

https://dspace.org/

Investigating allele sequence diversity at the Huntington disease loci HTT and JPH3 in African ancestry individuals

Levesley, Jessica

2021
https://hdl.handle.net/10539/34576

Downloaded from a DSpace repository. https://dspace.org/

Investigating allele sequence diversity at the Huntington disease loci *HTT* and *JPH3* in African ancestry individuals

Jessica Levesley

Student Number: 600997



A Thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Supervisors:

Dr Fiona Savanhu-Baine, MSc (Med), PhD

Professor Amanda Krause, MBBCh, PhD

Professor Darren Monckton, BSc, PhD

Johannesburg, 2021

Declaration

I, Jessica Levesley, declare that this thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

_2_ day of_August_2021_in_Johannesburg

Presentations from this study

Levesley, J., Baine-Savanhu, F., Ciosi, M., Maxwell, A., Monckton, D G., Krause, A. Poster. A *cis*-acting modifier of age at diagnosis of Huntington disease in black South African patients. European Huntington Disease Network (EHDN) Plenary meeting, Austria Centre Vienna, Vienna, Austria, 14 - 16 September 2018.

- Levesley, J., Baine-Savanhu, F., Ciosi, M., Maxwell, A., Monckton, D G., Krause, A. Poster. C03 A *cis*-acting modifier of age at diagnosis of Huntington disease in black South African patients. *Journal of Neurology, Neurosurgery and Psychiatry*, 89, S1. 10.1136/jnnp-2018-EHDN.74.

Levesley, J., Baine-Savanhu, F., Ciosi, M., Maxwell, A., Monckton, D G., Krause, A. Oral. A *cis*-acting modifier of age at diagnosis of Huntington disease in black South African patients. Faculty of Health Sciences Biennial Research Day & Postgraduate Expo, Medical School, University of Witwatersrand, Johannesburg, South Africa, 6 September 2018.

Levesley, J.; Baine-Savanhu, F.; Ciosi, M.; Maxwell, A.; Monckton, DG.; Krause, A. Oral. Investigating the allele sequence diversity at the Huntington disease loci in an Africa population. 18th Biennial Congress of the Southern African Society for Human Genetics, Century City Conference Centre, Cape Town, South Africa. 3- 6 August 2019.

Levesley, J.; Baine-Savanhu, F.; Ciosi, M.; Monckton, DG.; Krause, A. Oral. Investigating the allele sequence diversity at the Huntington disease like-2 locus. Faculty of Health Sciences Virtual Research Day & Postgraduate Expo. 15 October 2020.

Levesley, J.; Baine-Savanhu, F.; Ciosi, M.; Monckton, DG.; Krause, A. Oral. The loss of the intervening proline sequence is a novel *cis*-acting modifier of Huntington disease in African ancestry patients. ASHG 2020 Virtual Meeting. 27-30 October 2020.

Abstract

Huntington disease (HD) is a dominantly inherited neurodegenerative disorder caused by the expansion of the CAG repeat in the huntingtin (*HTT*) gene. The clinical phenotype of HD is commonly described as a triad of movement disorder, cognitive decline and psychiatric complications. In South Africa, patients with Africa ancestry presenting with an HD phenotype may alternatively have Huntington disease like-2 (HDL2), caused by the expansion of the CTG repeat in the junctophilin-3 (*JPH3*) gene.

Although the underlying pathological mechanism in HD/HDL2 remains unclear, the main driver of disease severity is the expanded repeat. However, considerable additional variation is observed that is heritable. Although HD has been extensively studied in patients with European ancestry, very few studies have been dedicated to studying patients with African ancestry with even fewer studies conducted on HDL2. This study aimed to characterise the diversity at the two loci responsible for HD and HDL2 and explore potential genetic modifiers of disease.

In the *HTT* analysis, a high level of sequence diversity was observed in African ancestry individuals with the identification of three novel allele structures. The most common disease-associated allele structure Q^1-2-0-9-2 was defined as atypical, due to the loss of the CCGCCA sequence. Although the CAG repeat was confirmed to be the main driver of disease severity, the Q^1-2-0-9-2 allele structure was associated with an earlier age of diagnosis of approximately 7.1 years and occurred exclusively on haplotype B2. Although the somatic expansion was associated with an earlier age of diagnosis, the Q^1-2-0-9-2 allele structure displayed reduced somatic expansion. Thus, the Q^1-2-0-9-2 allele structure occurring on haplotype B2 was proposed as an African *cis*-acting modifier that appears to modify the age of diagnosis, through a mechanism that was not driven by somatic expansion. This was a particularly interesting finding, as it was hypothesised that these patients would have the highest amount of somatic expansion however the inverse was found to be true. These findings provide a different perspective on HD and emphasize the need for the development of African inclusive therapeutic approaches.

In the *JPH3* analysis, less sequence diversity was observed across all individuals affected and unaffected with HDL2. The disease alleles all had the same typical allele structure, which occurred on a single disease haplotype 2: CCATC. The haplotype 2 was also only identified in African ancestry non-disease alleles lending support of an African specific single origin of the HDL2 mutation. The CTG repeat was confirmed to be the main driver of disease severity although when compared with the CAG repeat, it accounted for considerably less of the

variation in the disease phenotype. When the broader haplotype background was considered, the larger haplotype 6: CCCATCG was associated with an earlier age of onset of disease by 17.2 years however the outer two tag-SNPs appear less informative, and its mechanism is uncertain. The CTG repeat showed considerably less somatic expansion than the CAG repeat and unlike HD, somatic expansion was not significantly associated with the age of onset in HDL2 although the sample size of 28 HDL2 patient was smaller than the 60 HD patients. A major limitation when studying rare diseases is the sample size which inevitably reduces the power to identify significant associations.

The comparison of the *HTT* and *JPH3* loci indicated that despite the clinical similarities between HD and HDL2, stark contrasts were identified that suggest a different potential mechanism of disease modification may be occurring in the two diseases. With the only similarity seen in the association with the disease-associated repeats, making the clinical similarity more intriguing. This study has emphasised the importance of studying genetically diverse populations and shown the locus heterogeneity within the *HTT* locus, the disease phenotype and highlighted differences among an HD phenocopy.

