEM-HD, 2019). Such genes include *FAN1*, *RRM2B*, *MLH1*, *MSH3*, *PMS2*, *PMS1* and *LIG1*. Many of the implicated genes are members of the mismatch repair pathway, which is of specific interest due to its influence on the somatic expansion of the CAG repeat. In HD mouse models, further somatic expansion is prevented upon deletion of the mismatch repair genes *Msh2*, *Msh3*, *Mlh1* and *Pms2* (Gomes-Pereira, 2004; Manley et al., 1999; Pinto et al., 2013; Tomé et al., 2013). These initial studies in mouse models have been recapitulated in human data with disease-associated SNPs identified near the *MLH1*, *MSH3*, *PMS2*, and *PMS1* loci (GEM-HD, 2019). *FAN1*, one of the lead modifying genes identified, encodes a nuclease involved in interstrand DNA crosslink repair, which is additionally recruited to stalled replication forks, physically interacts with *MLH1*, and is required for homologous recombination. The previous and recent GWAS have reported two independent signals in *FAN1* that alter HD age at onset with opposing effects; 6.1 year onset-hastening and 1.4 year onset-delaying variants (GEM-HD, 2015, 2019). In contrast to the mismatch repair genes, *FAN1* expression is associated with reduced somatic expansion of the CAG repeat and is associated with later onset (Goold et al., 2019). The knock-out of *Fan1* in the knock-in *Hdh* mouse models carrying 50 and 111 CAGs, representing adult- and juvenile-onset HD ranges, respectively, increased somatic expansion (Loupe et al., 2020). However, the simultaneous knock-out of *Mlh1* inhibited the *Fan1* knock-out-induced acceleration of the CAG repeat expansion. These results suggest that functional *Mlh1* is required for the increased instability associated with *Fan1* loss (Loupe et al., 2020).

The previously identified DNA repair gene modifiers associated with HD phenotype have been identified in a cohort of approximately 9,000 HD patients, which highlights our limited sample power. Nonetheless, the results reported here adhere to the previous literature and show that there is a significant association with age at onset across all SNPs when using the regression parameters from Bettencourt *et al*., 2016. No significant associations of individual SNPs were observed with age at onset after Bonferroni

correction. However, prior to correction, significant associations with earlier age at onset were identified with variants in *FAN1* (rs114136100, rs150393409), *PMS2* (rs12534423) and *ERCC3* (rs1566822). *FAN1*, *PMS2* and *ERCC3* all play a role in DNA repair, and are specifically involved in interstrand crosslink repair, mismatch repair and nucleotide excision repair, respectively (GEM-HD, 2019; Weeda et al., 1997). This report does not capture the most significant signal previously identified in *FAN1*, which is most likely due to the low power of the study and the low minor allele frequency of the SNP. However, we recapitulated a significant association with earlier age at onset for the rs150393409 SNP in *FAN1*, which was determined to have an onset hastening effect of 5.2 years in the latest GWAS (GEM-HD, 2019). Additionally, rs150393409 is the top ranked coding SNP from the analysis of GEM-HD, 2019, 2015 and Moss *et al*., 2017. This reinforces that genetic variation in DNA repair pathway genes plays a role in modifying the HD phenotype.

7.3.5 Trinucleotide Repeat Sequencing

Typically, trinucleotide repeat diseases are diagnosed by PCR amplification of the repeat regions and the fragment size is determined in combination with appropriate controls by capillary electrophoresis. However, previous studies have reported that 3% to 13% of alleles fall outside the error limits accredited by the best practice guidelines (Losekoot et al., 2013; Quarrell et al., 2012). Allelic dropout and misinterpretation of the genotype can occur from the amplification failure of alleles with large CAG repeat sizes, which can present as homozygous for the wild type allele. Additionally, heterozygosity of the flanking CCG repeat can contribute to the incorrect calling of CAG repeat sizes (Losekoot et al., 2013). The current genetic testing for HD does not give the base-by-base sequence configuration. Due to the errors and inconsistencies in sizing the CAG repeat via the previous methods, this suggests that sequencing the repeat could be more beneficial in obtaining an accurate CAG repeat size. Similarly, as interruptions have been reported in the trinucleotide repeat diseases and are known to play a role in clinical phenotype and disease heritability, it is crucial that their determination is incorporated into the diagnostic procedure (Fratta et al., 2014; GEM-HD, 2019; Menon et al., 2013).

In relation to FRDA, it is clear that the methods for GAA repeat sizing are not without their limitations. Long-range PCR and *Mbo*II digestion will only detect approximately 20 bp added to either of the flanking regions, or > 50 bp of internal interruptions within the GAA repeat. *Mbo*II digestion analysis will not detect smaller interruptions less than 50

bp or $(GAAGA)_n$ interruptions as such sequences will be cut by the enzyme. TP-PCR is limited to the last approximate 100 GAAs at the 3' end and neither method identifies the sequence at the base pair level. As such, further analysis is often needed when there are discrepancies between the clinical and molecular data. These pitfalls are usually due to the presence of unusual sequences, such as large deletions or interruptions (Santoro et al., 2020). This is evident in a recent study where a misdiagnosis occurred due to an undiscovered benign GAA repeat interruption that could not be captured by the current diagnostic methods (Santoro et al., 2020). Long-range PCR of two siblings reported two small pathologically expanded GAA repeat alleles in both after being tested for FRDA when the elder sister presented with symptoms. The elder sister was diagnosed with late-onset FRDA and the unaffected younger sister was diagnosed with pre-symptomatic late-onset FRDA. Further analysis was carried out on the parents and the two siblings including direct sequencing of the long-range PCR products. The sequencing results determined that the unaffected younger sister and their healthy mother carried an expanded GAA repeat allele and an uncommon $(GAAGGA)_{66\text{-}67}$ repeat interruption, which lacked pathogenicity and mimicked a GAA repeat expansion resulting in the wrong initial diagnosis of pre-symptomatic late-onset FRDA. Ohshima *et al.*, 1999 similarly reported a $(GAAGGA)_{65}$ GAA repeat interruption which was concluded to be a benign variant and further supports the lack of pathogenicity associated with this interruption in combination with a pure GAA repeat expansion (Ohshima et al., 1999). The intergenerational stability of this sequence, which can be seen in the healthy mother and the unaffected daughter reinforces the stabilising properties of repeat sequence interruptions. Furthermore, this cements the need for elucidating the exact sequence configuration in order to accurately determine the relationship between GAA repeat interruptions and FRDA phenotype.

Mutation detection in the trinucleotide repeat diseases has been revolutionised in the past few years by genomic sequencing technologies. Next generation sequencing has identified a plethora of *de novo* mutations, however, many rare diseases are not fully diagnosable due to its short read technology. Even with the most advanced bioinformatics algorithms, structural variants, repetitive regions, extreme GC content, or sequences with multiple homologous elements, are difficult to characterise (Höijer et al., 2018). These limitations provided the incentive to develop third generation sequencing platforms, such as PacBio and Oxford Nanopore Technologies, which provide long read sequencing using SMRT technology without the need for prior DNA amplification by PCR (van Dijk et al.,

2018). Removing the need for PCR removes any PCR related bias, such as PCR stutter, and leaves the DNA in its native state. Recent reports have highlighted the use of SMRT sequencing to generate high-quality *de novo* human genome assemblies and to resolve complex repetitive regions (Seo et al., 2016; Shi et al., 2016). In contrast, the bioinformatics pipeline used to analyse the Illumina MiSeq data, ScaleHD, aligns the sequence reads to a reference genome consisting of approximately 4,000 typical *HTT* references. Although alignment-based approaches are suitable in many cases, they are not optimal for situations where it is difficult to make *a priori* assumptions on the configuration and structure of the resolved sequence.

PacBio's No-Amp Targeted sequencing describes the sequencing of targeted regions without PCR amplification. This method was tested on 11 HD blood samples, which were previously sized by fragment analysis, using the RSII system to examine the CAG and CCG repeats in *HTT* (Höijer et al., 2018). An analysis pipeline was developed to compute the repeat counts on both alleles and visualise their configurations without aligning the sequence reads to a human reference. SMRT sequencing of the HD samples gave an average of 157 on-target reads for *HTT*. The most common CAG repeat size for each allele determined by SMRT sequencing agreed with the previous fragment analysis. Similarly, the population of CCG repeat sizes obtained from each allele per patient aligned with previous reports; 63% carried the commonest allele of seven CCGs, followed by 32% with 10 CCGs and 5% with nine CCGs (Höijer et al., 2018). However, in comparing the efficiency of the technique on DNA extracted from HEK 293 cells versus blood samples, the average number of on-target reads for *HTT* was 209 and 157, respectively. The discrepancy between the HEK 293 cells and the HD samples was attributed to the lower yield of DNA from human blood samples. As such, instead of using four restriction enzymes for the genome complexity reduction step, only two were used on the HD samples which could also influence the variability in enrichment results (Höijer et al., 2018).

The application of this method on our HD sample cohort has the potential to determine the exact sequence configuration and somatic mosaicism pattern with confidence. No-Amp Targeted sequencing requires at least 5 µg of input DNA, which limits its use to specific sample types, where it is easy to obtain large amounts of DNA (Höijer et al., 2018). Additionally, this technique is not currently ready for implementation into the clinic due to the cost, need to simplify the laboratory protocol, including the reduction of the required input DNA concentration, and the need to reduce on-target read number

variation. In addition to the work by Höijer *et al*., 2018 where the input DNA influenced the number of on target reads, this is evident in the results reported in this study. Out of the four HD patient blood-derived samples sequenced by PacBio's SMRT sequencing, two yielded enough on-target molecules to give convincing results; 39 molecules for HD patient 13 and 48 molecules for HD patient 3. HD patient 24 yielded 10 on-target molecules, which is sufficient for the repeat analysis tool but insufficient to give full confidence in the result. HD patient 8 was analysed on the Sequel System, which yielded no results, as the sample did not survive library preparation. This protocol is currently being optimised by PacBio to reduce the amount of DNA required, which would allow this technology to be applicable to more sensitive samples including DNA extracted from *post-mortem* brain tissue.

An additional possibility with No Amp Targeted sequencing is the ability to multiplex. During the sequencing of *HTT*, Höijer *et al*., 2018 examined additional loci harbouring repeat expansions; *FMR1*, *ATXN10* and *C9orf72*. Multiplexing theoretically allows the targeting of nearly any region in the genome and thus its utility could be applied in determining the influence of CAG repeat size variations in other polyglutamine disease-associated genes (PDAGs) in relation to HD age at onset. A previous report revealed that the age at onset in several SCAs is modulated by the CAG repeat sizes in the wild-type range in other PDAGs (Tezenas du Montcel et al., 2014). The age at onset in SCA3 patients was found to be modulated by the longer wild-type *HTT* CAG repeat size (the longer the CAG repeat, the later the age at onset), thus, the influence of CAG repeat size variations in other PDAGs was subsequently examined in relation to HD age at onset (Stuitje et al., 2017). Clinical data and DNA samples were obtained from manifest HD patients enrolled in the European Huntington's Disease Network REGISTRY Study. The $(CAG)_n$ for each examined PDAG (*ATN1, ATXN1, ATXN7, CACNA1A, HTT, AR, ATXN2, ATXN3,* and *TBP*) was determined and a multiple linear regression model was used to assess the association between the PDAGs $(CAG)_n$ and HD age at onset. The results revealed that the HD age at onset in this cohort was inversely associated with the $(CAG)_n$ of the *HTT* expanded allele, which accounted for 66.1% of the age at onset variation (Stuitje et al., 2017). Additionally, the larger *ATXN3* allele was associated with a later age at onset in the HD patients yet there was no significant interaction between either of the *ATXN3* alleles and the expanded *HTT* allele. A significant interaction was determined between the *HTT* expanded $(CAG)_n$ and the larger *CACNA1A* allele. This association revealed that for HD patients with a below median $(CAG)_n$ in their expanded *HTT* allele,

more CAGs in the longer *CACNA1A* allele resulted in a later age at onset. A significant interaction was also determined between the expanded *HTT* $(CAG)_n$ and the larger *AR* allele. Specifically, for HD patients with a below median expanded *HTT* $(CAG)_n$, more CAGs in the larger *AR* allele delayed age at onset, however for patients with an above median expansion, the larger *AR* CAG repeats advanced onset (Stuitje et al., 2017). Therefore, these data indicate that the age at onset in this HD cohort is modulated by the wild-type $(CAG)_n$ of *ATXN3*, *CACNA1A* and *AR* (Stuitje et al., 2017). This work further highlights the biological interaction between the PDAGs and suggests that the $(CAG)_n$ of the PDAGs could be another factor modifying the HD phenotype presented by the cohort in this thesis.

7.4 Limitations

7.4.1 Clone Sequencing

It is challenging to derive a firm conclusion on the faithfulness of the sequence alterations obtained from clone sequencing as our study is limited by both patient and clone number and by the methodology. The small number of clones disallowed some of the sequence configurations to be converted into percentages of occurrence per patient (number of clones/total number of clones per patient), which in turn made any conclusions speculative. Cloning from a PCR product captures one amplicon from the population and as such, there is a much higher chance of this molecule containing an error. Although steps have been taken to minimise artefacts including a proof-reading polymerase and recombination-deficient (Stbl3) *E.coli*, it is not possible to eliminate artefacts associated with polymerase slippage and mis-priming or template switching (Gao et al., 2012). Therefore, without a sufficient number of clones to confirm otherwise, any alterations identified in the forward or reverse sequence orientation alone are most likely due to experimental artefacts, such as slippage, which is known to occur towards the end of the repeat.

7.4.2 Third Generation Sequencing

The consistent limitation found during the attempt at sequencing 10 HD samples by PacBio SMRT sequencing was the required quality and quantity of the input DNA. Of the 10 HD samples sent for sequencing, six were analysed on the RSII instrument and four on the Sequel System. Only two samples analysed on the RSII instrument gave enough on-target molecules to have confidence in the result and none of the remaining four HD patient blood samples survived the library preparation step for sequencing on the Sequel System. PacBio amplification-free, CRISPR/Cas9 targeted enrichment SMRT sequencing does not use amplification techniques for the region of interest. Therefore, the input DNA quality is directly reflected in the sequencing results. DNA damage or contaminants within the input DNA will impair the performance of the system. More specifically, the input DNA must be double-stranded as single-stranded DNA will not be made into a SMRTbell template, it has to have undergone minimum freeze-thaw cycles, no exposure to > 65°C or extreme pH, a purity ratio of 1.8 to 2.0 is required, it cannot contain chelating agents, divalent metal cations, denaturants, detergents, insoluble material or RNA, it cannot be exposed to intercalating fluorescent dyes or UV radiation and it cannot contain carry over contamination from the starting organism or tissue. As

such, and in order to ensure the future compatibility of samples with this system, it is recommended to use the pre-sequencing methods advised by PacBio, including DNA clean-up methods and fluorescence-based quantification for enhanced sensitivity. Additionally, No-Amp Targeted sequencing requires at least 5 μg of input DNA, which limits the use of this method across different tissues.

The limitation of input DNA quality and quantity was further evidenced in the sizing results obtained from Nanopore sequencing for the HD patient blood and *post-mortem* brain samples. The percentage of read counts calculated for the modal wild-type and expanded allele in the HD patient blood samples were consistently low with a maximum 9.4% of total reads mapping to the progenitor allele size. Additionally in the *post-mortem* brain samples, the read counts per $(CAG)_n$ were inconsistent between samples which hindered any somatic mosaicism analysis being performed. Library preparation cannot be successfully achieved with insufficient DNA quality. This may account for the samples where the repeat counts could not be determined for the expanded allele and where the read counts are low, with one example being 7 reads for 42 CAGs in P72.10 occipital lobe compared to 564 reads for 43 CAGs in 72.10 temporal lobe. This further highlights the importance of having sufficient DNA quality.

7.4.3 *Post-mortem* Samples

A major limitation in using *post-mortem* samples is accessibility and availability. Our report includes six human HD *post-mortem* brains, however of these, only four had the striatum available for analysis. Sample quality is an inherent limitation when dealing with *post-mortem* tissues in which several factors have to be considered; cell death and autolysis, the *post-mortem* index (PMI), and the known effects of the fixative on the shape and magnetic resonance characteristics of the tissue (Alkemade et al., 2018). The condition of each brain used in this study varied in terms of PMI, fixation method and the presence of freeze-thaw artefacts. The PMI ranged from 10 to 96 hours and one HD *post-mortem* brain received by the UCL Queen Square Brain Bank was previously cut mid-sagittally with both halves frozen resulting in severe post-mortem freeze artefact. Additionally, the HD *post-mortem* brain samples used in this study are representative of a bulk tissue section and the role of changes in cellular composition is unknown. Therefore, we cannot exclude the fact that cellular atrophy contributes partially to the changes observed in our study. Insufficient resources prevented taking multiple cuts from the same region and performing multiple DNA extractions, which would give a more

representative environment of the region for subsequent analysis. In response to the ever growing need of examining the HD brain, the Huntington's Disease Society of America and the CHDI Foundation have collaborated to create HD LEGACY, which is aimed at promoting and supporting brain and other organ donations from HD affected families.

7.5 Future Work

7.5.1 Single-nuclei RNA Sequencing

The work presented in this study is sensitive to the bulk *post-mortem* tissue samples, thus ignoring the heterogeneity of individual cells as measurements are summed over the remaining cell populations. Therefore, future work in the *post-mortem* brain regions to deduce the single cell content and their relative contribution to the neuropathogenesis examined in this report is necessary. Single-cell transcriptome profiling by RNA sequencing has enhanced the information available on the complexity of cell types in the nervous system based on gene expression (Poulin et al., 2016). However, cells from the CNS have been under-characterised due to the difficulty of isolating intact, whole cells. Neurons are highly interconnected and separating them physically by such methods as laser-capture microdissection results in considerable damage and stress to the cells, which has the potential to alter their gene expression (Grindberg et al., 2013). Single nuclei RNA sequencing (snRNA-Seq) is an alternative approach that has been developed and takes advantage of the low levels of mRNA within nuclei. It has been reported that nuclei can substitute for whole cells as the gene expression signatures were proven to be equivalent between both (Krishnaswami et al., 2016). A method that obtains nuclear transcriptomes from *post-mortem* brain tissue stored at -80°C has been described, which makes brain archives, such as our samples, accessible for RNA sequencing (Krishnaswami et al., 2016).

There has been a surge in reports examining the transcriptome of HD *post-mortem* brains with technologies such as the 10X Chromium platform currently leading the field. The 10X Chromium single cell/nuclei gene expression solution allows whole transcriptome profiling through advanced microfluidics to perform single cell or nuclei partitions, each containing an identifying barcode for downstream analysis. snRNA-Seq using the 10X Chromium platform was performed on caudate and putamen from HD *post-mortem* brains that were characterised by Vonsattel grades 2, 3 or 4, and corresponding controls (Lee et al., 2020). Network based clustering revealed the cell types present and distinguished between medium spiny neurons, the cell type most affected in HD, from the direct and indirect pathway. This distinction was less obvious in the HD samples compared to controls and the number of medium spiny neurons from the indirect pathway decreased with HD grade progression, highlighting their selective vulnerability compared to those from the direct pathway. Differential gene expression analysis determined that several striatal cell types displayed a downregulation of several medium spiny neuronal marker

genes and an upregulation of mitochondrial-encoded RNAs (Lee et al., 2020). Comparison of the differentially expressed genes in the HD samples with the genes linked to the DNA repair associated genes in GEM-HD, 2019 showed that several genes; *TCERG1*, *PMS1*, *FAN1*, *MSH3* and *MLH3*, were downregulated in the medium spiny neurons (Lee et al., 2020). Although future studies are needed to correlate these findings to HD progression, it highlights the applicability of this method to our HD *post-mortem* brains to examine the microenvironment between different regions, affected versus unaffected, and has the potential to determine the remaining cell types and their relative contributions to specific neuropathogenesis.

7.5.2 Induced Pluripotent Stem Cells

HD *post-mortem* brain samples have contributed invaluable insights into the profile of somatic mosaicism and the specific neuropathogenesis of HD, however, the translational value of this model could be further enhanced when combined with *in vitro* models derived from patient-specific iPSCs. The potential of iPSCs to differentiate into any cell type offers the opportunity to study specific neuronal subtypes at a defined disease stage, without having to exogenously overexpress the disease-specific pathogenic proteins. In the trinucleotide repeat diseases described in this study, the loss of disease-specific primary neurons results in severe atrophy of the related brain region. In HD specifically, by end-stage disease there is often extreme striatal atrophy, which results in no sample material. This is evident in our cohort as only four out of the six HD *post-mortem* brains had striatal tissue available. The neuropathogenesis in these diseases is complex and some underlying mechanisms remain to be elucidated. Accumulating evidence has revealed the generation of various types of brain cells in 2D cultures and further advances in 3D culture systems are paving the way for the development of organoids in which multiple brain cell types and specific brain regions are differentiated to recapitulate the more complex features of the brain (Conforti et al., 2018). Additionally, the advent of CRISPR/Cas9 has enhanced the efficiency of genome editing and accelerated the generation of isogenic controls, which retain the genetic background of the patients and make precise genotype-phenotype correlations possible. Either in combination with *post-mortem* samples or singularly, iPSCs present unprecedented opportunities to model the trinucleotide repeat diseases.

7.5.2.1 HD iPSCs

The degeneration of two basal ganglia structures: the caudate nucleus and the putamen, which form the neostriatum, is the primary characteristic of HD neuropathogenesis. However, additional brain regions including the cortex also degenerate as the pathology progresses (Rosas et al., 2008). Selective neuronal populations in the HD striatum, such as the GABAergic medium spiny neurons, are the most vulnerable cell type, whereas the large cholinergic interneurons, the medium sized GABAergic interneurons and glial cells, are relatively spared (Cicchetti et al., 1996). In this report, we have examined the relationship between the somatic instability profile and the selective neurodegeneration in HD *post-mortem* brains. Somatic instability was determined by two contrasting methods; Illumina MiSeq, which calculates the proportion of small CAG repeat changes relative to the mode of the progenitor allele, and SP-PCR, which quantifies the presence of extremely large CAG repeats. Common to both of these techniques was that they were performed on bulk tissue samples from eight HD *post-mortem* brain regions; frontal lobe, temporal lobe, occipital lobe, putamen, caudate nucleus, cerebellum, pons, and medulla. As such, the exact cellular composition is unknown and we cannot exclude that cellular atrophy contributes partially to the results observed in this study. Therefore, differentiating HD-affected iPSCs into the specifically vulnerable and spared cell types would allow us to determine their inherent rate of somatic instability and the contribution of somatic mosaicism in defined neuronal populations to the HD phenotype.

Advances in genome editing has enabled the exploration of CAG repeat-dependent and cell type-specific effects that might contribute to the neuropathogenesis of HD. Ooi *et al*., 2019 genetically engineered human embryonic stem cell (hESC) lines to carry 30, 45, 65 and 81 CAGs (Ooi et al., 2019). Creating an isogenic HD allelic series ensures that the cells contain the same genetic background. Therefore, any differences in functional and molecular measurements can be directly attributed to the length of the CAG repeat. This cell series was differentiated into cell types with varying vulnerability to mutant *HTT* including neural progenitor cells, neurons, hepatocytes and skeletal muscle myotubes (Ooi et al., 2019). Genome-wide RNA sequencing was performed on all cell types carrying each CAG repeat length. Transcription profiles were distinguishable between CAG repeat lengths and cell types. Differential gene expression analysis revealed strong cell type specificity with only eight genes differentially expressed between all cell types. Fold changes of differentially expressed genes between the 45, 65 and 81 CAG repeat lengths relative to the 30 CAG repeat length in each cell type were determined to analyse

CAG size-dependent transcriptional signatures. The fold changes were clustered and scaled between each cell type and the clusters were subsequently analysed for significant functional enrichments (Ooi et al., 2019). In neurons specifically, there was significant enrichment for biological process terms relating to cell cycle progression. These results highlight the potential to explore cell-type specific and CAG repeat-dependent transcriptional changes relevant to HD. Delineating the diversity of intermingled striatal cell types within human HD *post-mortem* brains has been difficult due to technical challenges and limited availability. HD affected hESCs that are differentiated into specific cell types presents a desirable alternative approach, however, certain considerations must be noted. The transcriptomic data may not solely represent CAG repeat-related differences but also cellular heterogeneity if the analysis is performed on bulk cellular populations and not purified cells expressing a defined set of markers. Additionally, the differentiated cell type must possess the HD phenotype in which you are examining. In contrast to the age-dependent CAG repeat instability present in the striatum of HD patients and mouse models, minimal CAG repeat instability was observed in the isogenic HD panel after cell-type specific differentiation (Ooi et al., 2019).

7.6 Conclusions

To investigate the role of modifiers in the trinucleotide repeat diseases, it is important to detect pathogenic repeat sizes accurately and determine their sequence configuration. Pathogenic allele repeat length is critically associated with disease severity and the age at symptom onset, which suggests that any alterations have the potential to modify phenotype. In accordance with the current literature, this report identifies sequences interruptions as disease modifiers in a cohort of HD patients with similarly sized pathogenic alleles but extreme phenotypic variation, and further highlights the diagnostic benefit of sequencing the repeat (GEM-HD, 2019). Sequencing technologies are advancing for the better, especially in terms of high throughput, read length, and cost. However, the new technologies are not without their teething problems. The major barrier experienced with PacBio's No-Amp SMRT sequencing on our *post-mortem* brain samples especially, was the high quantity and quality of input DNA needed. In order for this technology to be used more routinely, PacBio are currently optimising this technique to be more forgiving in relation to the extent of DNA quantity and quality required. This will make the platform more translatable to clinic and in the general laboratory setting in the future.

Illumina MiSeq sequencing proved to be most efficient in identifying the HD sequence configuration at the base-pair level. This protocol prepares sequencing libraries of the *HTT* CAG repeat by a single PCR using locus-specific primers incorporating sequencing adaptors for sequencing on the MiSeq platform (Ciosi et al., 2018). The MiSeq sequencing output allowed a validation of the clone sequencing results and also quantified somatic mosaicism by determining the proportion of common variants in our cohort of HD *post-mortem* brains. Therefore, not only does this technology have the potential to improve HD diagnosis by accurately sizing the CAG repeat and revealing modifying variants, it also aids in elucidating the somatic mosaicism profile. The somatic mosaicism results determined in this report by MiSeq and SP-PCR, reinforce the hypothesis that cells carrying the largest expansions are primarily lost and indicates that somatic instability is additionally a modifier of HD. This report contributes to the identification of trinucleotide repeat disease modifiers which will ultimately enable a more precise diagnosis and investigation of novel therapeutic targets.

References

Alkemade, A., Groot, J.M., Forstmann, B.U., 2018. Do We Need a Human post mortem Whole-Brain Anatomical Ground Truth in in vivo Magnetic Resonance Imaging? Front. Neuroanat. 12. https://doi.org/10.3389/fnana.2018.00110

Al-Mahdawi, S., Ging, H., Bayot, A., Cavalcanti, F., La Cognata, V., Cavallaro, S., Giunti, P., Pook, M.A., 2018. Large Interruptions of GAA Repeat Expansion Mutations in Friedreich Ataxia Are Very Rare. Front. Cell. Neurosci. 12. https://doi.org/10.3389/fncel.2018.00443

Andrew, S.E., Paul Goldberg, Y., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M.A., Graham, R.K., Hayden, M.R., 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat. Genet. 4, 398–403. https://doi.org/10.1038/ng0893-398

Aronin, N., Chase, K., Young, C., Sapp, E., Schwarz, C., Matta, N., Kornreich, R., Landwehrmeyer, B., Bird, E., Beal, M.F., 1995. CAG expansion affects the expression of mutant Huntingtin in the Huntington's disease brain. Neuron 15, 1193–1201. https://doi.org/10.1016/0896-6273(95)90106-x

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. https://doi.org/10.1212/WNL.0000000000005690

Bates, G., Tabrizi, S., Jones, L. (Eds.), 2014. Huntington's disease, 4th edition. ed, Oxford monographs on medical genetics. Oxford University Press, Oxford ; New York.

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. Nat. Rev. Dis. Primer 15005. https://doi.org/10.1038/nrdp.2015.5

Behjati, S., Tarpey, P.S., 2013. What is next generation sequencing? Arch. Dis. Child. - Educ. Pract. Ed. 98, 236–238. https://doi.org/10.1136/archdischild-2013-304340

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, L.E., Alonso-Vilatela, M.E., Lima, M., Raposo, M., Traynor, B., Sweeney, M., Wood, N., Giunti, P., The SPATAX Network, 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: DNA Repair Pathways Modify polyQ Disease Onset. Ann. Neurol. 79, 983–990. https://doi.org/10.1002/ana.24656

Bowden, R., Davies, R.W., Heger, A., Pagnamenta, A.T., de Cesare, M., Oikkonen, L.E., Parkes, D., Freeman, C., Dhalla, F., Patel, S.Y., Popitsch, N., Ip, C.L.C., Roberts, H.E., Salatino, S., Lockstone, H., Lunter, G., Taylor, J.C., Buck, D., Simpson, M.A., Donnelly, P., 2019. Sequencing of human genomes with nanopore technology. Nat. Commun. 10. https://doi.org/10.1038/s41467-019-09637-5

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, 1202–1210.

Brown, M.B., 1975. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. Biometrics 31, 987. https://doi.org/10.2307/2529826

Campbell, I.M., Shaw, C.A., Stankiewicz, P., Lupski, J.R., 2015. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. TIG 31, 382–392. https://doi.org/10.1016/j.tig.2015.03.013

Campuzano, V., Montermini, L., Moltò, M.D., Pianese, L., Cossée, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Cañizares, J., Koutnikova, H., Bidichandani, S.I., Gellera, C., Brice, A., Trouillas, P., De Michele, G., Filla, A., De Frutos, R., Palau, F., Patel, P.I., Di Donato, S., Mandel, J.L., Cocozza, S., Koenig, M., Pandolfo, M., 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271, 1423–1427.

Caron, N.S., Desmond, C.R., Xia, J., Truant, R., 2013. Polyglutamine domain flexibility mediates the proximity between flanking sequences in huntingtin. Proc. Natl. Acad. Sci. 110, 14610–14615. https://doi.org/10.1073/pnas.1301342110

Castaldo, I., De Rosa, M., Romano, A., Zuchegna, C., Squitieri, F., Mechelli, R., Peluso, S., Borrelli, C., Del Mondo, A., Salvatore, E., Vescovi, L.A., Migliore, S., De Michele, G., Ristori, G., Romano, S., Avvedimento, E.V., Porcellini, A., 2018. DNA damage signatures in peripheral blood cells (PBMC) as biomarkers in prodromal Huntington's disease. Ann. Neurol. ana.25393. https://doi.org/10.1002/ana.25393

Castillo-Lizardo, M., Henneke, G., Viguera, E., 2014. Replication slippage of the thermophilic DNA polymerases B and D from the Euryarchaeota Pyrococcus abyssi. Front. Microbiol. 5. https://doi.org/10.3389/fmicb.2014.00403

Chong, S.S., McCall, A.E., Cota, J., Subramony, S.H., Orr, H.T., Hughes, M.R., Zoghbi, H.Y., 1995. Gametic and somatic tissue–specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. Nat. Genet. 10, 344–350. https://doi.org/10.1038/ng0795-344

Chung, M.Y., Ranum, L.P., Duvick, L.A., Servadio, A., Zoghbi, H.Y., Orr, H.T., 1993. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nat. Genet. 5, 254–258. https://doi.org/10.1038/ng1193-254

Cicchetti, F., Gould, P.V., Parent, A., 1996. Sparing of striatal neurons coexpressing calretinin and substance P (NK1) receptor in Huntington's disease. Brain Res. 730, 232–237. https://doi.org/10.1016/0006-8993(96)00307-1

Ciccia, A., Elledge, S.J., 2010. The DNA Damage Response: Making It Safe to Play with Knives. Mol. Cell 40, 179–204. https://doi.org/10.1016/j.molcel.2010.09.019

Ciosi, M., Ciosi, M., Cumming, S.A., Mubarak, A., Symeonidi, E., Herzyk, P., McGuinness, D., Galbraith, J., Hamilton, G., Monckton, D.G., 2018. Library preparation and MiSeq sequencing for the genotyping-by-sequencing of the Huntington disease HTT exon one trinucleotide repeat and the quantification of somatic mosaicism. Protoc. Exch. https://doi.org/10.1038/protex.2018.089

Ciosi, M., Maxwell, A., Cumming, S.A., Hensman Moss, D.J., Alshammari, A.M., Flower, M.D., Durr, A., Leavitt, B.R., Roos, R.A.C., Holmans, P., Jones, L., Langbehn, D.R., Kwak, S., Tabrizi, S.J., Monckton, D.G., 2019. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. EBioMedicine 48, 568–580. https://doi.org/10.1016/j.ebiom.2019.09.020

Ciotti, P., Di Maria, E., Bellone, E., Ajmar, F., Mandich, P., 2004. Triplet Repeat Primed PCR (TP PCR) in Molecular Diagnostic Testing for Friedreich Ataxia. J. Mol. Diagn. 6, 285–289. https://doi.org/10.1016/S1525-1578(10)60523-5

Cocozza, S., Costabile, T., Pontillo, G., Lieto, M., Russo, C., Radice, L., Pane, C., Filla, A., Brunetti, A., Saccà, F., 2020. Cerebellum and cognition in Friedreich ataxia: a voxel-based morphometry and volumetric MRI study. J. Neurol. 267, 350–358. https://doi.org/10.1007/s00415-019-09582-9

Cossée, M., Schmitt, M., Campuzano, V., Reutenauer, L., Moutou, C., Mandel, J.L., Koenig, M., 1997. Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations. Proc. Natl. Acad. Sci. U. S. A. 94, 7452–7457.

de la Monte, S.M., Vonsattel, J.P., Richardson, E.P., 1988. Morphometric demonstration of atrophic changes in the cerebral cortex, white matter, and neostriatum in Huntington's disease. J. Neuropathol. Exp. Neurol. 47, 516–525. https://doi.org/10.1097/00005072-198809000-00003

De Michele, G., Perrone, F., Filla, A., Mirante, E., Giordano, M., De Placido, S., Campanella, G., 1996. Age of onset, sex, and cardiomyopathy as predictors of disability and survival in Friedreich's disease: A retrospective study on 119 patients. Neurology 47, 1260–1264. https://doi.org/10.1212/WNL.47.5.1260

Desai, A., Gerson, S., 2014. Exo1 independent DNA mismatch repair involves multiple compensatory nucleases. DNA Repair 21, 55–64. https://doi.org/10.1016/j.dnarep.2014.06.005

Dexheimer, T.S., 2013. DNA Repair Pathways and Mechanisms, in: Mathews, L.A., Cabarcas, S.M., Hurt, E.M. (Eds.), DNA Repair of Cancer Stem Cells. Springer Netherlands, Dordrecht, pp. 19–32. https://doi.org/10.1007/978-94-007-4590-2_2

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., Dürr, A., Buée, L., Krystkowiak, P., Sablonnière, B., Blum, D., Huntington French Speaking Network, 2009. A genetic variation in the ADORA2A gene modifies age at onset in Huntington's disease. Neurobiol. Dis. 35, 474–476. https://doi.org/10.1016/j.nbd.2009.06.009

Djoussé, 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.A., McCusker, E., Gómez-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, 279–282. https://doi.org/10.1002/ajmg.a.20190

Dragatsis, I., Levine, M.S., Zeitlin, S., 2000. Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. Nat. Genet. 26, 300–306. https://doi.org/10.1038/81593

Dürr, A., Cossee, M., Agid, Y., Campuzano, V., Mignard, C., Penet, C., Mandel, J.-L., Brice, A., Koenig, M., 1996. Clinical and Genetic Abnormalities in Patients with Friedreich's Ataxia. N. Engl. J. Med. 335, 1169–1175. https://doi.org/10.1056/NEJM199610173351601

Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., 1993. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat. Genet. 4, 387–392. https://doi.org/10.1038/ng0893-387

Filla, A., De Michele, G., Cavalcanti, F., Pianese, L., Monticelli, A., Campanella, G., Cocozza, S., 1996. The relationship between trinucleotide (GAA) repeat length and clinical features in Friedreich ataxia. Am. J. Hum. Genet. 59, 554–560.

Flower, M., Lomeikaite, V., Ciosi, M., Cumming, S., Morales, F., Lo, K., Hensman Moss, D., Jones, L., Holmans, P., Monckton, D.G., Tabrizi, S.J., TRACK-HD Investigators, Kraus, P., Hoffman, R., Tobin, A., Borowsky, B., Keenan, S., Whitlock, K.B., Queller, S., Campbell, C., Wang, C., Langbehn, D., Axelson, E., Johnson, H., Acharya, T., Cash, D.M., Frost, C., Jones, R., Jurgens, C., 't Hart, E.P., van der Grond, J., Witjes- Ane, M.-N.N., Roos, R.A.C., Dumas, E.M., van den Bogaard, S.J.A., Stopford, C., Craufurd, D., Callaghan, J., Arran, N., Rosas, D.D., Lee, S., Monaco, W., O'Regan, A., Milchman, C., Frajman, E., Labuschagne, I., Stout, J., Campbell, M., Andrews, S.C., Bechtel, N., Reilmann, R., Bohlen, S., Kennard, C., Berna, C., Hicks, S., Durr, A., Pourchot, C., Bardinet, E., Nigaud, K., Valabre, R., gue, `, Lehericy, S., Marelli, C., Jauffret, C., Justo, D., Leavitt, B., Decolongon, J., Sturrock, A., Coleman, A., Santos, R.D., Patel, A., Gibbard, C., Whitehead, D., Wild, E., Owen, G., Crawford, H., Malone, I., Lahiri, N., Fox, N.C., Hobbs, N.Z., Scahill, R.I., Ordidge, R., Pepple, T., Read, J., Say, M.J., Landwehrmeyer, B., OPTIMISTIC Consortium, Daidj, F., Bassez, G., Lignier, B., Couppey, F., Delmas, S., Deux, J.-F., Hankiewicz, K., Dogan, C., Minier, L., Chevalier, P., Hamadouche, A., Catt, M., van Hees, V., Catt, S., Schwalber, A., Dittrich, J., Kierkegaard, M., Wenninger, S., Schoser, B., Schüller, A., Stahl, K., Künzel, H., Wolff, M., Jellinek, A., Moreno, C.J., Gorman, G., Lochmüller, H., Trenell, M., van Laar, S., Wood, L., Cassidy, S., Newman, J., Charman, S., Steffaneti, R., Taylor, L., Brownrigg, A., Day, S., Atalaia, A., Raaphorst, J., Okkersen, K., van Engelen, B., Nikolaus, S., Cornelissen, Y., van Nimwegen, M., Maas, D., Klerks, E., Bouman, S., Knoop, H., Heskamp, L., Heerschap, A., Rahmadi, R., Groot, P., Heskes, T., Kapusta, K., Glennon, J., Abghari, S., Aschrafi, A., Poelmans, G., Treweek, S., Hogarth, F., Littleford, R., Donnan, P., Hapca, A., Hannah, M., McKenzie, E., Rauchhaus, P., Cumming, S.A., Monckton, D.G., Adam, B., Faber, C., Merkies, I., 2019. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. Brain 142, 1876–1886. https://doi.org/10.1093/brain/awz115