Acknowledgments

Looking back on the last four years and the body of research that is within this thesis I have an overwhelming sense of pride that has come from the grit and determination required to see this research to completion. I am immensely grateful for all the lessons I have learnt that go far beyond the growth I have experienced as a postgraduate student but into other aspects of my life. I have loved this research and all that I have been able to experience but it has not come without a lot of challenges and discomfort. Despite this, I do believe that all the doubt and sleepless nights have been worth it. All of this would not have been possible without a very long list of people who have supported me through this journey.

I must start by thanking my supervisors; Dr Fiona Baine-Savanhu, Prof Amanda Krause and Prof Darren Monckton. I am incredibly lucky to have had the guidance that you all provided me - it has been invaluable and ultimately allowed for better research to be accomplished. I am grateful for your kind encouragement, wealth of knowledge and understanding throughout the many years. Dr Baine-Savanhu and Prof Krause I am so thankful you took a chance on me and looking back at all that this research has allowed me to experience I am just so appreciative. The assistance and support you have provided, and the generosity of your time has allowed for so much growth. Prof Monckton thank you for allowing me to learn and develop my laboratory skills over the two research visits to the University of Glasgow and for the funding you generously provided that aided a lot of this research.

I also must thank the whole Monckton group at the Institute of Molecular Cell & Systems Biology, University of Glasgow, in particular to Dr Marc Ciosi. Dr Ciosi your continued help and advice in the laboratory and statistical analyses have been an absolute blessing. To the Division of Human Genetics, thank you for being my home for the last four years. I have grown so much through all that I have been exposed to, whether it be the weekly seminars or journal clubs, the various courses I have been given the opportunity to attend and participation in the conferences that the Division has funded. A special thanks must also go to the postgraduate students who have been a sounding board for me over the years and helped to keep me sane.

None of this would have been possible without the funders who have generously supported this research. I must acknowledge the following bodies for both personal and research funding - the University of the Witwatersrand postgraduate merit awards, the Faculty Research Committee grants, the National Health Laboratory Service research trust grant, the South African Medical Research Council funding from Prof Krause, Cure Huntington's Disease Initiative funding from

Prof Monckton and the National Research Foundation for multiple travel awards and postgraduate scholarships.

I would also like to thank all my wonderful friends and family for the unconditional love and support during these years. I hope I can regain some sense of normality once this is over and return all the care you have shown me. To my mother, thank you for being the rock of a parent you are and for providing me with every opportunity possible and your unwavering support. Lastly to my husband James, I thank you for all the love, kindness, and motivation you have provided me when I needed it the most.

This thesis is dedicated to my loving parents

Gemey Valentine Abrahams

and

the late Howard Anthony Levesley

Table of Contents

Declaration i
Presentations from this study ii
Abstract iii
Acknowledgments v
Table of Contents vii
List of Figures xi
List of Tables xiii
List of Abbreviations xiv
Nomenclature xvi
Terminology for the South African population groups xvii
Outline of the thesis sections xviii
Section 1 xix
1. Chapter 1: Introduction 1
1.1. Trinucleotide repeat disorders 1
1.2. Huntington disease 2
1.2.1. History of Huntington disease 2
1.2.2. HD molecular genetics 3
1.2.3. Mutant HTT protein and proposed pathogenesis 5
1.2.4. Clinical presentation and neuropathology of HD 6
1.2.5. HD prevalence 7
1.2.6. HD haplotypes 8
1.3. Huntington disease-like 2 11
1.3.1. History of HDL2 11
1.3.2. HDL2 molecular genetics 12
1.3.3. Mutant JPH3 protein and proposed pathogenesis 13
1.3.4. Clinical phenotype and neuropathology of HDL2 15
1.3.5. HDL2 prevalence 16
1.3.6. HDL2 haplotypes 17
1.4. Potential modifiers of HD 19
1.4.1. CAG repeat 19
1.4.2. *HTT* repeat sequence 19
1.4.3. *HTT* haplotypes 20

1.4.4. Somatic expansion of CAG repeat 21
1.5. Potential modifiers of HDL2 21
1.6. Genetic testing approaches 22
1.7. Potential treatments 22
1.8. Justification for the study 23
1.9. Aim and Objectives 25
Section 2 29
2. Chapter 2: Methodology 30
2.1. Sample selection 30
2.1.1. HD patients and unaffected controls 30
2.1.2. HDL2 patients and unaffected controls 30
2.2. Sample preparation 31
2.3. Library preparation and sequencing 32
2.3.1. *HTT* repeat tract PCR optimisation 32
2.3.2. *JPH3* repeat tract PCR optimisation 32
2.3.3. PCR amplification of the *HTT* and *JPH3* repeat tracts 33
2.3.4. Cleaning and assessment of sequencing libraries 34
2.3.5. Quantification of library and MiSeq sequencing 35
2.4. Sequencing analysis of repeat tracts 36
2.4.1. Data analysis pipeline 36
2.4.2. Repeat sizing and variant identification 37
2.5. Haplotype analysis 40
2.5.1. Selection of *HTT* and *JPH3* tag-SNPs 40
2.5.2. Genotyping of tag-SNPs 40
2.5.3. Haplotype phasing 41
2.6. Somatic expansion analysis 42
2.7. Analysis of potential modifiers of disease 44
2.7.1. Repeat tract analysis 44
2.7.2. Haplogroup/haplotype analysis 45
2.7.3. Somatic expansion analysis 45
Section 3 47
3. Chapter 3: *HTT* analysis 48

3.1. Results of *HTT* analysis 48
3.1.1. *HTT* repeat tract sequence diversity 48
3.1.2. *HTT* haplogroup/haplotype diversity 53
3.1.3. Potential modifiers of the HD phenotype 57
3.2. Discussion of *HTT* analysis 72
3.2.1. *HTT* repeat tract sequence diversity 72
3.2.2. *HTT* haplogroup/haplotype diversity 74
3.2.3. Potential modifiers of the HD phenotype 76
4. Chapter 4: *JPH3* analysis 81
4.1. Results of *JPH3* analysis 81
4.1.1. *JPH3* repeat tract sequence diversity 81
4.1.2. *JPH3* haplogroup/haplotype diversity 88
4.1.3. Potential modifiers of the HDL2 phenotype 93
4.2. Discussion of *JPH3* analysis 104
4.2.1. *JPH3* repeat tract sequence diversity 104
4.2.2. *JPH3* haplogroup/haplotype diversity 107
4.2.3. Potential modifiers of HDL2 108
Section 4 113
5. Chapter 5: Conclusion 114
5.1. Summary of findings 114
5.1.1. *HTT* analysis 114
5.1.2. *JPH3* analysis 116
5.2. Limitations and unanswered questions 119
5.2.1. *HTT* analysis 119
5.2.2. *JPH3* analysis 119
5.3. Future prospects 120
5.3.1. *HTT* analysis 120
5.3.2. *JPH3* analysis 123
5.4. Concluding remarks 124
Reference list 126
Appendices 140
Appendix 1. Ethics clearance certificate 141