Fratta, P., Collins, T., Pemble, S., Nethisinghe, S., Devoy, A., Giunti, P., Sweeney, M.G., Hanna, M.G., Fisher, E.M.C., 2014. Sequencing analysis of the spinal bulbar muscular

atrophy CAG expansion reveals absence of repeat interruptions. Neurobiol. Aging 35, 443.e1–3. https://doi.org/10.1016/j.neurobiolaging.2013.07.015
Gao, R., Zhao, A.H., Du, Y., Ho, W.T., Fu, X., Zhao, Z.J., 2012. PCR artifacts can explain the reported biallelic JAK2 mutations. Blood Cancer J. 2, e56. https://doi.org/10.1038/bcj.2012.2
GEM-HD, G. M. O. H. S. D. G.-H. C.-. 2015. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell,* 162**,** 516-26. https://doi.org/10.1016/j.cell.2015.07.003
GEM-HD, G. M. O. H. S. D. G.-H. C.-. 2019. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell,* 178**,** 887-900.e14. https://doi.org/10.1016/j.cell.2019.06.036
Goldberg, Y.P., McMurray, C.T., Zeisler, J., Almqvist, E., Sillence, D., Richards, F., Gacy, A.M., Buchanan, J., Telenius, H., Hayden, M.R., 1995. Increased instability of intermediate alleles in families with sporadic Huntington disease compared to similar sized intermediate alleles in the general population. Hum. Mol. Genet. 4, 1911–1918.
Gomes-Pereira, M., 2004. Pms2 is a genetic enhancer of trinucleotide CAG·CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. Hum. Mol. Genet. 13, 1815–1825. https://doi.org/10.1093/hmg/ddh186
Gomes-Pereira, M., Hilley, J.D., Morales, F., Adam, B., James, H.E., Monckton, D.G., 2014. Disease-associated CAG·CTG triplet repeats expand rapidly in non-dividing mouse cells, but cell cycle arrest is insufficient to drive expansion. Nucleic Acids Res. 42, 7047–7056. https://doi.org/10.1093/nar/gku285
Gonitel, R., Moffitt, H., Sathasivam, K., Woodman, B., Detloff, P.J., Faull, R.L.M., Bates, G.P., 2008. DNA instability in postmitotic neurons. Proc. Natl. Acad. Sci. 105, 3467–3472. https://doi.org/10.1073/pnas.0800048105
Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. https://doi.org/10.1038/nrg.2016.49
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., 2018. FAN1 modifies Huntington's disease progression by stabilising the expanded HTT CAG repeat. Hum. Mol. Genet. https://doi.org/10.1093/hmg/ddy375
Gorbunova, V., Seluanov, A., Mao, Z., Hine, C., 2007. Changes in DNA repair during aging. Nucleic Acids Res. 35, 7466–7474. https://doi.org/10.1093/nar/gkm756
Graffelman, J. 2020. *CRAN - Package HardyWeinberg* [Online]. Comprehensive R Archive Network (CRAN). Available: https://cran.r-project.org/web/packages/HardyWeinberg/index.html [Accessed].
Grindberg, R.V., Yee-Greenbaum, J.L., McConnell, M.J., Novotny, M., O'Shaughnessy, A.L., Lambert, G.M., Araúzo-Bravo, M.J., Lee, J., Fishman, M., Robbins, G.E., Lin, X., Venepally, P., Badger, J.H., Galbraith, D.W., Gage, F.H., Lasken, R.S., 2013. RNA-sequencing from single nuclei. Proc. Natl. Acad. Sci. U. S. A. 110, 19802–19807. https://doi.org/10.1073/pnas.1319700110
Grondin, R., Kaytor, M.D., Ai, Y., Nelson, P.T., Thakker, D.R., Heisel, J., Weatherspoon, M.R., Blum, J.L., Burright, E.N., Zhang, Z., Kaemmerer, W.F., 2012. Six-month partial suppression of Huntingtin is well tolerated in the adult rhesus striatum. Brain J. Neurol. 135, 1197–1209. https://doi.org/10.1093/brain/awr333
Haddad, L.A., Mingroni-Netto, R.C., Vianna-Morgante, A.M., Pena, S.D., 1996. A PCR-based test suitable for screening for fragile X syndrome among mentally retarded males. Hum. Genet. 97, 808–812. https://doi.org/10.1007/BF02346194
Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003
Höijer, I., Tsai, Y.-C., Clark, T.A., Kotturi, P., Dahl, N., Stattin, E.-L., Bondeson, M.-L., Feuk, L., Gyllensten, U., Ameur, A., 2018. Detailed analysis of *HTT* repeat elements in human blood using targeted amplification-free long-read sequencing. Hum. Mutat. https://doi.org/10.1002/humu.23580

Holloway, T.P., Rowley, S.M., Delatycki, M.B., Sarsero, J.P., 2011. Detection of interruptions in the GAA trinucleotide repeat expansion in the FXN gene of Friedreich ataxia. BioTechniques 50, 182–186. https://doi.org/10.2144/000113615

Hood, L., Rowen, L., 2013. The human genome project: big science transforms biology and medicine. Genome Med. 5, 79. https://doi.org/10.1186/gm483

Hsiao, K.M., Lin, H.M., Pan, H., Li, T.C., Chen, S.S., Jou, S.B., Chiu, Y.L., Wu, M.F., Lin, C.C., Li, S.Y., 1999. Application of FTA sample collection and DNA purification system on the determination of CTG trinucleotide repeat size by PCR-based Southern blotting. J. Clin. Lab. Anal. 13, 188–193. https://doi.org/10.1002/(sici)1098-2825(1999)13:4<188::aid-jcla8>3.0.co;2-g

Ip, C.L.C., Loose, M., Tyson, J.R., de Cesare, M., Brown, B.L., Jain, M., Leggett, R.M., Eccles, D.A., Zalunin, V., Urban, J.M., Piazza, P., Bowden, R.J., Paten, B., Mwaigwisya, S., Batty, E.M., Simpson, J.T., Snutch, T.P., Birney, E., Buck, D., Goodwin, S., Jansen, H.J., O'Grady, J., Olsen, H.E., MinION Analysis and Reference Consortium, 2015. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research 4, 1075. https://doi.org/10.12688/f1000research.7201.1

Jama, M., Millson, A., Miller, C.E., Lyon, E., 2013. Triplet Repeat Primed PCR Simplifies Testing for Huntington Disease. J. Mol. Diagn. 15, 255–262. https://doi.org/10.1016/j.jmoldx.2012.09.005

Jones, L., Houlden, H., Tabrizi, S.J., 2017. DNA repair in the trinucleotide repeat disorders. Lancet Neurol. 16, 88–96. https://doi.org/10.1016/S1474-4422(16)30350-7

Kay, C., Collins, J.A., Miedzybrodzka, Z., Madore, S.J., Gordon, E.S., Gerry, N., Davidson, M., Slama, R.A., Hayden, M.R., 2016. Huntington disease reduced penetrance alleles occur at high frequency in the general population. Neurology 87, 282–288. https://doi.org/10.1212/WNL.0000000000002858

Kennedy, L., 2003. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. Hum. Mol. Genet. 12, 3359–3367. https://doi.org/10.1093/hmg/ddg352

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. Genet. 9, 2539–2544.

Koeppen, A.H., Morral, J.A., Davis, A.N., Qian, J., Petrocine, S.V., Knutson, M.D., Gibson, W.M., Cusack, M.J., Li, D., 2009. The dorsal root ganglion in Friedreich's ataxia. Acta Neuropathol. (Berl.) 118, 763–776. https://doi.org/10.1007/s00401-009-0589-x

Kovtun, I.V., McMurray, C.T., 2008. Features of trinucleotide repeat instability in vivo. Cell Res. 18, 198–213. https://doi.org/10.1038/cr.2008.5

Kraus-Perrotta, C., Lagalwar, S., 2016. Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. Cerebellum Ataxias 3. https://doi.org/10.1186/s40673-016-0058-y

Krishnaswami, S.R., Grindberg, R.V., Novotny, M., Venepally, P., Lacar, B., Bhutani, K., Linker, S.B., Pham, S., Erwin, J.A., Miller, J.A., Hodge, R., McCarthy, J.K., Kelder, M., McCorrison, J., Aevermann, B.D., Fuertes, F.D., Scheuermann, R.H., Lee, J., Lein, E.S., Schork, N., McConnell, M.J., Gage, F.H., Lasken, R.S., 2016. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat. Protoc. 11, 499–524. https://doi.org/10.1038/nprot.2016.015

La Spada, A.R., Taylor, J.P., 2003. Polyglutamines Placed into Context. Neuron 38, 681–684. https://doi.org/10.1016/S0896-6273(03)00328-3

Langbehn, D.R., Brinkman, R.R., Falush, D., Paulsen, J.S., Hayden, M.R., International Huntington's Disease Collaborative Group, 2004. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. Clin. Genet. 65, 267–277. https://doi.org/10.1111/j.1399-0004.2004.00241.x

Lee, D.-Y., McMurray, C.T., 2014. Trinucleotide expansion in disease: why is there a length threshold? Curr. Opin. Genet. Dev. 26, 131–140. https://doi.org/10.1016/j.gde.2014.07.003

Lee, H., Fenster, R.J., Pineda, S.S., Gibbs, W.S., Mohammadi, S., Davila-Velderrain, J., Garcia, F.J., Therrien, M., Novis, H.S., Gao, F., Wilkinson, H., Vogt, T., Kellis, M., LaVoie, M.J., Heiman, M., 2020. Cell Type-Specific Transcriptomics Reveals that Mutant

Huntingtin Leads to Mitochondrial RNA Release and Neuronal Innate Immune Activation. Neuron 107, 891-908.e8. https://doi.org/10.1016/j.neuron.2020.06.021
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. https://doi.org/10.1093/hmg/ddx286
Lee, J.-M., Zhang, J., Su, A.I., Walker, J.R., Wiltshire, T., Kang, K., Dragileva, E., Gillis, T., Lopez, E.T., Boily, M.-J., Cyr, M., Kohane, I., Gusella, J.F., MacDonald, M.E., Wheeler, V.C., 2010. A novel approach to investigate tissue-specific trinucleotide repeat instability. BMC Syst. Biol. 4, 29. https://doi.org/10.1186/1752-0509-4-29
Leggett, R.M., Clark, M.D., 2017. A world of opportunities with nanopore sequencing. J. Exp. Bot. https://doi.org/10.1093/jxb/erx289
Levene, M.J., 2003. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science 299, 682–686. https://doi.org/10.1126/science.1079700
Liu, Q., Zhang, P., Wang, D., Gu, W., Wang, K., 2017. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. Genome Med. 9. https://doi.org/10.1186/s13073-017-0456-7
Losekoot, M., van Belzen, M.J., Seneca, S., Bauer, P., Stenhouse, S.A.R., Barton, D.E., Barton, D.E., 2013. EMQN/CMGS best practice guidelines for the molecular genetic testing of Huntington disease. Eur. J. Hum. Genet. 21, 480–486. https://doi.org/10.1038/ejhg.2012.200
Loupe, J.M., Pinto, R.M., Kim, K.-H., Gillis, T., Mysore, J.S., Andrew, M.A., Kovalenko, M., Murtha, R., Seong, I., Gusella, J.F., Kwak, S., Howland, D., Lee, R., Lee, J.-M., Wheeler, V.C., MacDonald, M.E., 2020. Promotion of somatic CAG repeat expansion by Fan1 knock-out in Huntington's disease knock-in mice is blocked by Mlh1 knock-out. Hum. Mol. Genet. ddaa196. https://doi.org/10.1093/hmg/ddaa196
Lu, H., Giordano, F., Ning, Z., 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 14, 265–279. https://doi.org/10.1016/j.gpb.2016.05.004
Lu, X.-H., Mattis, V.B., Wang, N., Al-Ramahi, I., van den Berg, N., Fratantoni, S.A., Waldvogel, H., Greiner, E., Osmand, A., Elzein, K., Xiao, J., Dijkstra, S., de Pril, R., Vinters, H.V., Faull, R., Signer, E., Kwak, S., Marugan, J.J., Botas, J., Fischer, D.F., Svendsen, C.N., Munoz-Sanjuan, I., Yang, X.W., 2014. Targeting ATM ameliorates mutant Huntingtin toxicity in cell and animal models of Huntington's disease. Sci. Transl. Med. 6, 268ra178-268ra178. https://doi.org/10.1126/scitranslmed.3010523
Lynch, D.R., Farmer, J., Hauser, L., Blair, I.A., Wang, Q.Q., Mesaros, C., Snyder, N., Boesch, S., Chin, M., Delatycki, M.B., Giunti, P., Goldsberry, A., Hoyle, C., McBride, M.G., Nachbauer, W., O'Grady, M., Perlman, S., Subramony, S.H., Wilmot, G.R., Zesiewicz, T., Meyer, C., 2019. Safety, pharmacodynamics, and potential benefit of omaveloxolone in Friedreich ataxia. Ann. Clin. Transl. Neurol. 6, 15–26. https://doi.org/10.1002/acn3.660
Lyon, E., Laver, T., Yu, P., Jama, M., Young, K., Zoccoli, M., Marlowe, N., 2010. A simple, high-throughput assay for Fragile X expanded alleles using triple repeat primed PCR and capillary electrophoresis. J. Mol. Diagn. JMD 12, 505–511. https://doi.org/10.2353/jmoldx.2010.090229
MacDonald, M.E., 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, 971–983.
Maiuri, T., Mocle, A.J., Hung, C.L., Xia, J., van Roon-Mom, W.M.C., Truant, R., 2016. Huntingtin is a scaffolding protein in the ATM oxidative DNA damage response complex. Hum. Mol. Genet. ddw395. https://doi.org/10.1093/hmg/ddw395
Manley, K., Shirley, T.L., Flaherty, L., Messer, A., 1999. Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat. Genet. 23, 471–473. https://doi.org/10.1038/70598
McDaniel, D.O., Keats, B., Vedanarayanan, V.V., Subramony, S.H., 2001. Sequence variation in GAA repeat expansions may cause differential phenotype display in Friedreich's ataxia. Mov. Disord. Off. J. Mov. Disord. Soc. 16, 1153–1158.

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. https://doi.org/10.1371/journal.pgen.1003648

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, 453–459. https://doi.org/10.1007/s00439-010-0873-9

Metzker, M.L., 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. 11, 31–46. https://doi.org/10.1038/nrg2626

Mirkin, S.M., 2007. Expandable DNA repeats and human disease. Nature 447, 932–940. https://doi.org/10.1038/nature05977

Monckton, D.G., Wong, L.-J.C., Ashizawa, T., Caskey, C.T., 1995. Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. Hum. Mol. Genet. 4, 1–8. https://doi.org/10.1093/hmg/4.1.1

Montermini, L., Andermann, E., Labuda, M., Richter, A., Pandolfo, M., Cavalcanti, F., Pianese, L., Iodice, L., Farina, G., Monticelli, A., Turano, M., Filla, A., De Michele, G., Cocozza, S., 1997. The Friedreich ataxia GAA triplet repeat: premutation and normal alleles. Hum. Mol. Genet. 6, 1261–1266.

Montermini, L., Rodius, F., Pianese, L., Moltò, M.D., Cossée, M., Campuzano, V., Cavalcanti, F., Monticelli, A., Palau, F., Gyapay, G., 1995. The Friedreich ataxia critical region spans a 150-kb interval on chromosome 9q13. Am. J. Hum. Genet. 57, 1061–1067.

Morales, F., Vásquez, M., Santamaría, 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 40, 57–66. https://doi.org/10.1016/j.dnarep.2016.01.001

Moss, D.J.H., Pardiñas, A.F., Langbehn, D., Lo, K., Leavitt, B.R., Roos, R., Durr, A., Mead, S., Holmans, P., Jones, L., Tabrizi, S.J., Coleman, A., Santos, R.D., Decolongon, J., Sturrock, A., Bardinet, E., Ret, C.J., Justo, D., Lehericy, S., Marelli, C., Nigaud, K., Valabrègue, R., van den Bogaard, S., Dumas, E M, van der Grond, J., t'Hart, E., Jurgens, C., Witjes-Ane, M.-N., Arran, N, Callaghan, J, Stopford, C, Frost, C., Jones, R., Hobbs, N., Lahiri, N, Ordidge, R., Owen, G., Pepple, T., Read, J., Say, M., Wild, E, Patel, A, Fox, N.C., Gibbard, C., Malone, I., Crawford, H., Whitehead, D., Keenan, S., Cash, D.M., Berna, C., Bechtel, N, Bohlen, S, Man, A.H., Kraus, P., Axelson, E., Wang, C., Acharya, T., Lee, S., Monaco, W., Campbell, C., Queller, S., Whitlock, K., Campbell, C., Campbell, M., Frajman, E., Milchman, C., O'Regan, A., Labuschagne, I., Stout, J., Landwehrmeyer, B, Craufurd, D, Scahill, R., Hicks, S., Kennard, C., Johnson, H., Tobin, A., Rosas, H., Reilmann, R, Borowsky, B., Pourchot, C., Andrews, S.C., Bachoud-Lévi, A.-C., Bentivoglio, A.R., Biunno, I., Bonelli, R., Burgunder, J.-M., Dunnett, S., Ferreira, J., Handley, O., Heiberg, A., Illmann, T., Landwehrmeyer, G.B., Levey, J., Ramos-Arroyo, M.A., Nielsen, J., Koivisto, S.P., Päivärinta, M., Roos, R.A.C., Sebastián, A Rojo, Tabrizi, S., Vandenberghe, W., Verellen-Dumoulin, C., Uhrova, T., Wahlström, J., Zaremba, J., Baake, V., Barth, K., Garde, M.B., Betz, S., Bos, R., Callaghan, Jenny, Come, A., Guedes, L.C., Ecker, D., Finisterra, A.M., Fullam, R., Gilling, M., Gustafsson, L., Handley, O.J., Hvalstedt, C., Held, C., Koppers, K., Lamanna, C., Laurà, M., Descals, A.M., Martinez-Horta, S., Mestre, T., Minster, S., Monza, D., Mütze, L., Oehmen, M., Orth, M., Padieu, H., Paterski, L., Peppa, N., Koivisto, S.P., Di Renzo, M., Rialland, A., Røren, N., Šašinková, P., Timewell, E., Townhill, J., Cubillo, P.T., da Silva, W.V., van Walsem, M.R., Whalstedt, C., Witjes-Ané, M.-N., Witkowski, G., Wright, A., Zielonka, D., Zielonka, E., Zinzi, P., Bonelli, R.M., Lilek, S., Hecht, K., Herranhof, B., Holl, A., Kapfhammer, H.-P., Koppitz, M., Magnet, M., Müller, N., Otti, D., Painold, A., Reisinger, K., Scheibl, M., Schöggl, H., Ullah, J., Braunwarth, E.-M., Brugger, F., Buratti, L., Hametner, E.-M., Hepperger, C., Holas, C., Hotter, A., Hussl, A., Müller, C., Poewe, W., Seppi, K., Sprenger, F., Wenning, G., Boogaerts, A., Calmeyn, G., Delvaux, I., Liessens, D., Somers, N.,