Appendix 2.1. Demographic information for patients affected with HD. 142

Appendix 2.2. Demographic information for patients affected with HDL2. 144

Appendix 3. *HTT*/*JPH3* PCR reagents and amplification conditions. 145

Appendix 4.1. *HTT* MiSeq PCR primers. 146

Appendix 4.2. *JPH3* MiSeq PCR primers. 147

Appendix 5. *HTT*/*JPH3* tag-SNPs used for haplotype construction. 148

Appendix 6. Haplotype B2 frequencies in populations from the 1,000 Genomes Project. . 149

Appendix 7. *JPH3* repeat tract sequence and variants that cause atypical structures. 150

Appendix 8. *JPH3* larger haplotypes identified in disease and non-disease alleles. 151

Appendix 9. Plagiarism form. 153

Appendix 10. Submissible article. 154

List of Figures

Figure 1.1. Trinucleotide repeat expansion disorders and corresponding repeat expansions. 1
Figure 1.2. The *HTT* CAG repeat size ranges for unaffected and affected individuals with HD. 3
Figure 1.3. The *HTT* repeat tract showing the reference gene structure (LRG_763).....4
Figure 1.4. The 22 tag-SNPs spanning the *HTT* gene used to construct haplotypes.....10
Figure 1.5.The *JPH3* gene structure, showing the full-length transcript cDNA and the three open reading frames containing the CTG repeats. 12
Figure 1.6. The *JPH3* haplotypes were constructed using the three SNPs and three microsatellites upstream and downstream of the *JPH3* gene. 18
Figure 2.1. The *HTT* and *JPH3* repeat tracts showing the composition of the MiSeq-compatible primers used for library preparation and sequencing. 34
Figure 2.2. The procedural steps followed in the bioinformatics pipelines used for the sequencing analysis of the *HTT* and *JPH3* repeat tracts. 39
Figure 2.3. The sequencing read count distribution for the CAG/CCG repeat of an individual affected with HD. 43
Figure 3.1. Schematic representation of the *HTT* disease and non-disease allele structures defined for this study. 51
Figure 3.2. Haplotype B2 estimated frequencies for populations in the 1,000 Genomes Project, using the LDhap tool. 56
Figure 3.3. Linear regression analysis testing the association between the log transformed AoD and AoO.....57
Figure 3.4. Linear regression analysis testing the association between the log transformed AoD and the disease associated CAG repeats.....58
Figure 3.5. Linear regression analysis testing the association between the log transformed AoD and the CAG repeats for each disease allele structure. 60
Figure 3.6.The estimated marginal mean AoD for the disease allele structures, corrected for repeat size. 62
Figure 3.7. The estimated marginal mean AoD for the haplotypes, corrected for CAG repeat size. 65
Figure 3.8. The frequency distribution of the R-squared difference between the Q^1-2-0-9-2 allele structure and haplotype B2 models. 66
Figure 3.9. Linear regression analysis testing the association between the log transformed ratio of somatic expansion and CAG repeats for each disease allele structure. 67
Figure 3.10. The estimated marginal mean expansion score for the allele structures, corrected for CAG repeat size and age at sampling. 71
Figure 4.1. Schematic representation of the JPH3 repeat tract allele structures in the non-disease and disease alleles.....84
Figure 4.2. The CTG repeat distributions and estimated mean repeat size in the non-disease alleles from various Africa, European and Asian populations. 87
Figure 4.3. Linear regression analysis testing the association between the disease associated CTG repeat and the log transformed AoO.....94
Figure 4.4. Linear regression analysis testing the association between the disease associated CTG repeat and the between log transformed AoD. 95

Figure 4.5. The estimated marginal mean AoO for each of the larger haplotypes, corrected for CTG repeat size. 99
Figure 4.6. Linear regression analysis testing the association between CTG repeat and the ratio of somatic expansion. 101
Figure 4.7. The comparison of the CAG/CTG repeat association with somatic expansion in the individuals affected with HD and HDL2. 103

List of Tables

Table 2.1. The African ancestry individuals affected and unaffected with HD. 30
Table 2.2. The individuals affected and unaffected with HDL2 from various worldwide populations. 31
Table 3.1. A summary of the African ancestry *HTT* disease and non-disease allele structures. 50
Table 3.2. A summary of the *HTT* haplogroups/haplotypes and associated allele structures. .. 54
Table 3.3. Linear model testing the effect of the individual components of the *HTT* repeat tract on the AoD. 59
Table 3.4. Linear model testing the effect of the allele structures on the AoD. 61
Table 3.5. Linear model testing the effect of the haplogroup background on the AoD. 63
Table 3.6. Linear model testing the effect of the haplotype background on the AoD. 64
Table 3.7. Linear model testing of the effect of the CAG repeat and age at sampling on the ratio of somatic expansion. 68
Table 3.8. Linear model testing of the effect of the CAG repeat, age at sampling and the allele structures on the ratio of somatic expansion. 69
Table 3.9. Linear model testing of the effect of the CAG repeats, allele structure and the expansion score on the AoD. 70
Table 4.1. A summary of the *JPH3* disease and non-disease associated allele structures. 83
Table 4.2. A summary of the *JPH3* non-disease allele structures in the various populations. . 85
Table 4.3. A summary of the *JPH3* haplogroups in the non-disease and disease alleles. 89
Table 4.4. A summary of the *JPH3* haplotypes in the non-disease and disease alleles. 90
Table 4.5. A summary of the *JPH3* haplotypes in the non-disease alleles for each population. 92
Table 4.6. Linear model testing the effect of the disease associated CTG repeat on the AoO. 96
Table 4.7. Linear model testing the effect of the disease and non-disease associated CTG repeats and the sex of the HDL2 patients on the AoO. 97
Table 4.8. Linear model testing the effect of the larger haplotypes on the AoO. 98
Table 4.9. Linear model testing the effect of the CTG repeat and age at sampling on the ratio of somatic expansion. 102
Table 4.10. Linear model testing the effect of the CTG repeat, larger haplotypes and the expansion score on the AoO. 102