Dupuit, M., Minet, C., van Paemel, D., Ribaï, P., Verellen-Dumoulin, C., Boogaerts, A., Vandenberghe, W., van Reijen, D., Klempír, J., Majerová, V., Roth, J., Stárková, I., Hjermind, L.E., Jacobsen, O., Nielsen, J.E., Larsen, I.U., Vinther-Jensen, T., Hiivola, H., Hyppönen, H., Martikainen, K., Tuuha, K., Allain, P., Bonneau, D., Bost, M., Gohier, B., Guérid, M.-A., Olivier, A., Prundean, A., Scherer-Gagou, C., Verny, C., Babiloni, B., Debruxelles, S., Duché, C., Goizet, C., Jameau, L., Lafoucrière, D., Spampinato, U., Barthélémy, R., De Bruycker, C., Carette, M.C.A.-S., Defebvre, E.D.L., Delliaux, M., Delval, A., Destee, A., Dujardin, K., Lemaire, M.-H., Manouvrier, S., Peter, M., Plomhouse, L., Sablonnière, B., Simonin, C., Thibault-Tanchou, S., Vuillaume, I., Bellonet, M., Berrissoul, H., Blin, S., Courtin, F., Duru, C., Fasquel, V., Godefroy, O., Krystkowiak, P., Mantaux, B., Roussel, M., Wannepain, S., Azulay, J.-P., Delfini, M., Eusebio, A., Fluchere, F., Mundler, L., Anheim, M., Julié, C., Boukbiza, O.L., Longato, N., Rudolf, G., Tranchant, C., Zimmermann, M.-A., Kosinski, C.M., Milkereit, E., Probst, D., Reetz, K., Sass, C., Schiefer, J., Schlangen, C., Werner, C.J., Gelderblom, H., Priller, J., Prüß, H., Spruth, E.J., Ellrichmann, G., Herrmann, L., Hoffmann, R., Kaminski, B., Kotz, P., Prehn, C., Saft, C., Lange, H., Maiwald, R., Löhle, M., Maass, A., Schmidt, S., Bosredon, C., Storch, A., Wolz, A., Wolz, M., Capetian, P., Lambeck, J., Zucker, B., Boelmans, K., Ganos, C., Heinicke, W., Hidding, U., Lewerenz, J., Münchau, A., Orth, M., Schmalfeld, J., Stubbe, L., Zittel, S., Diercks, G., Dressler, D., Gorzolla, H., Schrader, C., Tacik, P., Ribbat, M., Longinus, B., Bürk, K., Möller, J.C., Rissling, I., Mühlau, M., Peinemann, A., Städtler, M., Weindl, A., Winkelmann, J., Ziegler, C., Bechtel, Natalie, Beckmann, H., Bohlen, Stefan, Hölzner, E., Lange, H., Reilmann, Ralf, Rohm, S., Rumpf, S., Schepers, S., Weber, N., Dose, M., Leythäuser, G., Marquard, R., Raab, T., Wiedemann, A., Barth, K., Buck, A., Connemann, J., Ecker, D., Geitner, C., Held, C., Kesse, A., Landwehrmeyer, Bernhard, Lang, C., Lewerenz, J., Lezius, F., Nepper, S., Niess, A., Orth, M., Schneider, A., Schwenk, D., Süßmuth, S., Trautmann, S., Weydt, P., Cormio, C., Sciruicchio, V., Serpino, C., de Tommaso, M., Capellari, S., Cortelli, P., Galassi, R., Rizzo, G., Poda, R., Scaglione, C., Bertini, E., Ghelli, E., Ginestroni, A., Massaro, F., Mechi, C., Paganini, M., Piacentini, S., Pradella, S., Romoli, A.M., Sorbi, S., Abbruzzese, G., di Poggio, M.B., Ferrandes, G., Mandich, P., Marchese, R., Albanese, A., Di Bella, D., Castaldo, A., Di Donato, S., Gellera, C., Genitrini, S., Mariotti, C., Monza, D., Nanetti, L., Paridi, D., Soliveri, P., Tomasello, C., De Michele, G., Di Maio, L., Massarelli, M., Peluso, S., Roca, A., Russo, C.V., Salvatore, E., Sorrentino, P., Amico, E., Favellato, M., Griguoli, A., Mazzante, I., Petrollini, M., Squitieri, F., D'Alessio, B., Esposito, C., Bentivoglio, R., Frontali, M., Guidubaldi, A., Ialongo, T., Jacopini, G., Piano, C., Romano, S., Soleti, F., Spadaro, M., Zinzi, P., van Hout, M.S.E., Verhoeven, M.E., van Vugt, J.P.P., de Weert, A.M., Bolwijn, J.J.W., Dekker, M., Kremer, B., Leenders, K.L., van Oostrom, J.C.H., van den Bogaard, S.J.A., Bos, R., Dumas, Eve M., 't Hart, E.P., Roos, R.A.C., Kremer, Berry, Verstappen, C.C.P., Aaserud, O., C, J.F., Heiberg, A., van Walsem, M.R., Wehus, R., Bjørgo, K., Fannemel, M., Gørvell, P.F., Lorentzen, E., Koivisto, S.P., Retterstøl, L., Stokke, B., Bjørnevoll, I., Sando, S.B., Dziadkiewicz, A., Nowak, M., Robowski, P., Sitek, E., Slawek, J., Soltan, W., Szinwelski, M., Blaszcyk, M., Boczarska-Jedynak, M., Ciach-Wysocka, E., Gorzkowska, A., Jasinska-Myga, B., Klodowska-Duda, G., Opala, G., Stompel, D., Banaszkiewicz, K., Bocwinska, D., Bojakowska-Jaremek, K., Dec, M., Krawczyk, M., Rudzinska, M., Szczygiel, E., Szczudlik, A., Wasielewska, A., Wójcik, M., Bryl, A., Ciesielska, A., Klimberg, A., Marcinkowski, J., Samara, H., Sempolowicz, J., Zielonka, D., Gogol, A., Janik, P., Kwiecinski, H., Jamrozik, Z., Antczak, J., Jachinska, K., Krysa, W., Rakowicz, M., Richter, P., Rola, R., Ryglewicz, D., Sienkiewicz-Jarosz, H., Stepniak, I., Sulek, A., Witkowski, G., Zaremba, J., Zdzienicka, E., Zieora-Jakutowicz, K., Ferreira, J.J., Coelho, M., Guedes, L.C., Mendes, T., Mestre, T., Valadas, A., Andrade, C., Gago, M., Garrett, C., Guerra, M.R., Herrera, C.D., Garcia, P.M., Barbera, M.A., Guia, D.B., Hernanz, L.C., Catena, J.L., Ferrer, P.Q., Sebastián, Ana Rojo, Carruesco, G.T., Bas, J., Busquets, N., Calopa, M., Robert, M.F., Viladrich, C.M., Idiago, J.M.R., Riballo, A.V., Cubo, E., Polo, C.G., Mariscal, N., Rivadeneyra, P.J., Barrero, F., Morales, B., Fenollar, M., García, R.G.-R., Ortega, P., Villanueva, C., Alegre, J., Bascuñana, M., Caldentey,

J.G., Ventura, M.F., Ribas, G.G., de Yébenes, J.G., Moreno, J.L.L.-S., Cubillo, P.T., Alegre, J., Frech, F.A., de Yébenes, J.G., Ruíz, P.J.G., Martínez-Descals, A., Guerrero, R., Artiga, M.J.S., Sánchez, V., Perea, M.F.N., Fortuna, L., Manzanares, S., Reinante, G., Torres, M.M.A., Moreau, L.V., González González, S., Guisasola, L.M., Salvador, C., Martín, E.S.S., Ramirez, I.L., Gorospe, A., Lopera, M.R., Arques, P.N., Rodríguez, M.J.T., Pastor, B.V., Gaston, I., Martinez-Jaurrieta, M.D., Ramos-Arroyo, M.A., Moreno, J.M.G., Lucena, C.M., Damas, F., Cortegana, H.E.P., Peña, J.C., Redondo, L., Carrillo, F., Teresa Cáceres, M., Mir, P., Suarez, M.J.L., Vargas-González, L., Bosca, M.E., Brugada, F.C., Burguera, J.A., Campos, A., Vilaplana, G.C.P., Berglund, P., Constantinescu, R., Fredlund, G., Høsterey-Ugander, U., Linnsand, P., Neleborn-Lingefjärd, L., Wahlström, J., Wentzel, M., Loutfi, G., Olofsson, C., Stattin, E.-L., Westman, L., Wikström, B., Burgunder, J.-M., Stebler, Y., Kaelin, A., Romero, I., Schüpbach, M., Weber Zaugg, S., Hauer, M., Gonzenbach, R., Jung, H.H., Mihaylova, V., Petersen, J., Jack, R., Matheson, K., Miedzybrodzka, Z., Rae, D., Simpson, S.A., Summers, F., Ure, A., Vaughan, V., Akhtar, S., Crooks, J., Curtis, A., de Souza, J., Piedad, J., Rickards, H., Wright, J., Coulthard, E., Gethin, L., Hayward, B., Sieradzan, K., Wright, A., Armstrong, M., Barker, R.A., O'Keefe, D., Di Pietro, A., Fisher, K., Goodman, A., Hill, S., Kershaw, A., Mason, S., Paterson, N., Raymond, L., Swain, R., Guzman, N.V., Busse, M., Butcher, C., Callaghan, Jenny, Dunnett, S., Clenaghan, C., Fullam, R., Handley, O., Hunt, S., Jones, L., Jones, U., Khalil, H., Minster, S., Owen, M., Price, K., Rosser, A., Townhill, J., Edwards, M., Ho, C., Hughes, T., McGill, M., Pearson, P., Porteous, M., Smith, P., Brockie, P., Foster, J., Johns, N., McKenzie, S., Rothery, J., Thomas, G., Yates, S., Burrows, L., Chu, C., Fletcher, A., Gallantrae, D., Hamer, S., Harding, A., Klöppel, S., Kraus, A., Laver, F., Lewis, M., Longthorpe, M., Markova, I., Raman, A., Robertson, N., Silva, M., Thomson, A., Wild, S., Yardumian, P., Chu, C., Evans, C., Gallentrae, D., Hamer, S., Kraus, A., Markova, I., Raman, A., Chu, C., Hamer, S., Hobson, E., Jamieson, S., Kraus, A., Markova, I., Raman, A., Musgrave, H., Rowett, L., Toscano, J., Wild, S., Yardumian, P., Bourne, C., Clapton, J., Clayton, C., Dipple, H., Freire-Patino, D., Grant, J., Gross, D., Hallam, C., Middleton, J., Murch, A., Thompson, C., Alusi, S., Davies, R., Foy, K., Gerrans, E., Pate, L., Andrews, T., Dougherty, A., Golding, C., Kavalier, F., Laing, H., Lashwood, A., Robertson, D., Ruddy, D., Santhouse, A., Whaite, A., Andrews, T., Bruno, S., Doherty, K., Golding, C., Haider, S., Hensman, D., Lahiri, Nayana, Lewis, M., Novak, M., Patel, Aakta, Robertson, N., Rosser, E., Tabrizi, S., Taylor, R., Warner, T., Wild, Edward, Arran, Natalie, Bek, J., Callaghan, Jenny, Craufurd, David, Fullam, R., Hare, M., Howard, L., Huson, S., Johnson, L., Jones, M., Murphy, H., Oughton, E., Partington-Jones, L., Rogers, D., Sollom, A., Snowden, J., Stopford, Cheryl, Thompson, J., Trender-Gerhard, I., Verstraelen, N., Westmoreland, L., Armstrong, R., Dixon, K., Nemeth, A.H., Siuda, G., Valentine, R., Harrison, D., Hughes, M., Parkinson, A., Soltysiak, B., Bandmann, O., Bradbury, A., Gill, P., Fairtlough, H., Fillingham, K., Foustanos, I., Kazoka, M., O'Donovan, K., Peppa, N., Taylor, C., Tidswell, K., Quarrell, O., Burgunder, J.-M., Lau, P.N., Pica, E., Tan, L., 2017. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. Lancet Neurol. 16, 701–711. https://doi.org/10.1016/S1474-4422(17)30161-8

Myers, T.A., Chanock, S.J., Machiela, M.J., 2020. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. Front. Genet. 11, 157. https://doi.org/10.3389/fgene.2020.00157

Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., Shinzato, M., Minami, M., Nakanishi, T., Teruya, K., Satou, K., Hirano, T., 2017. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum. Cell 30, 149–161. https://doi.org/10.1007/s13577-017-0168-8

Nethisinghe, S., Pigazzini, M.L., Pemble, S., Sweeney, M.G., Labrum, R., Manso, K., Moore, D., Warner, J., Davis, M.B., Giunti, P., 2018. PolyQ Tract Toxicity in SCA1 is Length Dependent in the Absence of CAG Repeat Interruption. Front. Cell. Neurosci. 12. https://doi.org/10.3389/fncel.2018.00200

Oh, K.-S., Imoto, K., Boyle, J., Khan, S.G., Kraemer, K.H., 2007. Influence of XPB helicase on recruitment and redistribution of nucleotide excision repair proteins at sites of UV-induced DNA damage. DNA Repair 6, 1359–1370. https://doi.org/10.1016/j.dnarep.2007.03.025

Ohshima, K., Montermini, L., Wells, R.D., Pandolfo, M., 1998. Inhibitory effects of expanded GAA.TTC triplet repeats from intron I of the Friedreich ataxia gene on transcription and replication in vivo. J. Biol. Chem. 273, 14588–14595.

Ohshima, K., Sakamoto, N., Labuda, M., Poirier, J., Moseley, M.L., Montermini, L., Ranum, L.P., Wells, R.D., Pandolfo, M., 1999. A nonpathogenic GAAGGA repeat in the Friedreich gene: implications for pathogenesis. Neurology 53, 1854–1857.

Ooi, J., Langley, S.R., Xu, X., Utami, K.H., Sim, B., Huang, Y., Harmston, N.P., Tay, Y.L., Ziaei, A., Zeng, R., Low, D., Aminkeng, F., Sobota, R.M., Ginhoux, F., Petretto, E., Pouladi, M.A., 2019. Unbiased Profiling of Isogenic Huntington Disease hPSC-Derived CNS and Peripheral Cells Reveals Strong Cell-Type Specificity of CAG Length Effects. Cell Rep. 26, 2494-2508.e7. https://doi.org/10.1016/j.celrep.2019.02.008

Pandolfo, M., 2020. Neurologic outcomes in Friedreich ataxia: Study of a single-site cohort. Neurol. Genet. 6, e415. https://doi.org/10.1212/NXG.0000000000000415

Pandolfo, M., 2009. Friedreich ataxia: The clinical picture. J. Neurol. 256, 3–8. https://doi.org/10.1007/s00415-009-1002-3

Parkinson, M.H., Boesch, S., Nachbauer, W., Mariotti, C., Giunti, P., 2013. Clinical features of Friedreich's ataxia: classical and atypical phenotypes. J. Neurochem. 126, 103–117. https://doi.org/10.1111/jnc.12317

Pêcheux, C., Mouret, J.F., Dürr, A., Agid, Y., Feingold, J., Brice, A., Dodé, C., Kaplan, J.C., 1995. Sequence analysis of the CCG polymorphic region adjacent to the CAG triplet repeat of the HD gene in normal and HD chromosomes. J. Med. Genet. 32, 399–400.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L., Mayer, G., 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci. Rep. 8, 10950. https://doi.org/10.1038/s41598-018-29325-6

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. https://doi.org/10.1371/journal.pgen.1003930

Poirier, J., Ohshima, K., Pandolfo, M., 1999. Heteroduplexes may confuse the interpretation of PCR-based molecular tests for the Friedreich ataxia GAA triplet repeat. Hum. Mutat. 13, 328–330. https://doi.org/10.1002/(SICI)1098-1004(1999)13:4<328::AID-HUMU10>3.0.CO;2-J

Pollard, L.M., 2004. Replication-mediated instability of the GAA triplet repeat mutation in Friedreich ataxia. Nucleic Acids Res. 32, 5962–5971. https://doi.org/10.1093/nar/gkh933

Potapov, V., Ong, J.L., 2017. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLOS ONE 12, e0169774. https://doi.org/10.1371/journal.pone.0169774

Quan, F., Janas, J., Popovich, B.W., 1995. A novel CAG repeat configuration in the SCA1 gene: implications for the molecular diagnostics of spinocerebellar ataxia type 1. Hum. Mol. Genet. 4, 2411–2413.

Quarrell, O.W., Handley, O., O'Donovan, K., Dumoulin, C., Ramos-Arroyo, M., Biunno, I., Bauer, P., Kline, M., Landwehrmeyer, G.B., European Huntington's Disease Network, 2012. Discrepancies in reporting the CAG repeat lengths for Huntington's disease. Eur. J. Hum. Genet. EJHG 20, 20–26. https://doi.org/10.1038/ejhg.2011.136

Quarrell, O.W., Nance, M.A., Nopoulos, P., Paulsen, J.S., Smith, J.A., Squitieri, F., 2013. Managing juvenile Huntington's disease. Neurodegener. Dis. Manag. 3, 267–276. https://doi.org/10.2217/nmt.13.18

Quigley, J., 2017. Juvenile Huntington's Disease: Diagnostic and Treatment Considerations for the Psychiatrist. Curr. Psychiatry Rep. 19, 9. https://doi.org/10.1007/s11920-017-0759-9

Ragno, M., De Michele, G., Cavalcanti, F., Pianese, L., Monticelli, A., Curatola, L., Bollettini, F., Cocozza, S., Caruso, G., Santoro, L., Filla, A., 1997. Broadened Friedreich's ataxia

phenotype after gene cloning: Minimal GAA expansion causes late-onset spastic ataxia. Neurology 49, 1617–1620. https://doi.org/10.1212/WNL.49.6.1617
Rang, F.J., Kloosterman, W.P., de Ridder, J., 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 19, 90. https://doi.org/10.1186/s13059-018-1462-9
Reetz, K., Dogan, I., Costa, A.S., Dafotakis, M., Fedosov, K., Giunti, P., Parkinson, M.H., Sweeney, M.G., Mariotti, C., Panzeri, M., Nanetti, L., Arpa, J., Sanz-Gallego, I., Durr, A., Charles, P., Boesch, S., Nachbauer, W., Klopstock, T., Karin, I., Depondt, C., vom Hagen, J.M., Schöls, L., Giordano, I.A., Klockgether, T., Bürk, K., Pandolfo, M., Schulz, J.B., 2015. Biological and clinical characteristics of the European Friedreich's Ataxia Consortium for Translational Studies (EFACTS) cohort: a cross-sectional analysis of baseline data. Lancet Neurol. 14, 174–182. https://doi.org/10.1016/S1474-4422(14)70321-7
Reiner, A., Dragatsis, I., Dietrich, P., 2011. Genetics and neuropathology of Huntington's disease. Int. Rev. Neurobiol. 98, 325–372. https://doi.org/10.1016/B978-0-12-381328-2.00014-6
Reisman, S.A., Gahir, S.S., Lee, C.-Y.I., Proksch, J.W., Sakamoto, M., Ward, K.W., 2019. Pharmacokinetics and pharmacodynamics of the novel Nrf2 activator omaveloxolone in primates. Drug Des. Devel. Ther. Volume 13, 1259–1270. https://doi.org/10.2147/DDDT.S193889
Roberts, R.J., Carneiro, M.O., Schatz, M.C., 2013. The advantages of SMRT sequencing. Genome Biol. 14. https://doi.org/10.1186/gb-2013-14-7-405
Rosas, H.D., Salat, D.H., Lee, S.Y., Zaleta, A.K., Pappu, V., Fischl, B., Greve, D., Hevelone, N., Hersch, S.M., 2008. Cerebral cortex and the clinical expression of Huntington's disease: complexity and heterogeneity. Brain 131, 1057–1068. https://doi.org/10.1093/brain/awn025
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, 204–216. https://doi.org/10.1038/nrneurol.2014.24
Rüb, U., Seidel, K., Heinsen, H., Vonsattel, J.P., den Dunnen, W.F., Korf, H.W., 2016. Huntington's disease (HD): the neuropathology of a multisystem neurodegenerative disorder of the human brain: The brain in Huntington's disease. Brain Pathol. 26, 726–740. https://doi.org/10.1111/bpa.12426
Rubinsztein, D.C., Leggo, J., Coles, R., Almqvist, E., Biancalana, V., Cassiman, J.J., Chotai, K., Connarty, M., Crauford, D., Curtis, A., Curtis, D., Davidson, M.J., Differ, A.M., Dode, C., Dodge, A., Frontali, M., Ranen, N.G., Stine, O.C., Sherr, M., Abbott, M.H., Franz, M.L., Graham, C.A., Harper, P.S., Hedreen, J.C., Hayden, M.R., 1996. Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. Am. J. Hum. Genet. 59, 16–22.
Saccà, F., Puorro, G., Antenora, A., Marsili, A., Denaro, A., Piro, R., Sorrentino, P., Pane, C., Tessa, A., Brescia Morra, V., Cocozza, S., De Michele, G., Santorelli, F.M., Filla, A., 2011. A combined nucleic acid and protein analysis in Friedreich ataxia: implications for diagnosis, pathogenesis and clinical trial design. PloS One 6, e17627. https://doi.org/10.1371/journal.pone.0017627
Sakamoto, N., Larson, J.E., Iyer, R.R., Montermini, L., Pandolfo, M., Wells, R.D., 2001. GGA·TCC-interrupted Triplets in Long GAA·TTC Repeats Inhibit the Formation of Triplex and Sticky DNA Structures, Alleviate Transcription Inhibition, and Reduce Genetic Instabilities. J. Biol. Chem. 276, 27178–27187. https://doi.org/10.1074/jbc.M101852200
Sakamoto, N., Ohshima, K., Montermini, L., Pandolfo, M., Wells, R.D., 2001b. Sticky DNA, a Self-associated Complex Formed at Long GAA·TTC Repeats in Intron 1 of the Frataxin Gene, Inhibits Transcription. J. Biol. Chem. 276, 27171–27177. https://doi.org/10.1074/jbc.M101879200