List of Abbreviations

ALS - Amyotrophic lateral sclerosis

AoD - Age of Diagnosis

AoO - Age of Onset

ASO - Antisense oligonucleotide

AAV - Adeno-Associated Virus

BAM - Binary Alignment Map

BLAST - Basic Local Alignment Search Tool

BWA-MEM - Burrows-Wheeler alignment tool

CAG - Cytosine Adenine Guanine Nucleotides (encoding Glutamine amino acid)

CAACAG - Polyglutamine Intervening Sequence in *HTT*

CCGCCA - Polyproline Intervening Sequence in *HTT*

CCG - Cytosine Cytosine Guanine Nucleotide (Proline amino acid)

cis-acting - Acting on the same DNA molecule as the gene transcribed

CTG - Cytosine Tyrosine Guanine Nucleotide (Leucine amino acid)

DNA - Deoxyribonucleic acid

dsDNA - Double-Stranded DNA

DDNTPs - Dideoxynucleoside triphosphate

DM1 - Myotonic dystrophy type 1

DMPK - DM1 protein kinase (gene)

DMSO - Dimethyl Sulfoxide

DRPLA - Dentatorubral-pallidoluysian atrophy

FTLD - Frontotemporal Lobar Degeneration

HD - Huntington Disease

HDL2 - Huntington Disease-like 2

HTT - Huntingtin (gene)

IA - Intermediate Allele

IT15 - Interesting transcript 15 (gene)

JPH3 - Junctophilin-3 (gene)

JHD - Juvenile Huntington Disease

MAPQ - Map Quality

MBNL1 - Muscleblind-like protein 1

$MgCl_2$ - Magnesium Chloride

NGS - Next-Generation Sequencing

NHLS - National Health Laboratory Service

NTC - No-Template Control

PCR - Polymerase Chain Reaction

qPCR - Quantitative Polymerase Chain Reaction

RAN - Repeat-Associated Non-ATG

RNA - Ribonucleic Acid

SAM - Sequence Alignment Map

SAP - Shrimp Alkaline Phosphatase

SBMA - Spinal and Bulbar Muscular Atrophy

SBS - Sequencing-By-Synthesis

SCA - Spinocerebellar Ataxia

SNP - Single Nucleotide Polymorphism

SNV - Single Nucleotide Variant

SPRI - Solid Phase Reversible Immobilisation

ssDNA - Single-Stranded DNA

Tm - Melting Temperature

TRD - Trinucleotide Repeat Disorders

tSNPs - Tag-Single Nucleotide Polymorphism

Wits - The University of The Witwatersrand

Nomenclature

Admixture - refers to the genetic lineage resulting from interbreeding of individuals from different genetic backgrounds.

Black African population group - refers to individuals who have genetic ancestral contributions from the African Khoisan and Bantu-speaking peoples of Africa.

Coloured population group - refers to South African individuals with genetic ancestral contributions from the African Khoisan, Bantu-speaking peoples and from both European and Asian populations.

Haplogroup - is a group of related haplotypes with a set of common core tag-SNPs.

Haplotype - is made up of a group of DNA polymorphisms that are closely linked and tend to be inherited together.

HD phenocopy - A disorder that presents with features characteristic of HD but is a result of a different genetic mutation.

Intermediate allele - the CAG repeat range between 27 and 35 repeats which exhibits meiotic instability on intergenerational transmission, and with the potential to expand into the HD-affected range in successive generations.

Modal repeat - the CAG repeat size with the highest read count and usually inferred to be the progenitor CAG repeat size.

Reduced penetrance *HTT* allele - the CAG repeat range between 36 to 39 repeats, with individuals likely to only show clinical symptoms at a very late age or remain asymptomatic.

Somatic expansion - the increase in repeat size in somatic cells due to instability of the repeats.

Tag-SNPs - single nucleotide polymorphisms selected that are in linkage disequilibrium and represent haplotypes.

PCR slippage - PCR artefacts that occur due to polymerase slippage during PCR amplification and typically generate repeats shorter than the modal repeat size.

Prevalence - the number of affected individuals with a condition in a given described group.

p-value - a measure of statistical significance. The estimated probability of rejecting the null hypothesis (H0) when that hypothesis is true. A p-value of ≤ 0.05 is significant and indicates evidence against the null hypothesis.

White population group - refers to individuals with genetic ancestral contributions from European ancestry populations.

Terminology for the South African population groups

South Africa is made up of various population groups which have been defined by the South African Government (Statistics South Africa: Census 2016) and self-identified by the South African people into the following: Black African, White, Coloured and Indian/Asian. These self-identified terms have been used throughout the thesis to explore the underlying molecular genetics of HD in South Africa and are not intended to hold any negative connotations.

Outline of the thesis sections

Section 1 incorporates the Introduction and Aims and Objectives chapter.

Section 2 incorporates the Methodology chapter.

Section 3 incorporates the Result and Discussion chapters.

Section 4 incorporates the Conclusion chapter.

Section 1

Chapter 1 of the thesis introduces and expands on Huntington disease and Huntington disease-like 2 and the various aspects of the underlying molecular genetics. These include the diversity of the repeat tracts and haplogroup/haplotypes and then the effect of these as potential modifiers of disease phenotype and also includes the effect of somatic expansion. The overall aim and objectives of the research are also outlined and broken down for each aspect.

1. Chapter 1: Introduction

1.1. Trinucleotide repeat disorders

Trinucleotide repeat disorders (TRDs) are a group of genetic diseases where the underlying mutation is the expansion of a trinucleotide repeat sequence (Harper, 1997; Orr and Zoghbi, 2007). Disease symptoms are manifest when the number of trinucleotide repeats exceeds a gene-specific threshold (Harper, 1997). In many TRDs, the expanded repeat length is inversely correlated with age of onset of disease symptoms (Hofferbert et al., 1997; Squitieri et al., 1994). Some of the TRDs with the corresponding triplet repeat sequence in the regions of expansion are shown in Figure 1.1.

The expanded repeat tracts associated with TRDs are highly unstable in nature with larger repeats being more likely to expand on transmission (Duyao et al., 1993). The discovery of this unstable nature of repeat expansions provided a molecular explanation for the variability in phenotype seen in most TRDs (Richards and Sutherland, 1997). However, the reasons for the instability of the expanded repeat tracts are poorly understood. To date at least 17 TRDs have been identified, with 10 resulting from CAG repeat expansion mutations located within a polyglutamine-containing open reading frame. The polyglutamine disorders include dentatorubral-pallidoluysian atrophy (DRPLA), Huntington disease (HD), spinal and bulbar muscular atrophy (SBMA) and several types of spinocerebellar ataxia (SCA) 1, 2, 3, 6, 7, 12 and 17 (Holmes et al., 1999; Koide et al., 1999; Ross, 1995; Zoghbi and Orr, 2000). All are neurodegenerative disorders, characterised by neuronal dysfunction, which progresses to neuronal loss in different regions of the brain (Zoghbi and Orr, 2000). Huntington disease (HD) was the fourth condition in which a triplet repeat expansion was found to be the disease-causing mutation (The Huntington's Disease Collaborative Research Group, 1993).

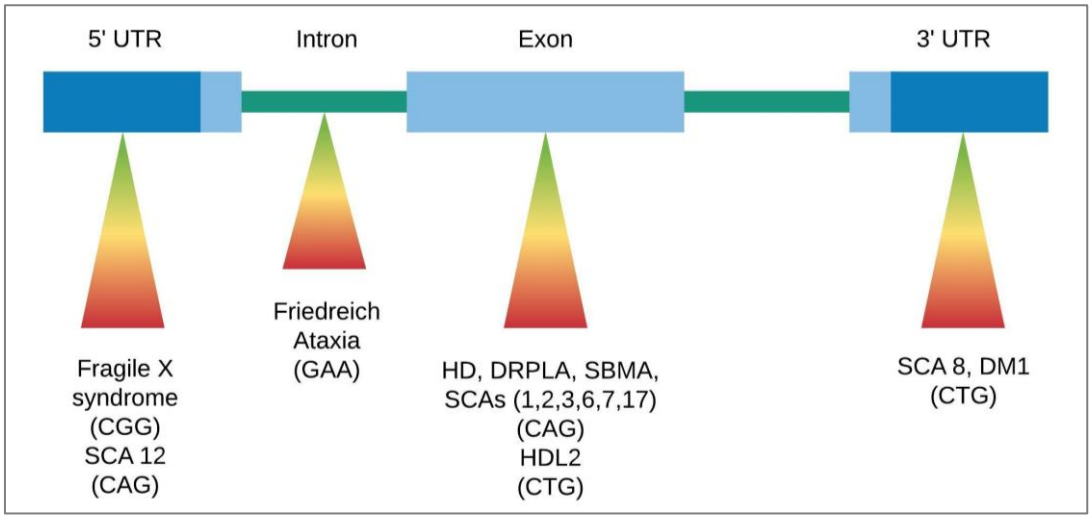


Figure 1.1. Trinucleotide repeat expansion disorders and corresponding repeat expansions. Figure adapted from (Budworth and McMurray, 2013).

1.2. Huntington disease

1.2.1. History of Huntington disease

The first succinct description of Huntington disease was in 1872 by Dr George Huntington, an American physician practising in East Hampton, New York (Huntington, 1872). After his graduation in 1871 from the College of Physicians and Surgeons of Columbia University, Huntington started to report cases of middle-aged patients presenting with dementia and chorea (Bhattacharyya, 2011). These patients were part of several large families which he studied in detail; some of whom had been previously described by his father and grandfather, both general practitioners in East Hampton over a 78 year period (Huntington, 1872).

In 1872, Huntington presented his findings entitled "On chorea" to the Meigs and Mason Academy of Medicine, in which he reported a disease he termed "hereditary chorea". The paper described adult patients experiencing a gradual increase in spasmodic movement to every muscle in the body and cognitive decline until death (Huntington, 1872). Huntington also noted the hereditary nature of the disease, which did not skip a generation to manifest in another. Although reports of these clinical features were made as early as 1816, it was Huntington's observation of the key features which earned him the eponym of Huntington chorea, later changed to Huntington disease (Bhattacharyya, 2011).

Despite the description of HD, the underlying disease-causing mutation remained unknown for over 120 years. Gene mapping efforts were made in the 1970s; however, these were unsuccessful due to a lack of power of the markers and insufficient large families to study (Harper, 1999). It was not until 1983 when a team from Harvard Medical School identified, through linkage analysis, a marker located on the short arm of chromosome 4 which segregated with HD (Gusella et al., 1983). This was made possible through the advent of methods to detect DNA polymorphisms and the study of the single largest population of HD-affected individuals from the Lake Maracaibo region in Venezuela, all of whom shared a common ancestor (Gusella et al., 1983; Harper, 1999). The founder effect could be traced back to one woman who lived in the 18th century, Maria Concepcion Soto (Bhattacharyya, 2011). It was this study of 18,149 individuals from 10 generations of Venezuelan HD kindreds by the Hereditary Disease Foundation that allowed for the localisation and discovery of the mutation that causes HD (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004).

The specific location of the HD mutation was narrowed down to chromosome 4p16.3, through linkage analysis. A 700 kb region between markers D4S180 and D4S182 was shown to be in linkage disequilibrium with HD (MacDonald et al., 1992). An exon amplification approach targeting this region allowed the identification of Interesting transcript 15 (*IT15*); later renamed

the Huntingtin *(HTT)* gene (The Huntington's Disease Collaborative Research Group, 1993). The *HTT* gene is large, spanning approximately 210 kb and contains a polymorphic CAG trinucleotide repeat tract (The Huntington's Disease Collaborative Research Group, 1993).

1.2.2. HD molecular genetics

Huntington disease is caused by the expansion of a CAG repeat tract in exon 1 of the *HTT* gene on chromosome 4p16.3 (Harper, 1999; The Huntington's Disease Collaborative Research Group, 1993). HD follows an autosomal dominant inheritance pattern; therefore, children of an affected individual are at 50% risk of inheriting the disease and only one expanded *HTT* gene is required to be affected. Compound heterozygotes for the expansion have however been identified, with a similar phenotype and only the rate of disease progression is enhanced and is best described by the larger of the two CAG expansions (Walker, 2007).