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74, 5463–5467.
Santoro, M., Perna, A., La Rosa, P., Petrillo, S., Piemonte, F., Rossi, S., Riso, V., Nicoletti, T.F., Modoni, A., Pomponi, M.G., Chiurazzi, P., Silvestri, G., 2020. Compound heterozygosity for an expanded (GAA) and a (GAAGGA) repeat at FXN locus: from a diagnostic pitfall to potential clues to the pathogenesis of Friedreich ataxia. neurogenetics 21, 279–287. https://doi.org/10.1007/s10048-020-00620-7
Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19, R227–R240. https://doi.org/10.1093/hmg/ddq416
Schmucker, S., Puccio, H., 2010. Understanding the molecular mechanisms of Friedreich's ataxia to develop therapeutic approaches. Hum. Mol. Genet. 19, R103–R110. https://doi.org/10.1093/hmg/ddq165
Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., Schatz, M.C., 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods 15, 461–468. https://doi.org/10.1038/s41592-018-0001-7
Seo, J.-S., Rhie, A., Kim, Junsoo, Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, Jihye, Kuk, J., Park, G.H., Kim, Juhyeok, Ryu, H., Kim, Jongbum, Roh, M., Baek, J., Hunkapiller, M.W., Korlach, J., Shin, J.-Y., Kim, C., 2016. De novo assembly and phasing of a Korean human genome. Nature 538, 243–247. https://doi.org/10.1038/nature20098
Sharma, R., De Biase, I., Gómez, M., Delatycki, M.B., Ashizawa, T., Bidichandani, S.I., 2004. Friedreich ataxia in carriers of unstable borderline GAA triplet-repeat alleles: FRDA Unstable Borderline Alleles. Ann. Neurol. 56, 898–901. https://doi.org/10.1002/ana.20333
Shelbourne, P.F., Keller-McGandy, C., Bi, W.L., Yoon, S.-R., Dubeau, L., Veitch, N.J., Vonsattel, J.P., Wexler, N.S., US-Venezuela Collaborative Research Group, Arnheim, N., Augood, S.J., 2007. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Hum. Mol. Genet. 16, 1133–1142. https://doi.org/10.1093/hmg/ddm054
Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., Lintner, K.E., Ding, Q., Wang, Z., Hu, J., Wang, D., Wang, F., Wang, L., Lyon, G.J., Guan, Y., Shen, Y., Evgrafov, O.V., Knowles, J.A., Thibaud-Nissen, F., Schneider, V., Yu, C.-Y., Zhou, L., Eichler, E.E., So, K.-F., Wang, K., 2016. Long-read sequencing and de novo assembly of a Chinese genome. Nat. Commun. 7, 12065. https://doi.org/10.1038/ncomms12065
Stolle, C.A., Frackelton, E.C., McCallum, J., Farmer, J.M., Tsou, A., Wilson, R.B., Lynch, D.R., 2008. Novel, complex interruptions of the GAA repeat in small, expanded alleles of two affected siblings with late-onset Friedreich ataxia. Mov. Disord. 23, 1303–1306. https://doi.org/10.1002/mds.22012
Stuitje, G., van Belzen, M.J., Gardiner, S.L., van Roon-Mom, W.M.C., Boogaard, M.W., Tabrizi, S.J., Roos, R.A.C., Aziz, N.A., 2017. Age of onset in Huntington's disease is influenced by CAG repeat variations in other polyglutamine disease-associated genes. Brain 140, e42–e42. https://doi.org/10.1093/brain/awx122
Sun, Y.-M., Zhang, Y.-B., Wu, Z.-Y., 2017. Huntington's Disease: Relationship Between Phenotype and Genotype. Mol. Neurobiol. 54, 342–348. https://doi.org/10.1007/s12035-015-9662-8
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, 3039–3047. https://doi.org/10.1093/hmg/ddp242
Tabrizi, S.J., Leavitt, B.R., Landwehrmeyer, G.B., Wild, E.J., Saft, C., Barker, R.A., Blair, N.F., Craufurd, D., Priller, J., Rickards, H., Rosser, A., Kordasiewicz, H.B., Czech, C., Swayze, E.E., Norris, D.A., Baumann, T., Gerlach, I., Schobel, S.A., Paz, E., Smith, A.V., Bennett, C.F., Lane, R.M., 2019. Targeting Huntingtin Expression in Patients with Huntington's Disease. N. Engl. J. Med. 380, 2307–2316. https://doi.org/10.1056/NEJMoa1900907

Team, R. C. 2013. *R: A language and environment for statistical computing* [Online]. R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org/

Telenius, H., Kremer, B., Goldberg, Y.P., Theilmann, J., Andrew, S.E., Zeisler, J., Adam, S., Greenberg, C., Ives, E.J., Clarke, L.A., 1994. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. Nat. Genet. 6, 409–414. https://doi.org/10.1038/ng0494-409

Tezenas du Montcel, S., Durr, A., Bauer, P., Figueroa, K.P., Ichikawa, Y., Brussino, A., Forlani, S., Rakowicz, M., Schöls, L., Mariotti, C., van de Warrenburg, B.P.C., 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), EUROSCA network, Goto, J., Ashizawa, T., Cazeneuve, C., Tsuji, S., Pulst, S.-M., Brusco, A., Riess, O., Brice, A., Stevanin, G., 2014. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. Brain J. Neurol. 137, 2444–2455. https://doi.org/10.1093/brain/awu174

Thu, D.C.V., Oorschot, D.E., Tippett, L.J., Nana, A.L., Hogg, V.M., Synek, B.J., Luthi-Carter, R., Waldvogel, H.J., Faull, R.L.M., 2010. Cell loss in the motor and cingulate cortex correlates with symptomatology in Huntington's disease. Brain J. Neurol. 133, 1094–1110. https://doi.org/10.1093/brain/awq047

Tippett, L.J., Waldvogel, H.J., Thomas, S.J., Hogg, V.M., van Roon-Mom, W., Synek, B.J., Graybiel, A.M., Faull, R.L.M., 2007. Striosomes and mood dysfunction in Huntington's disease. Brain J. Neurol. 130, 206–221. https://doi.org/10.1093/brain/awl243

Tomé, S., Simard, J.P., Slean, M.M., Holt, I., Morris, G.E., Wojciechowicz, K., te Riele, H., Pearson, C.E., 2013. Tissue-specific mismatch repair protein expression: MSH3 is higher than MSH6 in multiple mouse tissues. DNA Repair 12, 46–52. https://doi.org/10.1016/j.dnarep.2012.10.006

Tsai, Y.-C., Greenberg, D., Powell, J., Hoijer, I., Ameur, A., Strahl, M., Ellis, E., Jonasson, I., Mouro Pinto, R., Wheeler, V., Smith, M.L., Gyllensten, U., Sebra, R., Korlach, J., Clark, T.A., 2017. Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions. https://doi.org/10.1101/203919

van den Broek, W.J.A.A., Nelen, M.R., Wansink, D.G., Coerwinkel, M.M., te Riele, H., Groenen, P.J.T.A., 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. Genet. 11, 191–198.

van den Ouweland, A.M.W., van Minkelen, R., Bolman, G.M., Wouters, C.H., Becht-Noordermeer, C., Deelen, W.H., Deelen-Manders, J.M.C., Ippel, E.P.F., Saris, J., Halley, D.J.J., 2012. Complete FXN deletion in a patient with Friedreich's ataxia. Genet. Test. Mol. Biomark. 16, 1015–1018. https://doi.org/10.1089/gtmb.2012.0012

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., Thermes, C., 2018. The Third Revolution in Sequencing Technology. Trends Genet. TIG. https://doi.org/10.1016/j.tig.2018.05.008

Vankan, P., 2013. Prevalence gradients of Friedreich's Ataxia and R1b haplotype in Europe co-localize, suggesting a common Palaeolithic origin in the Franco-Cantabrian ice age refuge. J. Neurochem. 126, 11–20. https://doi.org/10.1111/jnc.12215

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 6, 789–796. https://doi.org/10.1016/j.dnarep.2007.01.002

Vonsattel, J.P., Myers, R.H., Stevens, T.J., Ferrante, R.J., Bird, E.D., Richardson, E.P., 1985. Neuropathological classification of Huntington's disease. J. Neuropathol. Exp. Neurol. 44, 559–577.

Vonsattel, J.P.G., Keller, C., Cortes Ramirez, E.P., 2011. Huntington's disease – neuropathology, in: Handbook of Clinical Neurology. Elsevier, pp. 83–100. https://doi.org/10.1016/B978-0-444-52014-2.00004-5

Warner, J.P., Barron, L.H., Goudie, D., Kelly, K., Dow, D., Fitzpatrick, D.R., Brock, D.J., 1996. A general method for the detection of large CAG repeat expansions by fluorescent PCR. J. Med. Genet. 33, 1022–1026.

Weeda, G., Eveno, E., Donker, I., Vermeulen, W., Chevallier-Lagente, O., Taïeb, A., Stary, A., Hoeijmakers, J.H., Mezzina, M., Sarasin, A., 1997. A mutation in the XPB/ERCC3 DNA repair transcription gene, associated with trichothiodystrophy. Am. J. Hum. Genet. 60, 320–329.

Wexler, N.S., Lorimer, J., Porter, J., Gomez, F., Moskowitz, C., Shackell, E., Marder, K., Penchaszadeh, G., Roberts, S.A., Gayán, 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., Landwehrmeyer, B., U.S.-Venezuela Collaborative Research Project, 2004. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proc. Natl. Acad. Sci. U. S. A. 101, 3498–3503. https://doi.org/10.1073/pnas.0308679101

Weydt, P., Soyal, S.M., Gellera, C., DiDonato, S., Weidinger, C., Oberkofler, H., Landwehrmeyer, G.B., Patsch, W., 2009. The gene coding for PGC-1α modifies age at onset in Huntington's Disease. Mol. Neurodegener. 4, 3. https://doi.org/10.1186/1750-1326-4-3

Wheeler, V.C., Lebel, L.-A., Vrbanac, V., Teed, A., te Riele, H., MacDonald, M.E., 2003. Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. Hum. Mol. Genet. 12, 273–281.

Wild, E.J., Tabrizi, S.J., 2017. Therapies targeting DNA and RNA in Huntington's disease. Lancet Neurol. 16, 837–847. https://doi.org/10.1016/S1474-4422(17)30280-6

Wilson, D.J., 2019. The harmonic mean *p* -value for combining dependent tests. Proc. Natl. Acad. Sci. 116, 1195–1200. https://doi.org/10.1073/pnas.1814092116

Wright, G.E.B., Collins, J.A., Kay, C., McDonald, C., Dolzhenko, E., Xia, Q., Bečanović, K., Semaka, A., Nguyen, C.M., Trost, B., Richards, F., Bijlsma, E.K., Squitieri, F., Scherer, S.W., Eberle, M.A., Yuen, R.K.C., Hayden, M.R., 2019. Length of uninterrupted CAG repeats, independent of polyglutamine size, results in increased somatic instability and hastened age of onset in Huntington disease. bioRxiv. https://doi.org/10.1101/533414

Xu, E., Tang, Y., Li, D., Jia, J., 2009. Polymorphism of HD and UCHL-1 genes in Huntington's disease. J. Clin. Neurosci. Off. J. Neurosurg. Soc. Australas. 16, 1473–1477. https://doi.org/10.1016/j.jocn.2009.03.027

Yang, Y.-I., Jung, D.-W., Bai, D.-G., Yoo, G.-S., Choi, J.-K., 2001. Counterion-dye staining method for DNA in agarose gels using crystal violet and methyl orange. ELECTROPHORESIS 22, 855–859. https://doi.org/10.1002/1522-2683()22:5<855::AID-ELPS855>3.0.CO;2-Y

Yuan, F., Song, L., Liu, F., Gu, L., Zhang, Y., 2012. Eukaryotic DNA Mismatch Repair In Vitro, in: Bjergbæk, L. (Ed.), DNA Repair Protocols. Humana Press, Totowa, NJ, pp. 149–162.

Zhao, J., Jain, A., Iyer, R.R., Modrich, P.L., Vasquez, K.M., 2009. Mismatch repair and nucleotide excision repair proteins cooperate in the recognition of DNA interstrand crosslinks. Nucleic Acids Res. 37, 4420–4429. https://doi.org/10.1093/nar/gkp399

Zhao, X.-N., Usdin, K., 2018. FAN1 protects against repeat expansions in a Fragile X mouse model. DNA Repair 69, 1–5. https://doi.org/10.1016/j.dnarep.2018.07.001

Zhou, A., Lin, T., Xing, J., 2019. Evaluating nanopore sequencing data processing pipelines for structural variation identification. Genome Biol. 20, 237. https://doi.org/10.1186/s13059-019-1858-1

Supplementary Data

Table 1. FRDA patient and carrier summary

	Code	GAA 1/2 sizes	AAO	MboII bands	Extra band	MboII comments	TP-PCR comments
BRUNEL Samples							
1	FA1	1023/1258	8	✓			Double peaks until peak 10
2	FA11	720/760	8	✓			Pure
3	FA15	WTC/+		✓			Decrease after 14 peaks, double peaks
4	FA16	WTC/720		✓			Pure
5	FA12	WTC/520		✓			Gap from peak 4-9 after 3 GAAs
6	FA13	10/10		✓			**23 GAA**
7	FA14	WT/500	7	✓			Increased intensity in first 3 peaks, decreased intensity from peak 4-9, double peak at 9-13
8	FA17	720 /720	22	✓			Starting 3 GAA late, double peak at GAA 5 until the end
9	FA18	500/720	25	✓			Starting 7 GAA late
10	FA19	WTC/900		✓			Pure
11	FA20	WTC/+		✓			Pure
12	SCA121	730/1040	9	✓			Double peak from GAA 3
13	FA31	750/900	10	✓			Double peak from 1-9 GAA
14	FA35	630/730	12	✓			Double peak from GAA 6
15	FA36	630/1040	13		✓	high band at 1kb	Double peak at GAA 6
16	FA47	680/840	7	✓			Starting 6 GAA late
17	FA49	763/1043	23	✓			Pure
18	FA53	850/1000	3		✓	stutter	Pure
19	FA61	WTC/+		✓			Pure
20	FA62	+/+	8	✓			Pure
21	FA63	567/752	17	✓			Starting 2 GAA late, double peak from start
22	FA64	+/+	17	✓			Starting 2 GAA late
23	FA66	500/730	13	✓			Pure
24	FA 75	+/+	8	✓			Pure
25	FA 76	+/+	14	✓			Pure
26	FA 77	+/+	20	✓			Double peak at GAA 15
27	FA78	765/765	34		✓	High band	Starting 3 GAA late
28	FA79	460/765	44	✓			Double peak at GAA 5
29	FA85	+/+	18	✓			Pure
30	FA88	765/765	32		✓	high band at 1kb	Decreased intensity
31	FA90	765/1100	16	✓			Pure
32	FA96	430/1245	30	✓			Increased intensity at GAA 4 and 5, increased intensity at GAA 14
33	FA98	1250/1465	8	✓			Double peak at GAA 9

34	FA102	+/+	11	✓			Double peak starting at GAA 6
35	FA103	550/10		✓			**3 GAA**
36	FA104	782/782	20	✓			Starting 2 GAA late, increasing to GAA 9 and decreasing
37	FA106	1040/1040	6	✓			Pure
38	FA107	163/600	35	✓			Double peak at GAA 10-16
39	FA108	WTC/1000			✓	3 bands 8/12/15 19/1/16	Increased intensity at GAA 3
40	FA109	760/890	20	✓			6 GAA gap after GAA 2, double peak at GAA 9
41	FA110	890/890	15	✓			6 GAA gap after GAA 2
42	FA113	1045/1045	6	✓			Increased intensity at GAA 3
43	FA114	765/1065	17	✓			Double peak starting from GAA 17
44	FA115	112/940	45		✓	80bp insert	Increased intensity starting at GAA 27, double peaks at GAA 22-28 and GAA 36
45	FA121	905/965	7	✓			Starting 2 GAA late
46	FA123	700/1040	16	✓			Pure
47	FA131	900/1300	12	✓			Starting 3 GAA earlier, double peak starting from GAA 8
48	FA132	WT/330	65			extra band - ? HTX	2 GAA gap after GAA 3
49	FA134	930/930	7	✓			Increased intensity starting at GAA3
50	FA142	350/750	5	✓			Increased intensity starting at GAA3
51	FA150	500/800	30	✓			Double peak starting at GAA 9
52	FA152	1070/1460	10				Pure
53	FA153	400/1000	25	✓			Increased intensity starting at GAA3, double peak starting at GAA 21
54	FA154	633/760	24		✓	stutter	Double peak starting at GAA 12
55	FA156	740/1200	17	✓			Double peak starting at GAA 15
56	FA163	700/1000	11	✓			Increased intensity starting at GAA 5, double peak at GAA 24
57	FA164	108/1040	45	✓			Decreased intensity at GAA 17, double peak at GAA 24
58	FA165	765/1045	17	✓			Starting 5 GAA late
59	FA167	1000/1000	6	✓			Pure
60	FA173	230/10		✓			**15 GAA**
61	FA174	1000/1000	6				Pure
62	FA176	780/780	20				Starting 7 GAA late
63	FA178	249/559	35	✓			Increased intensity at GAA 4-5, GAA 14

64	FA179	906/906	10	✓			Pure
65	FA181	536/809	13	✓			Gap from GAA 4-9
66	FA188	WT/1180					**19 GAA**
67	FA191	150/573	74	✓			Gap from GAA 2 -8, double peak starting from GAA 21
68	FA195	77/127	51	✓			Starting 3 GAA late, Gap from GAA 11-22
69	FA196	328/1194	39	✓			Increased intensity at GAA 18, double peak from GAA 8-11
70	FA197	478/1257	30	✓			Double peak starting at GAA 30
71	SCA70	358/358	45	✓			Increased intensity at GAA 2,3 and GAA 12
72	SCA211	160/1040	31				**FAIL**
73	SCA321	696/800	4	✓			Double peak starting from GAA 8
74	SCA322	800/1013	20		✓	Extra small band	3 GAA late, double peak starting from GAA 3
75	SCA372	766/1046	25	✓			Starting 1 GAA late
76	SCA380	390/390	29	✓			Decreased intensity at GAA 6, double peak starting at GAA 12
77	SCA502	WT/+	63	✓			Gap at GAA 3-5
78	SCA596	WT/+		✓			Gap at GAA 3-5, double peak from GAA 1
79	SCA597	765/1045	15				Starting 2 GAA late
80	SCA612	+/+	36				Double peak from GAA 3-5, increased intensity at GAA 10, double peak from GAA 18
81	SCA671	485/485	32	✓			Starting 2 GAA early
82	SCA694	800/800	27	✓			Pure
83	SCA743	483/905	35	✓			Pure
84	SCA814	+/+	20	✓			Double peak from GAA 21, hedgehog effect
85	SCA922	765/1100	20	✓			Pure
86	SCA937	700/1000	22	✓			Starting 4 GAA late
87	SCA1013	400/400	18	✓			Double peak from GAA 10, increased intensity at GAA 3
88	SCA1120	+/WT			✓	2 extra bands ? HTX	**20 GAA**
89	SCA1305	1010/1207	25	✓			Starting 1 GAA late, double peak from GAA 6
90	SCA1306	1085/1165	20	✓			Starting 1 GAA late, double peak from GAA 6
91	SCA1311	1100/1400	8	✓			Starting 1 GAA late
92	SCA1404	347/1301	30	✓			Decreased intensity at GAA 7, double peak starting at GAA 14

UCL Samples

1	6336	200/200	26	✓			Double peak starting at GAA 5, one shifted allele
2	9780	1020/1220		✓			Pure
3	9940	350/1020	13	✓			Double peak starting at GAA 13
4	10100	+/+		✓			Double peak starting at GAA 14
5	10325	350/885	25	✓			Gap ~9 GAA's, third GAA weak
6	10466	850/1050	13	✓			Double peak at GAA 8
7	10722	400/1000	15	✓			Gap ~9 GAA's, third GAA weak
8	10905	683/983	12	✓			Pure
9	11437	67/1100	13	✓			Pure
10	11912	+/+		✓			After 5 GAA's, 5 GAA Gap, double peak starting at GAA 12
11	12451	834/834			✓	extra band at 350 bp	**13 GAA**
12	12941	600/767	5		✓	extra band at 350 bp	Pure
13	13037	520/850	20		✓	extra band at 350 bp	GAA 3 start, double peak starting at GAA 11
14	14805	850/1180	7	✓		only visible on gel	Pure
15	15657	467/667		✓			After GAA 1, Gap of 5 GAA, double peak starting at GAA 11
16	16852	967/1100	16	✓			Starting 3 GAA before GAA 1, 1 GAA Gap
17	17494	834/1167	7	✓			Pure
18	17652	480/880					**FAIL**
19	20886	1167/1167	2				Pure
20	53297	WT/612	15		✓	extra band at 300 bp	Double peak starting at GAA 12
21	26162	400/534	26	✓			Double peak starting at GAA 17
22	27643	667/1100	15				Double peak starting at GAA 11
23	27884	867/1134	4.5				Pure
24	29897	150/850	2	✓			Pure
25	30670	150/534	25		✓	extra band at 300 bp	14 GAA then drop off, double peak starting at GAA 12
26	34215	367/1100	27		✓	extra band at 300 bp	Starting 2 GAA earlier, double peak starting at GAA 14
27	34655	1020/1220	7	✓			Double peak starting at GAA 3
28	35594	567/834	14	✓			Double peak starting at GAA 8
29	39232	1000/1000	39	✓			Double peak starting at GAA 11
30	40908	1100/1200	10				Pure
31	41805	450/720	16	✓			Double peak starting at GAA 8
32	42181	685/920	15	✓			Pure
33	43286	1067/1167					Pure

34	44134	520/1050	16	✓			Pure
35	44293	1200/1200	17		✓	apoptotic ladder? Bands the whole way down	Starting 2 GAA earlier, double peak starting at GAA 3 and stopping at GAA 8
36	44655	134/1134	49	✓			Double peak starting at GAA 5, stopping at GAA 13, starting again at GAA 23
37	47084	767/1000	12				Double peak starting at GAA 9
38	47553	267/1100	13		✓	extra band at 300 bp	Double peak starting at GAA 5, stopping at GAA 10, starting again at GAA 22
39	47689	750/912	14	✓			Double peak starting at GAA 17
40	48978	750/850	9	✓			Pure
41	49823	300/700	34	✓			Increased intensity from GAA 10 (Dips, and higher peak again) Interruption??
42	51041	800/1000		✓			Double peak starting at GAA 8
43	44295	867/1100	24	✓			Starting 2 GAA earlier, double peak starting at GAA 2 and stopping at GAA 7
44	52046	+/+		✓			Increased intensity from GAA 4
45	52999	200/1000	29	✓			Increased intensity from GAA 14, double peak only at GAA 6/7
46	53084	767/967	4		✓	extra band (thick) at 400 bp	Pure
47	53085	700/1100	17	✓			Pure
48	53964	680/880	18	✓			Double peak starting at GAA 10
49	54278	645/845		✓			Pure
50	55057	1167/1500	11		✓	extra band at 300 bp	Pure
51	55070	580/745	1	✓			Pure
52	55718	600/967	19		✓	extra band at 300 bp	Double peak starting at GAA 13
53	55749	500/1000	18	✓			Double peak starting at GAA 9
54	55830	480/780	16	✓			Double peak starting at GAA 13
55	55837	667/900	13		✓	extra band (thick) at 400 bp	Pure
56	56603	845/845		✓			Pure - low intensity peak at GAA 1
57	56994	834/1100	2	✓			Starting 3 GAA before GAA 1, increased intensity at GAA 4
58	56999	734/1067	4	✓			Double peak at GAA 2
59	57261	920/1120	10	✓			Pure

60	57683	767/900	7				Starting 3 GAA before GAA 1, increased intensity at GAA 4
61	58035	785/1020	6	✓			Pure
62	58666	800/867	14	✓			Pure
63	59258	450/980	18	✓			Double peak starting at GAA 11
64	59345	1020/1250	6.5				Pure
65	59923	820/820	7	✓			Starting 3 GAA before GAA 1, double peak starting at GAA 6
66	59992	885/1050			✓	extra band at 300 bp	Pure
67	60541	500/667	13		✓	extra band at 250 bp	Starting 7 GAA before GAA 1, double peak starting at GAA 8
68	67580	612/912	9	✓			Double peak starting at GAA 9
69	69777	+/+			✓	extra band at 250 bp	Gap at GAA 23 - GAA 30, double peak from GAA 23
70	71074	312/780		✓			Double peak starting at GAA 18
71	71891	645/880		✓			Double peak starting at GAA 4
72	72843	712/900		✓			Pure
73	73047	645/812	6	✓			Pure
74	73066	785/785	1.5	✓ - very faint			Pure
75	73341	+/+		✓			Double peak starting at GAA 17
76	74809	680/745		✓			Pure
77	75641	+/+		✓			Starting 3 GGA before GAA 1, double peak starting at GAA 17
78	75836	+/+		✓			Pure
79	76333	+/+		✓			Pure
80	FRDA 6	167/500	51	✓			Double peak starting at GAA 8
81	FRDA 11	720/920	3	✓			Pure
82	FRDA 14	583/1183	22	✓			Double peak starting at GAA 7 - intense GAA 1
83	FRDA 15	+/+	1	✓			Pure
84	FRDA 18	1100/1134	6				Pure
85	FRDA 22	634/767	9	✓			Increased intensity from GAA 4
86	FRDA 23	167/834	28	✓			Gap from GAA 3 - GAA 7
87	FRDA 26	100/1100	55	✓			Double peak at GAA 14/15
88	FRDA 27	412/850	33	✓			Gap from GAA 4 - GAA 5, increased intensity from GAA 16, double peak at GAA 7, double peak starting from GAA 16
89	FRDA 28	380/780	42	✓			Gap from GAA 4 - GAA 5, increased

							intensity from GAA 16, double peak at GAA 7, double peak starting from GAA 16
90	FRDA 33	834/1034	12	✓			Pure
91	FRDA 37	585/1250	14	✓			Starting 2 GAA before GAA 1, double peak starting at GAA 3 and decreasing at GAA 8, increased intensity from GAA 13
92	FRDA 39	785/785	5	✓			Increased intensity at GAA 3
93	FRDA 40	400/834	17	✓			Double peak from GAA 9
94	FRDA 41	100/500	15	✓			Starting 7 GAA before GAA 1, double peak starting at GAA 11
95	FRDA 42	780/980	7		✓	extra band at 350 bp	Starting 2 GAA before GAA 1, double peak starting at GAA 4
96	FRDA 43	334/900	20		✓	band present above the ladder	Double peak starting at GAA 14
97	FRDA 50	467/667	12	✓			Gap from GAA 2 - GAA 6, double peak starting at GAA 11
98	FRDA 54	650/850	13				Low intensity for GAA 1, double peak starting at GAA 11
99	FRDA 55	700/1000	15	✓			Pure
100	FRDA 56	800/1000	8	✓			Double peak starting at GAA 8
101	FRDA 57	1100/1234	10	✓			Pure
102	FRDA 58	200/1000	30	✓			Double peak at GAA 6/7 only, increased intensity from GAA 14
103	FRDA 61	834/1200	3	✓			Pure
104	FRDA 74	800/867	8	✓			Double peak starting at GAA 8
105	FRDA 76	1000/1200	8	✓			Pure
106	FRDA 78	720/920	19	✓			Pure
107	FRDA 81	720/1020	19	✓			Double peak starting at GAA 7
108	FRDA 84	567/1000	22	✓			GAA 3 missing, double peak starting at GAA 14
109	FRDA 86	850/1150	3	✓			Pure
110	FRDA 87	850/850	2	✓			Starting 1 GAA later, double peak from GAA 4
111	FRDA 88	685/1120	6	✓			Starting 2 GAA before, double peak from GAA 3
112	FRDA 89	750/850	4	✓			Double peak starting at GAA 4
113	FRDA 92	380/520	19	✓			Low intensity from GAA 1- GAA 6, Increased intensity at GAA 7 and at GAA 9

114	FRDA 93	720/885	18	✓			Double peak starting from GAA 10
115	FRDA 94	900/1200	14	✓			Double peak starting from GAA 10
116	FRDA 97	450/820	16	✓			Double peak starting at GAA 7
117	FRDA 98	920/920	10	✓			Pure
118	FRDA 99	450/985	19	✓			Double peak starting from GAA 11
119	FRDA 100	850/1150	7	✓			GAA starting from GAA 4
120	FRDA 101	1185/1185	12	✓			Double peak starting from GAA 4
121	FRDA 102	985/1120	12	✓			Pure
122	FRDA 104	1050/1050	8	✓			Pure
123	FRDA 105	WT/867	15	✓			Pure
124	FRDA 106	800/1134	6		✓	extra band at 300 bp	Pure
125	FRDA 107	834/834	4	✓			Pure
126	FRDA 108	700/800	11	✓			Double peak starting at GAA 8
127	FRDA 109	834/1100	10				Starting 3 GAA before GAA 1, double peak starting at GAA 3
128	FRDA 110	200/1100	36	✓			Double peak starting at GAA 5, stopping at GAA 9 and starting from GAA 16 again
129	FRDA 111	No GAA		✓			**13 GAA**
130	FRDA 112	734/900	7	✓			Pure
131	FRDA 113	720/720	10	✓			Pure
132	FRDA 114	400/667	11	✓			Pure
133	FRDA 115	600/834	12	✓			Pure
134	FRDA 116	634/1100	10	✓			Pure
135	FRDA 117	767/1134	15	✓			Double peak starting from GAA 2
136	FRDA 119	700/1000	5		✓	band present above the ladder	Double peak starting from GAA 8
137	FRDA 122	785/850	9	✓			Pure
138	FRDA 123	700/1200	12	✓			Double peak starting at GAA 9
139	FRDA 124	467/967	15	✓			Double peak starting from GAA 8
140	FRDA 125	567/900	13	✓			Pure
141	FRDA 126	767/867	8	✓			Pure
142	FRDA 127	600/1100	17	✓			Starting 3 GAA before GAA 1, double peak starting at GAA 18
143	FRDA 128	434/600	22	✓			Low intensity GAA 1, 5 GAA gap, increased intensity at GAA 7
144	FRDA 129	734/900	13	✓			Pure
145	FRDA 132	667/767	3	✓			Pure
146	FRDA 133	745/945	5	✓			Low intensity GAA 1
147	FRDA 134	1080/1080	17	✓			Increased intensity from GAA 11

148	FRDA 135	445/780	15	✓			Double peak starting at GAA 9
149	FRDA 137	780/880	7	✓			Pure
150	FRDA 138	745/845	6	✓			Pure
151	FRDA 140	No GAA		✓			**4 GAA**
152	FRDA 141	645/780	6	✓			Pure
153	FRDA 144	412/645	14	✓			Increased intensity at GAA 4
154	FRDA 147	212/845	33		✓	band present above the ladder.	Double peak starting at GAA 22
155	FRDA 148	245/912	24	✓			Double peak starting at GAA 22
156	FRDA 149	780/1180	11	✓			Double peak starting at GAA 12
157	FRDA 150	45/745	13		✓	extra band at 350 bp	Pure
158	FRDA 151	650/980	6	✓			Pure
159	17786	+/+			✓	extra band at 250 bp	Low intensity at GAA 3 and GAA 9, gap from GAA 10 to GAA 14
160	18204	+/+		✓			Single low intensity peak around 3 GAA before GAA 1, double peak at GAA 5, 5 GAA gap, double peak starting at GAA 16
161	65331	+/+		✓			7 GAA gap after 4 GAA

The FRDA cohort consists of 246 FRDA patients and 7 carriers (N = 253). GAA1/2: GAA1 size/ GAA2 size; AAO: age at onset; WTC: carrier wild-type allele, < 44 GAA repeats; WT: wild-type allele, < 44 GAA repeats; +: expanded allele of undetermined size, ≥ 44 GAA repeats; bold and red highlight: GAA sizes < 40 GAA repeats as determined by TP-PCR; bold: sample failed TP-PCR analysis; HTX: heteroduplex formation; bp: base pair.

Table 2. DNA sequences determined by clone sequencing of HD patients

	$(CAG)_n$	CCG1	Inter-CCG repeat region	CCG2
Ref Seq	CAG_{21}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
Patient 1	CAG_{43}	CCG_{12}-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
Patient 2	CAG_{43}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_2-**A**CG-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{39}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{43}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{43}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
Patient 3	CAG_{39}	CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
Patient 4	CAG_{41}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-**A**CG-CCG-CCA-CCC
	CAG_{42}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-**A**CG-CCG-CCA-CCC
	CAG_{42}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-**A**CG-CCG-CCA-CCC
	CAG_{42}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-**A**CG-CCG-CCA-CCC
	CAG_{42}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-**A**CG-CCG-CCA-CCC
Patient 5	CAG_{35}-CAA-CAG	CCG-CCA-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{36}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{36}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{38}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC

	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
Patient 6	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-C**A**G-CCG-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-C**A**G-CCG-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-C**A**G-CCG-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-C**T**C-CCG$_2$-CCA-CCC
Patient 7	CAG$_{42}$	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
Patient 8	CAG$_{39}$-CAA-CAG	CCG$_{\underline{9}}$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG$_{\underline{9}}$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{44}$-CAA-CAG	CCG$_{\underline{9}}$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
Patient 9	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_2$-CC**A**-CCG$_4$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-C**T**G-CCG-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-C**T**G-CCG-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_2$-CC**A**-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{\underline{6}}$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-C**T**G-CCG-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_2$-**T**CG-CCG-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-C**T**C-CCG$_2$-CCA-CCC
Patient 10	CAG$_{36}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC

	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_{10}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 11	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-<u>**A**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-C<u>**T**</u>C-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-<u>**A**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-<u>**A**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-<u>**A**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-<u>**A**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
Patient 13	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG-<u>**A**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{\underline{\mathbf{6}}}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{10}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{2}$-<u>**T**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{10}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{2}$-<u>**T**</u>CG-CCG-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{45}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 14	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 15	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC

Patient 17	CAG$_{25}$-CA**A**-CAG$_{17}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{35}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{36}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{36}$-CAA-CAG	CCG-CCA$_{\mathbf{2}}$-CCG$_{\mathbf{6}}$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-C**T**C-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{10}$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
Patient 18	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{49}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
Patient 20	CAG$_{36}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{37}$-CA**A**-CAG$_2$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
Patient 23	CAG$_{37}$-CAA-CAG	CCG-CCA-CCG$_{10}$-CCT$_2$	CAG-CTT-CCT-CAG-CC**A**-CCG$_2$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_2$-**T**CG-CCC-CCG$_2$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_{\mathbf{6}}$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC
Patient 24	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_7$-CCT$_2$	CAG-CTT-CCT-CAG-CCG$_3$-CAG-GCA-CAG-CCG-CTG$_2$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_3$-CCC-CCG$_2$-CCA-CCC

	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{6}$-CCT$_{3}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{6}$-CCT$_{3}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 25	CAG$_{36}$-CAA-CAG	CCG-CCA-CCG$_{6}$-CCT$_{3}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC$_{2}$-CCG$_{2}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC$_{2}$-CCG$_{2}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 28	CAG$_{37}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{6}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$	CCG-CCA-CCG$_{10}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$	CCG-CCA-CCG$_{10}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 29	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{50}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 30	CAG$_{32}$-CA**A**-CAG$_{7}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
Patient 31	CAG$_{16}$-CA**A**-CAG$_{28}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC

	CAG_{43}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC

$(CAG)_n$: CAG repeat sequence including CAA if present; RefSeq: reference sequence; CCG1: CCG repeat region 1; CCG2: CCG repeat region 2; Inter-CCG repeat region: region in between CCG1 and CCG2; red, bold and underlined: Sequence variations detected in the CAG repeat, CCG repeats and inter-CCG repeat region of *HTT*; light blue fill: Novel sequences that have not previously been reported.

Table 3. DNA sequences determined by clone sequencing of HD *post-mortem* brains

	$(CAG)_n$	CCG1	Inter-CCG repeat region	CCG2
Ref Seq	CAG_{21}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
			P40.97	
FNT	CAG_{43}-CAA-CAG	CCG-CCA-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
TMP	CAG_{40}-CAA-CAG	CCG-CCA-CCG_{10}-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_{10}-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
OCC	CAG_{36}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
PUT	CAG_{43}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
CNU	CAG_{36}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{38}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
CBM	CAG_{44}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
MED	CAG_{36}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{42}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
			P2.03	
PUT	CAG_{44}-CAA-CAG	CCG-CCA-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CC**T**-CCG_2-CCA-CCC
CBM	CAG_{42}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
			P72.10	

FNT	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{35}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
OCC	CAG$_{35}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{6}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{45}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{45}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
PUT	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
MED	CAG$_{35}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{47}$	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
			P3.92	
TMP	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_{9}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
OCC	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
CBM	CAG$_{34}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{34}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC

	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA$_{\mathbf{2}}$-CCG$_{\mathbf{6}}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
PON	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
MED	CAG$_{34}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{\mathbf{6}}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
			P7.96	
FNT	CAG$_{37}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-**A**CG-CCG-CCA-CCC
	CAG$_{42}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{45}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
TMP	CAG$_{35}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG-CC**T**-CCA-CCC
	CAG$_{43}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
PUT	CAG$_{37}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{38}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{39}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{40}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC
	CAG$_{41}$-CAA-CAG	CCG-CCA-CCG$_{7}$-CCT$_{2}$	CAG-CTT-CCT-CAG-CCG$_{3}$-CAG-GCA-CAG-CCG-CTG$_{2}$-CCT-CAG-CCG-CAG	CCG-CCC-CCG$_{3}$-CCC-CCG$_{2}$-CCA-CCC

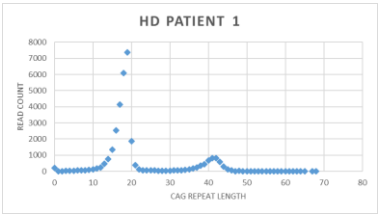
	CAG_{42}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{42}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{46}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{47}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{47}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{47}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
CNU	CAG_{38}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{39}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
CBM	CAG_{36}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{37}-CAA-CAG	CCG-CCA-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{38}-CAA-CAG	CCG-CCA_2-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{39}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{39}-CAA-CAG	CCG-CCA-CCG_3-CCT-CCG_3-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{40}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-ACG-CCG-CCA-CCC
PON	CAG_{36}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA_2-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
MED	CAG_{38}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{38}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{39}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{41}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{42}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{42}-CAA-CAG	CCG_2-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{43}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC

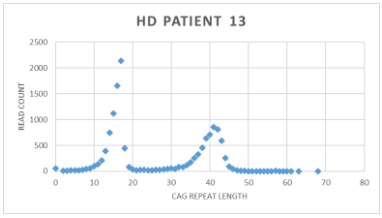
	CAG_{43}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
			P28.98	
FNT	CAG_{37}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{45}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{45}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{49}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
TMP	CAG_{43}	CCG-CCA_2-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{45}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{46}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{47}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{47}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
OCC	CAG_{38}-CAA-CAG	CCG-CCA_2-CCG_6-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{39}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{42}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{43}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{45}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
PUT	CAG_{42}-CAA-CAG	CCG-CCA_2-CCG_5-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{43}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC
	CAG_{44}-CAA-CAG	CCG-CCA-CCG_7-CCT_2	CAG-CTT-CCT-CAG-CCG_3-CAG-GCA-CAG-CCG-CTG_2-CCT-CAG-CCG-CAG	CCG-CCC-CCG_3-CCC-CCG_2-CCA-CCC

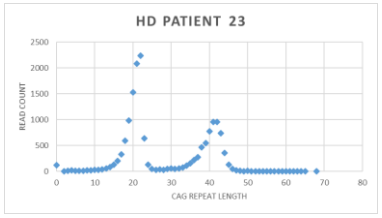
$(CAG)_n$: CAG repeat sequence including CAA if present; RefSeq: reference sequence; CCG1: CCG repeat region 1; CCG2: CCG repeat region 2; Inter-CCG repeat region: region in between CCG1 and CCG2; red, bold and underlined: Sequence variations detected in the CAG repeat, CCG repeats and inter-CCG repeat region of *HTT*; light blue fill: Novel sequences that have not previously been reported.

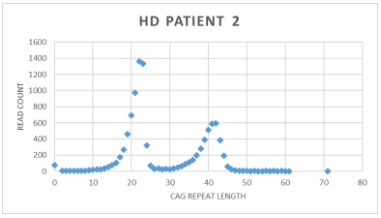
Figure 1. Sizing the CAG repeat in HD patients by Nanopore sequencing (overleaf)

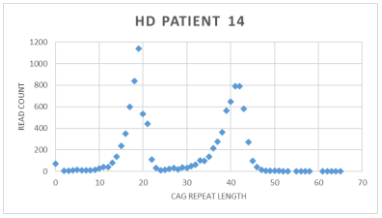
Y-axis: read count; x-axis: CAG repeat size; P82.10: control *post-mortem* brain.