The CAG repeats can be categorised based on repeat lengths, as shown in Figure 1.2. In unaffected individuals, the CAG repeat range is less than 26 CAG repeats, whereas affected individuals carry full penetrance alleles with 40 or more CAG repeats (Roos, 2010; Walker, 2007). Intermediate range alleles (IAs) have between 27 to 35 CAG repeats, exhibit meiotic instability on intergenerational transmission and have the potential to expand into the HD-affected range (Goldberg et al., 1993; Semaka et al., 2006). Lastly, there is a reduced penetrance range for alleles carrying between 36 and 39 CAG repeats. Individuals carrying CAG repeats in this range are likely to only show clinical symptoms at a very late age or may remain asymptomatic (Rubinsztein et al., 1996).

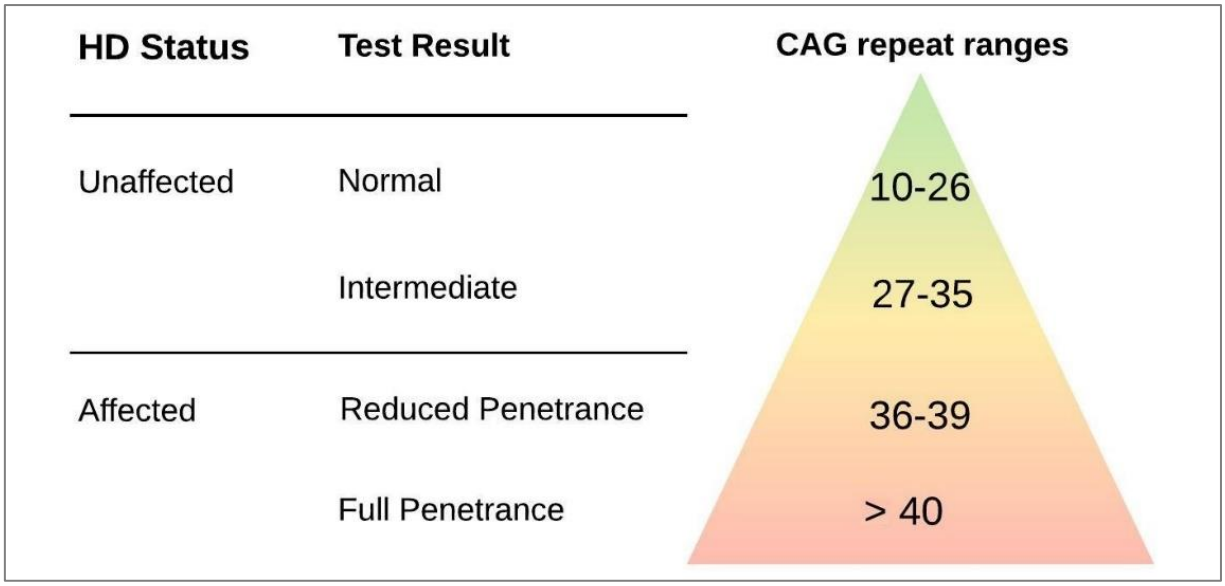


Figure 1.2. The *HTT* CAG repeat size ranges for unaffected and affected individuals with HD. Figure adapted from (Budworth and McMurray, 2013).

Typically, HD manifests as a late-onset disorder with diagnosis in adulthood between 30 and 50 years, with a mean age of onset of 40 years (Ross and Tabrizi, 2011). However, early-onset may occur below the age of 20 years and is termed juvenile Huntington disease (JHD) (Orr and Zoghbi, 2007; Roos, 2010), as a result of anticipation, where an earlier age of onset and increased disease severity occurs with successive generations (Harper, 1997; Ross, 1995). JHD has been shown to have a more rapid progression than adult-onset HD and CAG repeat size generally exceeds 60 repeats (Duyao et al., 1993). Approximately 80% of JHD cases are the result of a paternal transmission, which can be attributed to substantial increases in expansion length occurring during male gametogenesis (Merritt et al., 1969; Nance and Myers, 2001; Telenius et al., 1994).

Like most other TRDs, there is an inverse correlation between the length of the expanded CAG repeat and the age of onset; longer repeats are generally associated with earlier disease onset (The Huntington's Disease Collaborative Research Group, 1993). The length of expanded CAG repeat is unstable in the germline, with repeat length expansions and sometimes contractions in successive generations (Duyao et al., 1993). The CAG repeat expansions occur on both paternal and maternal transmissions however the rate of repeat expansion is lower for female than male transmissions, that tend to have much larger repeat expansion (Kremer et al., 1995). Repeat length instability is also present in somatic cells with contributing factors such as the length of the repeat, the age of the individual and the cell type (Kennedy et al., 2003; Telenius et al., 1994; Veitch et al., 2007).

Although HD is characterised by a CAG repeat expansion, the *HTT* repeat tract is composed of other sequence components. In the *HTT* repeat tract, the human reference sequence is composed of a polyglutamine and polyproline encoding region (locus reference genomic: LRG_763), as shown in Figure 1.3. The polyglutamine region includes a variable number of CAG repeats (Q^1) and a CAACAG (Q^2 = 2 triplets) sequence. This is followed by the region coding for polyproline which comprises a CCGCCA (P^1 = 2 triplets) sequence, a variable number of CCG (P^2) repeats and downstream CCT repeats (P^3).

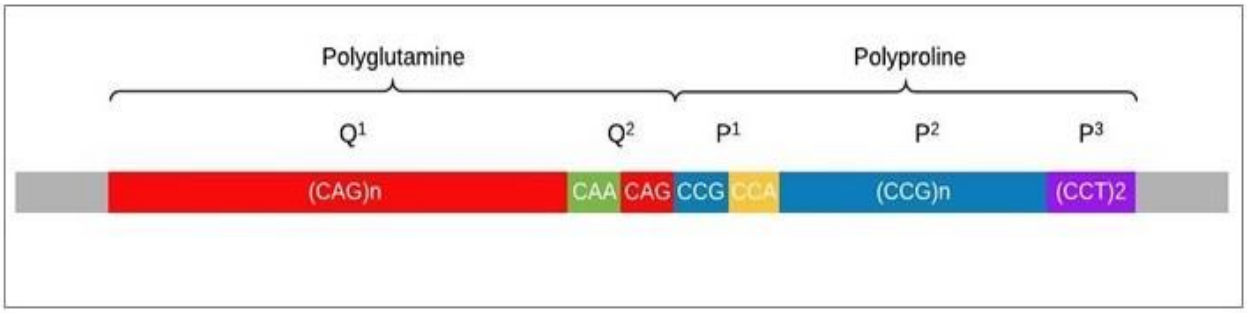


Figure 1.3. The *HTT* repeat tract showing the reference gene structure (LRG_763).

1.2.3. Mutant HTT protein and proposed pathogenesis

The wild-type huntingtin protein is large, approximately 350 kDa; however, the size varies depending on the number of glutamines. The N-terminus of the protein holds the polymorphic polyglutamine and polyproline repeat tracts (Orr and Zoghbi, 2007; The Huntington's Disease Collaborative Research Group, 1993). A structural feature of the N-terminus is the presence of HEAT (Huntingtin, Elongation factor 3, the regulatory A subunit of protein phosphatase 2A and TOR1) repeats (Andrade and Bork, 1995). The HEAT repeats are made-up of approximately 40 amino acid motifs, similar to those found in solenoid-like scaffolding proteins (Takano and Gusella, 2002) supporting the suggestion that the huntingtin (HTT) protein serves as a scaffold to coordinate complexes of other proteins (Li and Li, 2004).

Huntingtin protein is expressed throughout the central nervous system as well as in peripheral tissues. Although the exact function remains unknown, it is believed to mediate many physiological processes (Schulte and Littleton, 2011). Huntingtin is essential for normal development before birth, with targeted disruption resulting in embryonic lethality in mouse models (Nasir et al., 1995). Various animal homologues have shown that huntingtin is highly conserved through evolution, suggesting its importance in normal functioning (Walker, 2007).

All the polyglutamine TRDs share similar neurodegenerative features however, the expanded CAG repeat occurs in different genes, suggesting that the expanded polyglutamine protein is driving the pathogenesis of neurodegeneration (Harper, 1999). The precise molecular mechanisms underlying pathogenicity in HD have not yet been elucidated; however, several processes are known to contribute to pathogenesis (Walker, 2007). In HD, it has been proposed that the predominant pathological mechanism is through the mutant huntingtin conferring a toxic gain of function, involving proteolysis and polyglutamine aggregation which can then affect several nuclear and cytoplasmic proteins (Roos, 2002; Walker, 2007). The toxic gain of function also influences autophagy, vesicle transport, neurotransmitter release and mitochondrial function (Marxreiter, Stemick and Kohl, 2020; Ross and Tabrizi, 2011).

Other pathological mechanisms have also been proposed, including the toxic effects of the RNA and the loss of function of the *HTT* gene (Ross and Tabrizi, 2011). The loss of function mechanism, results in haploinsufficiency leading to inadequate protein production required for cell functioning (Walker, 2007). However, in other polyglutamine TRDs, the loss of function mechanism has not been shown to have a significant role in pathogenesis (Kratter and Finkbeiner, 2010), which is in line with HD, as heterozygous terminal deletion or the disruption of the *HTT* gene does not result in the HD phenotype in humans (Ambrose et al., 1994). Sense and antisense repeat-associated non-ATG (RAN) translation proteins have been found that occur

due to non-canonical translation initiation mechanisms (Bañez-Coronel et al., 2015). The presence of RAN proteins detected in cell models and in human autopsy brains, revealed the accumulation of senses and antisense, polyalanine, polyserine, polyleucine and polycysteine had toxic effects on neural cells (Bañez-Coronel et al., 2015). Although a recent study has suggested RAN proteins do not critically contribute to HD pathogenesis, as the RAN proteins were not detected in HD knock-in mice models, with only expanded polyglutamine associated with an HD phenotype (Yang et al., 2020).

1.2.4. Clinical presentation and neuropathology of HD

Huntington disease is a progressive neurodegenerative disorder with a clinical phenotype commonly described as a triad of movement disorder, cognitive decline and psychiatric complications (Harper, 1999; Walker, 2007). However, the presentation and severity of these have been shown to vary between patients (Sturrock and Leavitt, 2010). Although the occurrence of major cognitive decline and behavioural or psychiatric symptoms can be present up to 10 years prior to a clinical diagnosis, disease onset and clinical diagnosis is usually defined as the presence of the movement disorder (Sturrock and Leavitt, 2010). After onset of the motor symptoms, there is generally a progressive decline that leads to death between 10 to 20 years later (Sturrock and Leavitt, 2010; Walker, 2007).

The movement disorder is made up of a variety of motor features which include chorea, incoordination, motor impersistence and slowed saccadic eye movements. The motor features seen earlier in the course of disease include; general restlessness, hyperreflexia and fidgety fingers, hands and toes, as well as oculomotor abnormalities with the slowing of saccades (Wild and Tabrizi, 2007a). However, as the disease progresses, there are more obvious signs with choreiform movements presenting in 90% of adult-onset HD patients, accompanied by impairment of voluntary motor functioning, dystonia, parkinsonism and bradykinesia to varying degrees (Wild and Tabrizi, 2007a).

The cognitive decline associated with HD is seen across all patients with the impairment of executive functioning affecting judgment, ability to plan and can result in disorganisation and impulsive behaviour (Margolis and Holmes, 2003). HD patients may also exhibit apathy and loss of initiative (Montoya et al., 2006). The more common psychiatric complications associated with HD include depression, anxiety, irritability and aggression (Sturrock and Leavitt, 2010; Walker, 2007). As a result, the rate of suicide and suicide attempts in HD patients is between 5 to 10% higher than the general population (Baliko, Csala and Czopf, 2004; Di Maio et al., 1993).

The clinical phenotype of JHD is unique and has been termed the Westphal variant as it differs from adult-onset HD (Low, Allsop and Halmagyi, 1974). JHD is usually more severe, with patients having a shorter life expectancy. Chorea is uncommon in children who develop HD within the first decade (Nance and Myers, 2001). An early finding of JHD is a decline in motor milestones and school performance (Walker, 2007). The more common symptoms associated with JHD include; rigidity, bradykinesia, dystonia, dysarthria, seizures, ataxia, behavioural disturbances and cognitive decline (Nance and Myers, 2001; Sun, Zhang and Wu, 2016).

The neuropathology of HD is characterised by neuronal dysfunction and atrophy (Ross and Tabrizi, 2011), prominently seen in the caudate nucleus and putamen (Vonsattel and DiFiglia, 1998), leading to extensive neuronal loss and the destruction of the entire structure (*eg.* cerebral cortex) (Gusella and MacDonald, 1995; Roos, 2010). Another characteristic pathological feature of HD is the presence of nuclear and cytoplasmic inclusion bodies; these large aggregates containing mutant proteins and ubiquitin are present before disease onset (DiFiglia et al., 1997; Gutekunst et al., 1999).