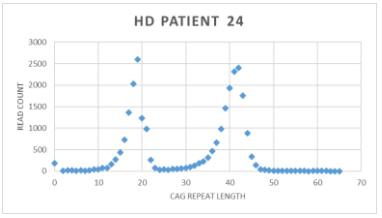

HD PATIENT 1
READ COUNT
CAG REPEAT LENGTH

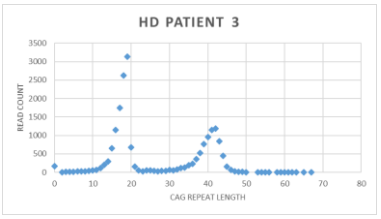

HD PATIENT 13
READ COUNT
CAG REPEAT LENGTH

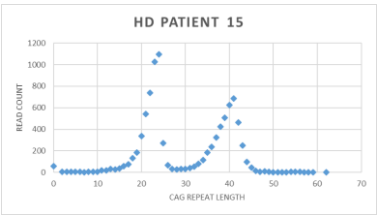

HD PATIENT 23
READ COUNT
CAG REPEAT LENGTH

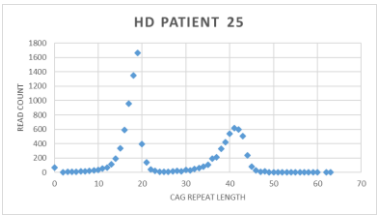

HD PATIENT 2
READ COUNT
CAG REPEAT LENGTH

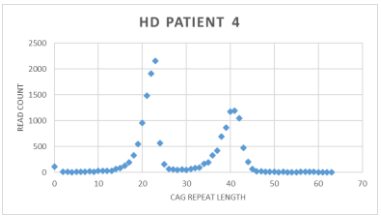

HD PATIENT 14
READ COUNT
CAG REPEAT LENGTH

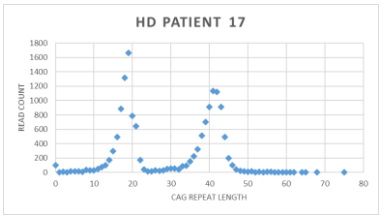

HD PATIENT 24
READ COUNT
CAG REPEAT LENGTH

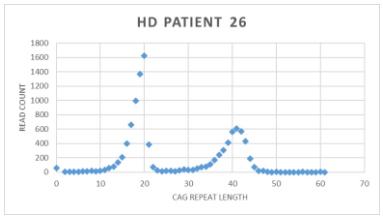

HD PATIENT 3
READ COUNT
CAG REPEAT LENGTH

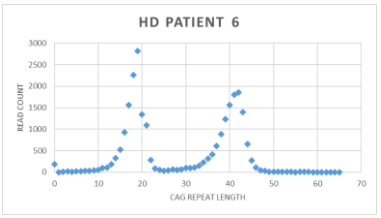

HD PATIENT 15
READ COUNT
CAG REPEAT LENGTH

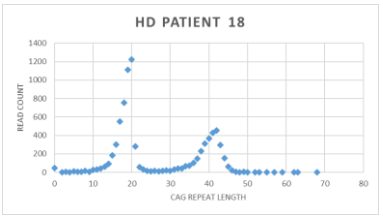

HD PATIENT 25
READ COUNT
CAG REPEAT LENGTH

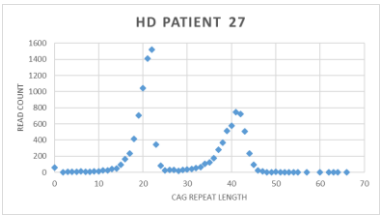

HD PATIENT 4
READ COUNT
CAG REPEAT LENGTH

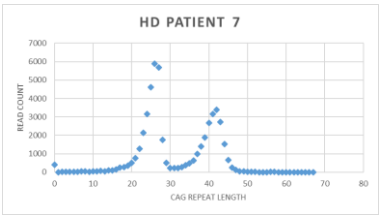

HD PATIENT 17
READ COUNT
CAG REPEAT LENGTH

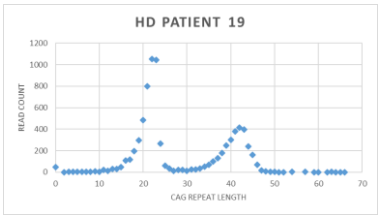

HD PATIENT 26
READ COUNT
CAG REPEAT LENGTH

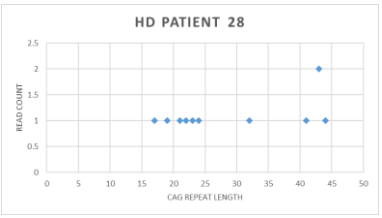

HD PATIENT 6
READ COUNT
CAG REPEAT LENGTH

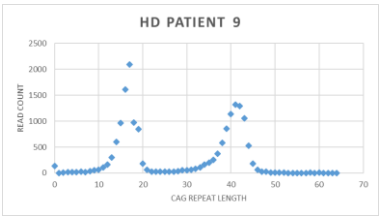

HD PATIENT 18
READ COUNT
CAG REPEAT LENGTH

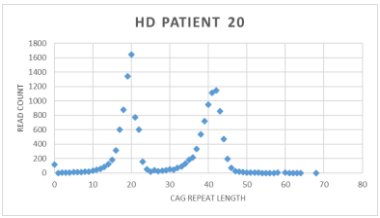

HD PATIENT 27
READ COUNT
CAG REPEAT LENGTH

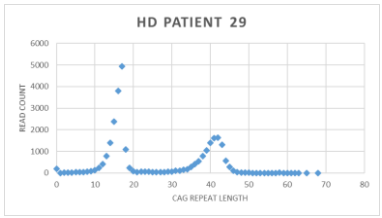

HD PATIENT 7
READ COUNT
CAG REPEAT LENGTH

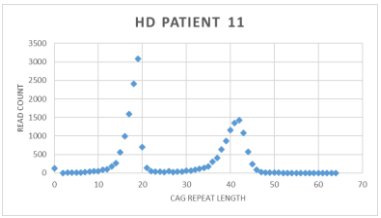

HD PATIENT 19
READ COUNT
CAG REPEAT LENGTH

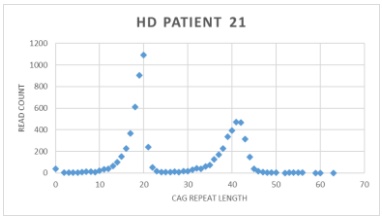

HD PATIENT 28
READ COUNT
CAG REPEAT LENGTH

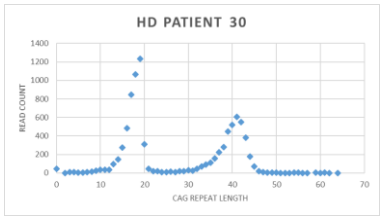

HD PATIENT 9
READ COUNT
CAG REPEAT LENGTH


HD PATIENT 20
READ COUNT
CAG REPEAT LENGTH


HD PATIENT 29
READ COUNT
CAG REPEAT LENGTH


HD PATIENT 11
READ COUNT
CAG REPEAT LENGTH


HD PATIENT 21
READ COUNT
CAG REPEAT LENGTH


HD PATIENT 30
READ COUNT
CAG REPEAT LENGTH

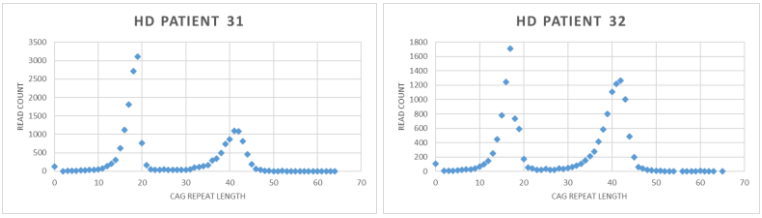
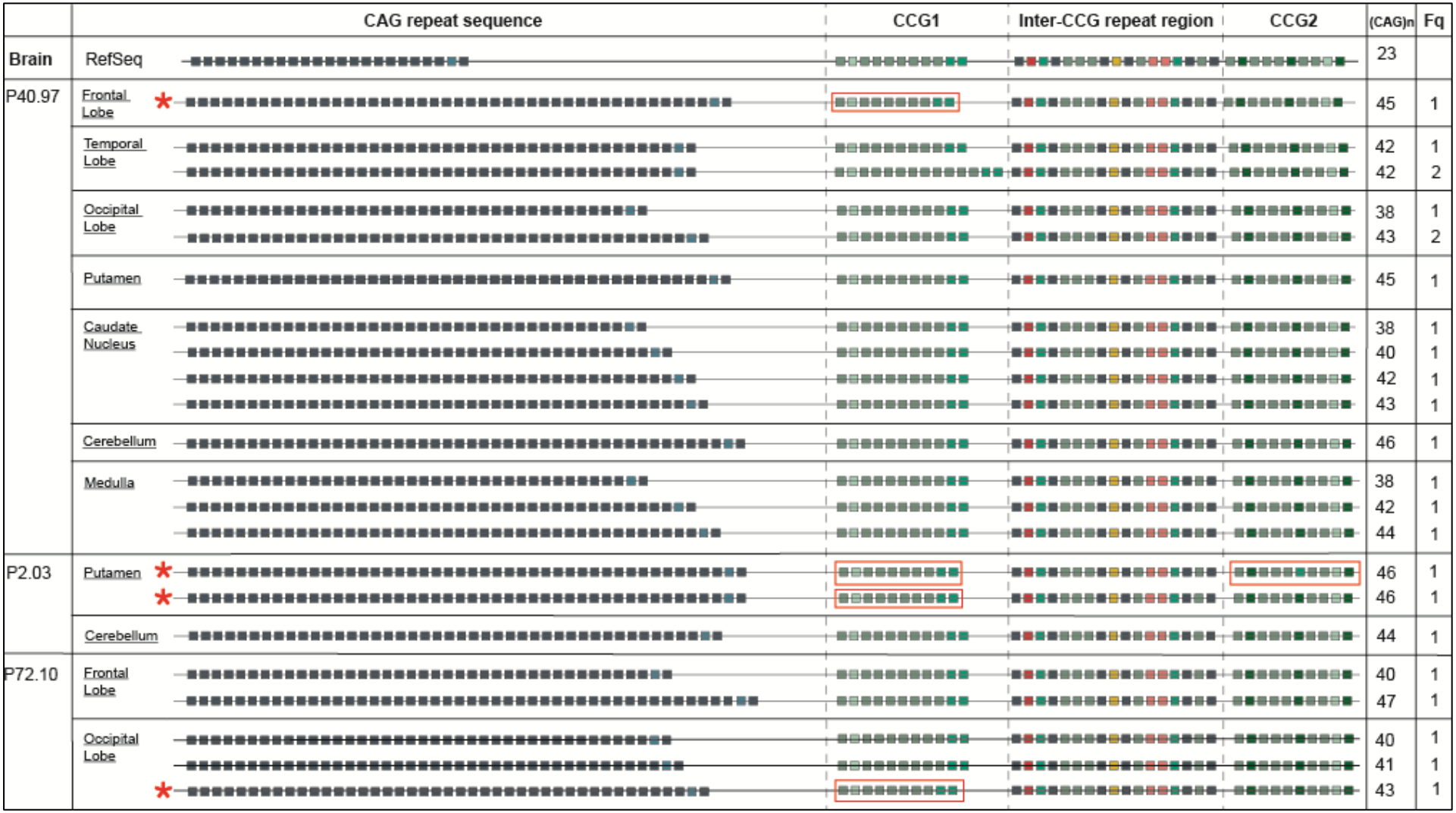


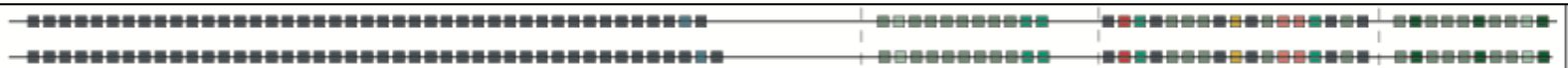
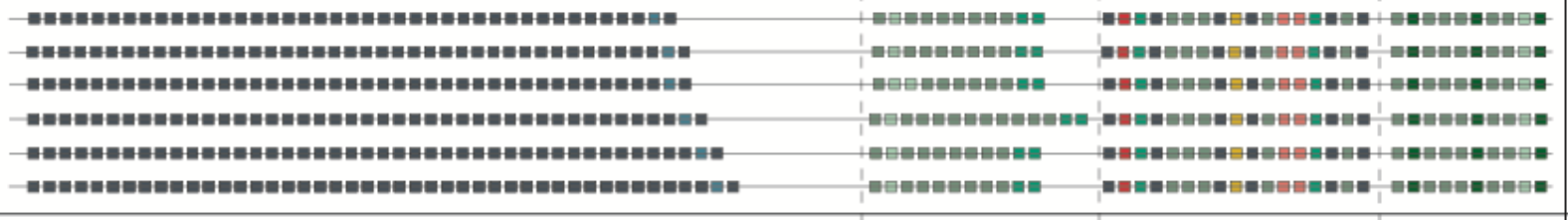
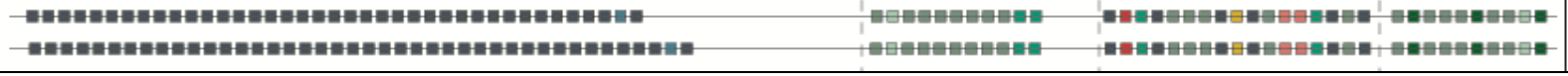
Figure 1. Sizing the CAG repeat in HD patients by Nanopore sequencing

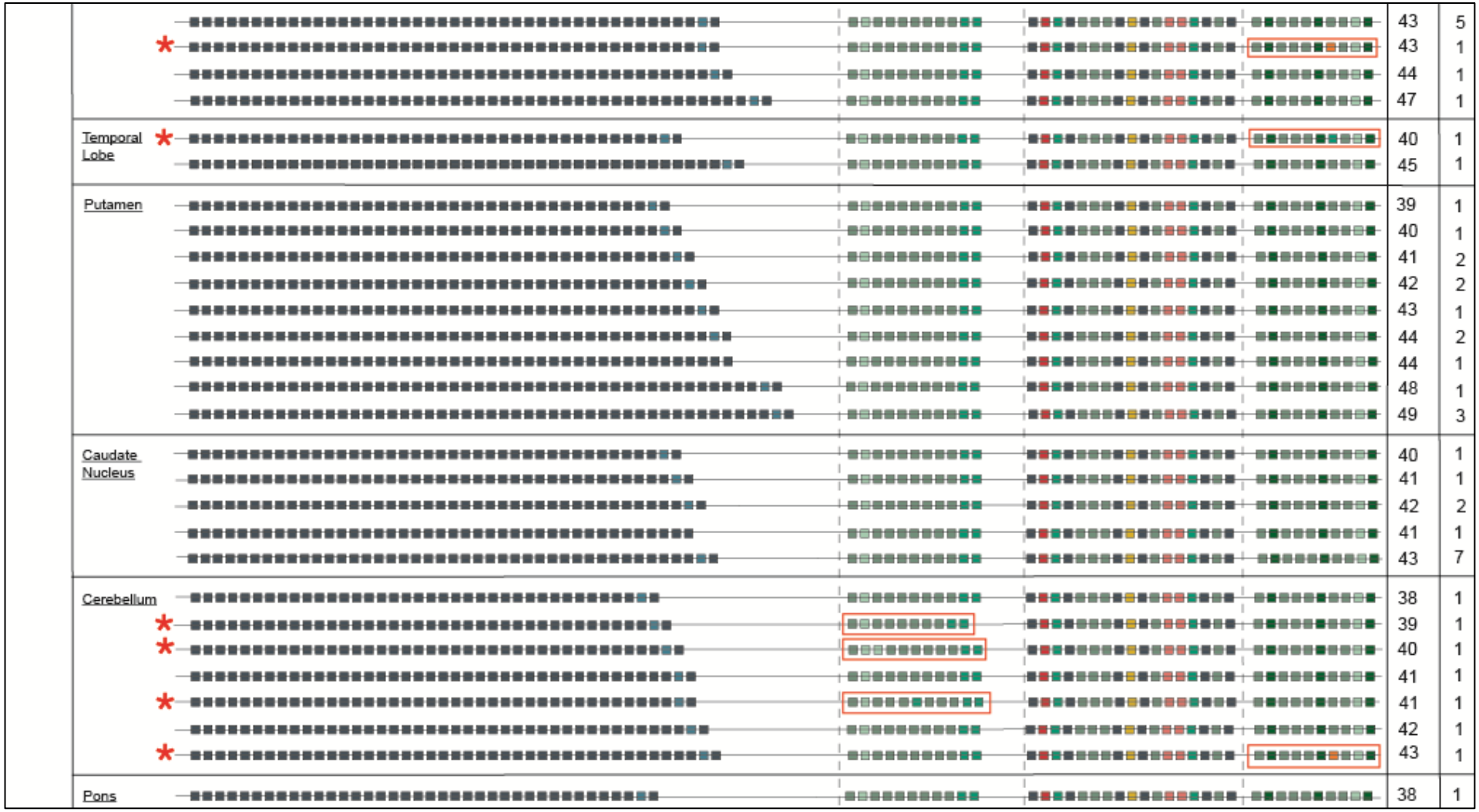
---

Figure 2. DNA sequences from successfully cloned HD *post-mortem* brains (overleaf)

Each square represents a codon and is colour-coded for visual ease. The frequency (Fq) represents the population of clones obtained per patient and quantifies the presence of each sequence. RefSeq: *HTT* reference sequence according to NCBI (https://www.ncbi.nlm.nih.gov/nuccore/NG_009378.1?from=5001&to=174286&report=fasta); CAG repeat sequence: codon composition including CAA trinucleotides; CCG1: codon composition of CCG repeat region 1; CCG2: codon composition of CCG repeat region 2; Inter-CCG repeat region: codon composition of the region in-between CCG1 and CCG2; $(CAG)_n$: CAG repeat tract of size n including CAA if present; *: HD patients with previously unreported sequence alterations; red rectangle: location of previously unreported sequence alterations.


CAG repeat sequence
CCG1
Inter-CCG repeat region
CCG2
(CAG)n
Fq
Brain
RefSeq
23
P40.97
Frontal Lobe
45 1
Temporal Lobe
42 1
42 2
Occipital Lobe
38 1
43 2
Putamen
45 1
Caudate Nucleus
38 1
40 1
42 1
43 1
Cerebellum
46 1
Medulla
38 1
42 1
44 1
P2.03
Putamen
46 1
46 1
Cerebellum
44 1
P72.10
Frontal Lobe
40 1
47 1
Occipital Lobe
40 1
41 1
43 1

			43 44 45 46	1 5 1 1
	Putamen		43 45	1 1
	Medulla		43 44 47	3 2 1
P3.92	Temporal Lobe		42 45	5 1
	Occipital Lobe		44	3
	Cerebellum		41 42 42 43 44 45	4 1 1 1 1 1
	Pons		41 42 44	1 1 1
	Medulla		41 43 43	1 2 1
P7.96	Frontal Lobe		39 42	1 5


Temporal Lobe
Putamen
Caudate Nucleus
Cerebellum
Pons
43 5
43 1
44 1
47 1
40 1
45 1
39 1
40 1
41 2
42 2
43 1
44 2
44 1
48 1
49 3
40 1
41 1
42 2
41 1
43 7
38 1
39 1
40 1
41 1
41 1
42 1
43 1
38 1

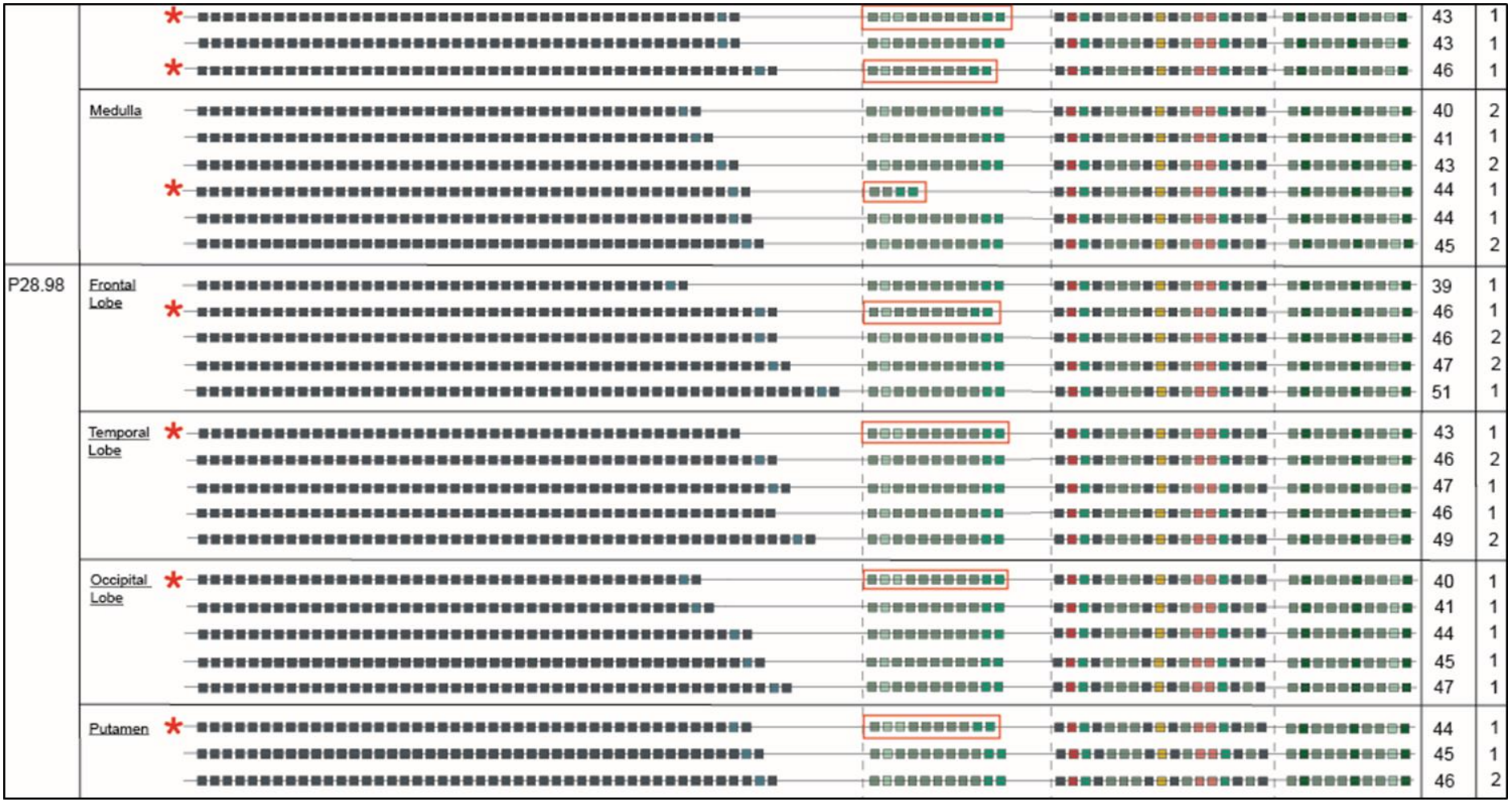
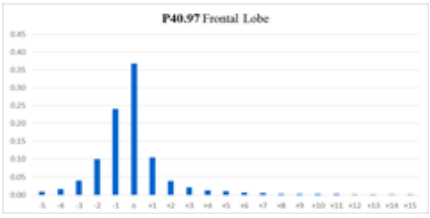


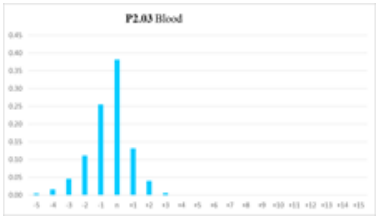
Figure 2. DNA sequences from successfully cloned HD *post-mortem* brains

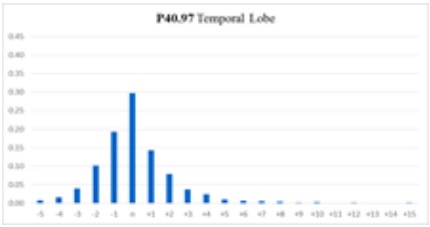
Figure 3. Read count distributions for the HD *post-mortem* brain and corresponding blood samples (overleaf)

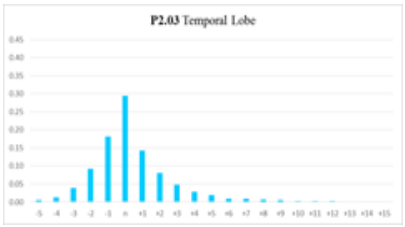
The x-axis represents the $(CAG)_n$ (number of CAGs in the modal allele) in the range of -5 to +15 CAGs from the $(CAG)_n$. The y-axis represents the relative peak height, which demonstrates the percentage of MiSeq reads present corresponding to the $(CAG)_n$.

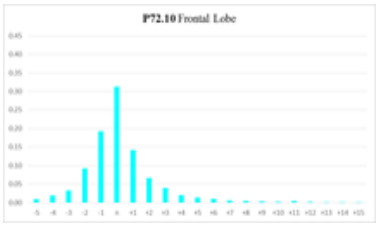

P40.97 Frontal Lobe

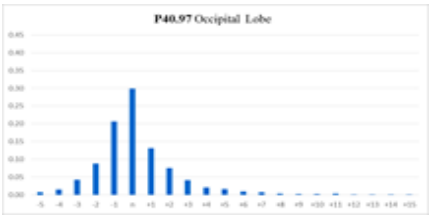

P2.03 Frontal Lobe


P2.03 Blood

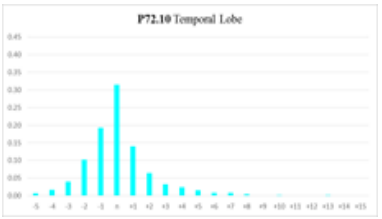

P40.97 Temporal Lobe

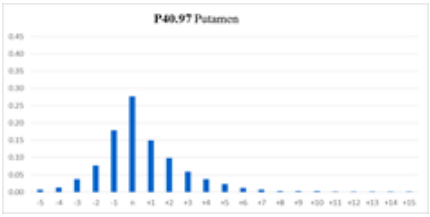

P2.03 Temporal Lobe

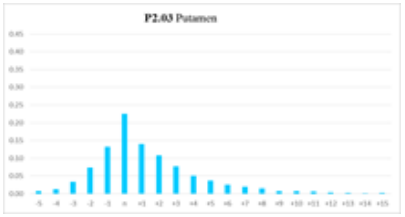

P72.10 Frontal Lobe

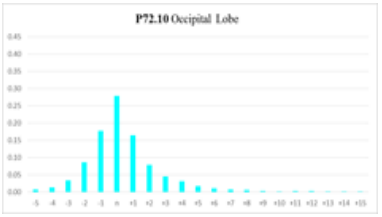

P40.97 Occipital Lobe

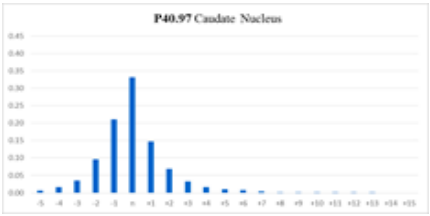

P2.03 Occipital Lobe

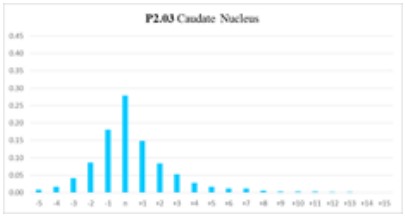

P72.10 Temporal Lobe

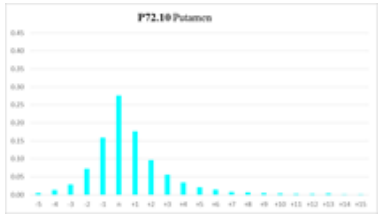

P40.97 Putamen


P2.03 Putamen

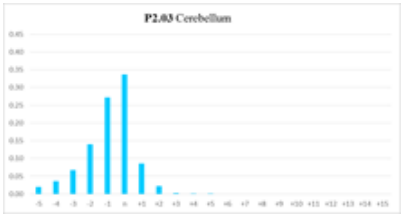

P72.10 Occipital Lobe

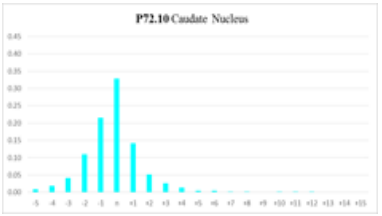

P40.97 Caudate Nucleus

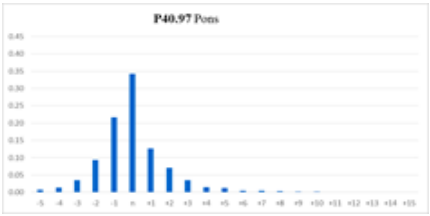

P2.03 Caudate Nucleus


P72.10 Putamen

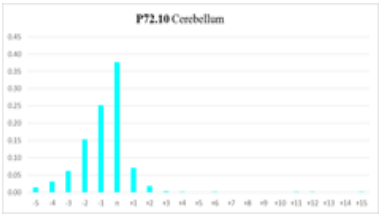

P40.97 Cerebellum

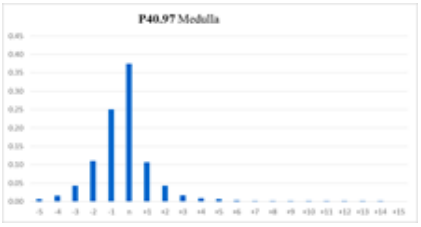

P2.03 Cerebellum

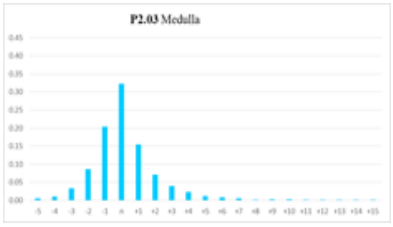

P72.10 Caudate Nucleus


P40.97 Pons

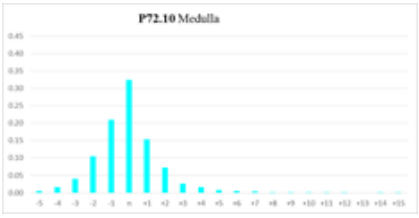

P2.03 Pons

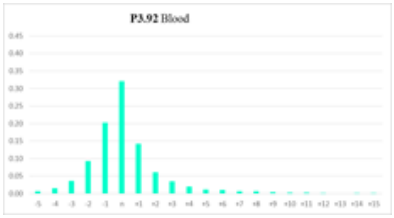

P72.10 Cerebellum

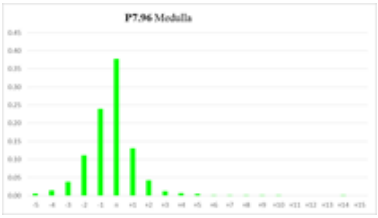

P40.97 Medulla

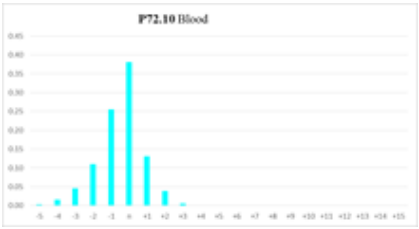

P2.03 Medulla

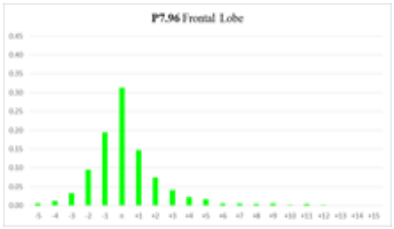

P72.10 Pons

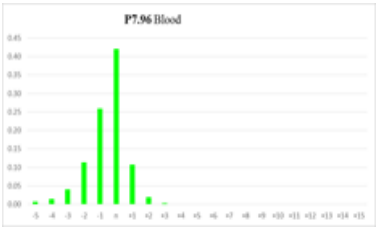

P72.10 Medulla

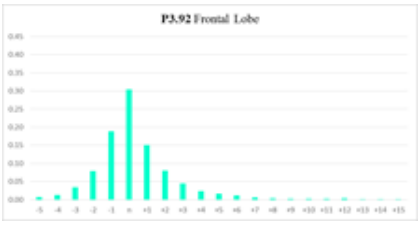

P3.92 Blood

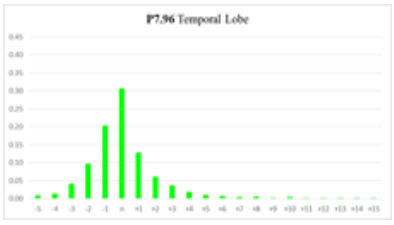

P7.96 Medulla

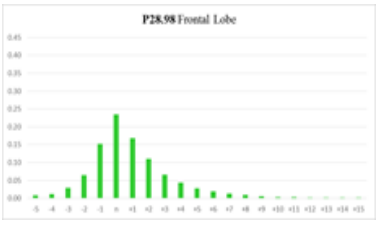

P72.10 Blood

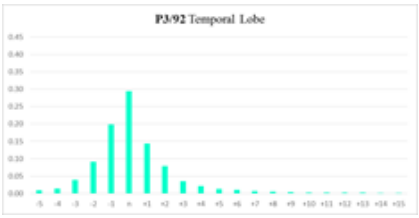

P7.96 Frontal Lobe

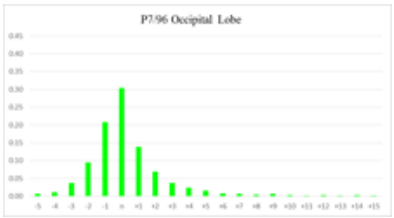

P7.96 Blood

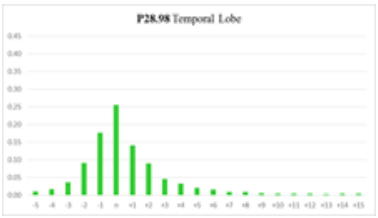

P3.92 Frontal Lobe

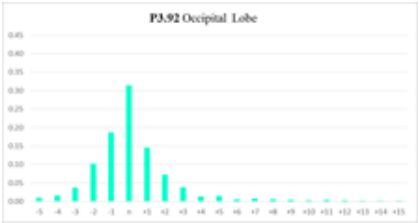

P7.96 Temporal Lobe

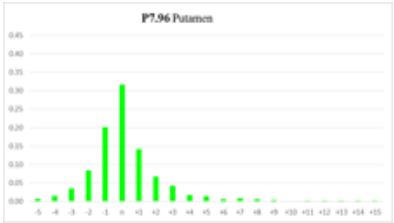

P28.98 Frontal Lobe


P3/92 Temporal Lobe

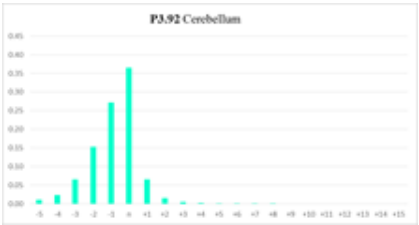

P7/96 Occipital Lobe

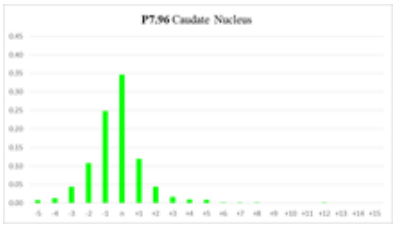

P28.98 Temporal Lobe

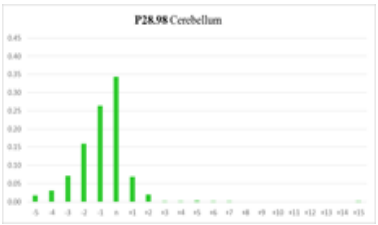

P3.92 Occipital Lobe

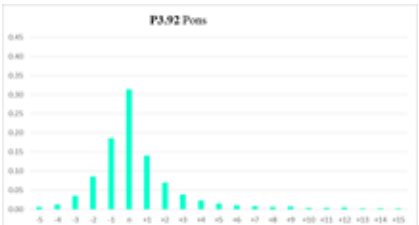

P7.96 Putamen


P28.98 Occipital Lobe


P3.92 Cerebellum


P7.96 Caudate Nucleus


P28.98 Cerebellum


P3.92 Pons


P7.96 Cerebellum


P28.98 Pons


P3.92 Medulla


P7.96 Pons


P28.98 Medulla

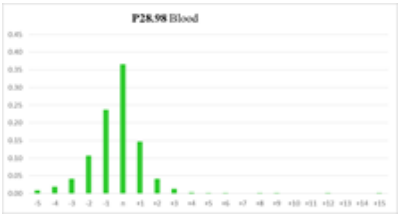


Figure 3. Read count distributions for the HD *post-mortem* brain and corresponding blood samples

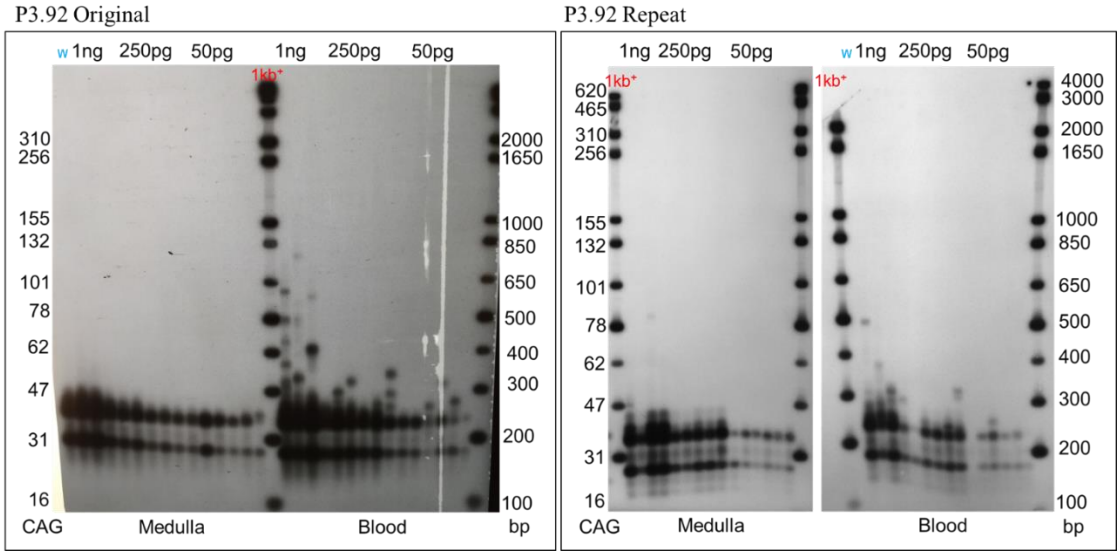


Figure 4. SP-PCR analysis of P3.92 medulla and blood

These samples were repeated due to the surprising instability of the CAG repeat in the blood. Subsequent investigation of the sample number in the Neurogenetics database revealed that it was DNA extracted from one of the post-mortem brain regions. The $(CAG)_n$ is labelled on the left and the base pair (bp) on the right according to the 1 kb^+ ladder. The first 3 lanes represent 1 ng of DNA, followed by 6 lanes each of 250 pg and 50 pg of DNA. Water (w) is used as a control to demonstrate that no PCR products are detected in the absence of a DNA template.

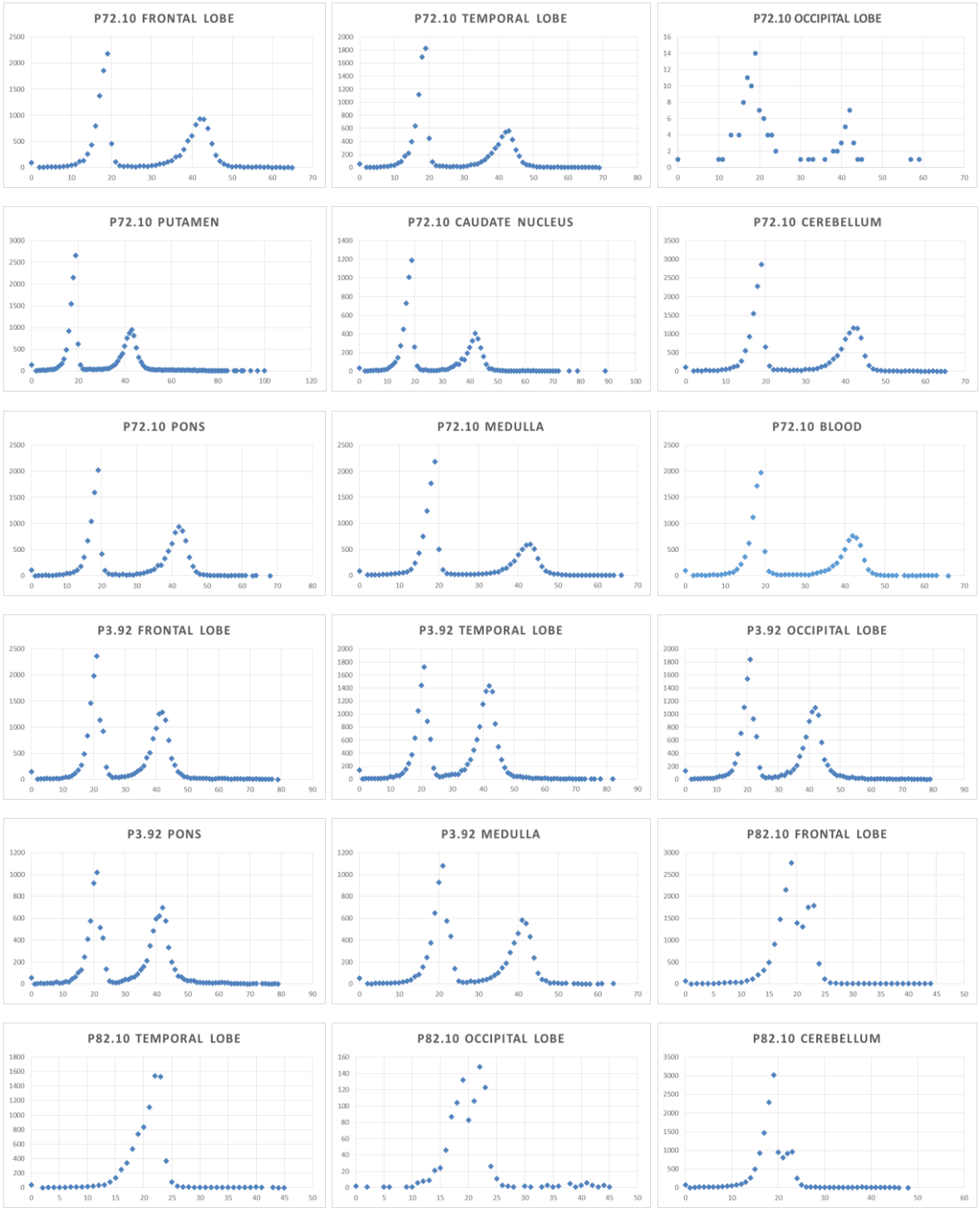


Figure 5. Sizing the CAG repeat in HD *post-mortem* brains by Nanopore sequencing

Y-axis: read count; x-axis: CAG repeat size; P82.10: control *post-mortem* brain.