1.2.5. HD prevalence

The prevalence of HD is thought to be determined by an interplay between the introduction of *de novo* HD mutations and the elimination of large HD expansions (Rubinsztein et al., 1994). Although HD is found worldwide, there are marked geographic differences in its prevalence (Harper, 1999; Warby et al., 2011). As mentioned previously the highest prevalence of HD is found amongst the inhabitants of the Lake Maracaibo region in Venezuela, where it is estimated at 700 per 100,000 individuals (Avila-Giron, 1973), as a result of a founder effect and isolation of the population.

Outside of Venezuela, the highest prevalences are reported in Western European populations and European ancestry populations (Harper, 1992; Pringsheim et al., 2012; Walker, 2007); whilst the lowest rates are seen in East Asian and African populations (Chang et al., 1994; Hayden, MacGregor and Beighton, 1980; Nakashima et al., 1996; Warby et al., 2011). A meta-analysis of HD prevalence revealed that in European ancestry populations, which combined North America, Europe and Australia; the prevalence estimate was 5.7 per 100,000 individuals (Pringsheim et al., 2012). More recent prevalence estimates incorporating both genetic and clinical diagnose are 10.6 - 13.7 per 100,000 individuals in European ancestry populations (Bates et al., 2015). The lowest prevalences are reported in East Asian populations, estimated at 0.4 per 100,000 individuals (Pringsheim et al., 2012; Warby et al., 2011). Despite the prevalence differences seen across the various ethnic groups and geographic locations, a worldwide estimate for HD was approximated at 2.71 per 100, 000 individuals (Pringsheim et al., 2012).

In a South African context, HD has been reported in three different population groups however with varying prevalence (Hayden, MacGregor and Beighton, 1980). The first prevalence estimates for HD in the South African population were estimated in the 1970s when HD was thought to be absent or rare in the black African population group and no molecular confirmation was available (Hayden, MacGregor and Beighton, 1980). The prevalence estimates were reported as 2.22, 2.17 and 0.01 per 100,000 individuals in the white, coloured and black African population groups respectively (Hayden, MacGregor and Beighton, 1980). No further prevalence estimates studies were conducted to confirm these findings until recently, where minimum frequencies were reported at 7.8, 2.2 and 0.5 per 100,000 individuals in the white, coloured and black African population groups respectively (Baine, Krause and Greenberg, 2016). These minimum estimates were considerably different from previous reports, with a 25 fold difference reported in the black African population (Baine, Krause and Greenberg, 2016). Despite this considerable increase, when the prevalence in black African population was compared with populations with European ancestry it remained much lower.

The low prevalence in the black African population group may be explained by the lower mean CAG repeat size in the general population (black African ancestry: 16.91 *vs* European ancestry: 18.23) which may make expansions into the pathogenic CAG repeat range less common (Baine et al., 2013; Squitieri et al., 1994). However, other factors such as under-diagnosis as a result of limited access to healthcare services and cultural beliefs may have also contributed to lower estimates (Baine et al., 2013; Krause and Greenberg, 2008; Warby et al., 2011).

In other African countries, an increasing number of HD cases has also been reported; however, prevalence estimates are largely based on anecdotal case reports and ascertainment is incomplete (Pringsheim et al., 2012). A 2014 systematic review of neurodegenerative disease in sub-Saharan Africa noted 19 publications of HD in Africa, with cases reported in Zimbabwe, Tanzania, Nigeria, Uganda, Senegal, Burkina-Faso, Togo and Sudan (Lekoubou, Echouffo-Tcheugui and Kengne, 2014). Initially, the low prevalence of HD in African ancestry individuals was hypothesised to be a result of the absence of HD in these populations, with cases of HD only a result of admixture with European ancestry individuals (Scrimgeour and Simpson, 1992). However, the various unique haplotypes seen in black South African patients (shown in section 1.2.6) suggest a distinct African origin of HD (Baine et al., 2013; Greenberg et al., 1991; Warby et al., 2009).

1.2.6. HD haplotypes

The differences in HD prevalence between populations are associated with differences in mutation rates and the geographic difference in the *HTT* haplotypes (Warby et al., 2011). The

effect of underlying *cis*-acting factors can be studied through analysis of haplotypes (comprised of DNA variants generally inherited as a unit) (Squitieri et al., 1994; Warby et al., 2009). These haplotypes were first studied before the identification of the *HTT* gene using multi-allele polymorphisms which revealed 26 different haplotypes and evidence for more than one origin of the *HTT* mutation (MacDonald et al., 1992). Haplotype analysis conducted in several populations and expanded to include additional polymorphisms along the gene, as well as the associated CAG and CCG repeat size (Rubinsztein et al., 1995; Squitieri et al., 1994), revealed different haplotype distributions and frequencies in different populations, which likely occurred as a result of multiple HD mutation events (Andrew et al., 1993).

A limitation of these studies was the number of markers used, as only a small number of polymorphisms had been characterised along the *HTT* gene (Warby et al., 2009). In 2009, a comprehensive assessment of the *HTT* haplotypes was conducted using 22 tag-single nucleotide polymorphisms (tag-SNPs). The study defined three broad haplogroups, A, B and C. Haplogroup A represented the most common haplogroup in European ancestry patients and is deemed to be the most highly associated with disease (Warby et al., 2009). The haplogroup A variants (A1 and A2) were also most common on chromosomes with intermediate alleles (27 to 35 CAG repeats) in populations of European ancestry, suggesting that these alleles are predisposed to expand into a pathogenic range (Warby et al., 2009). Haplogroup variants A1/A2 thus designated 'high-risk', were absent from the East Asian population, where haplogroup C is the most commonly occurring (Warby et al., 2011).

As with the prevalence, the diversity observed in HD haplogroup variation is specific to the different population groups in South Africa (Baine et al., 2013; Hayden, MacGregor and Beighton, 1980). Disease-associated alleles in the white and coloured population groups occur almost exclusively on the European haplogroup A (Baine et al., 2013; Warby et al., 2009). This contrasts with the black African population group where disease-associated alleles occur predominantly on haplogroups B and C (Baine et al., 2013). The 22 tag-SNPs and alleles with the corresponding haplotypes identified in the South African population groups are shown in Figure 1.4 (page 10). The study by Baine et al. (2013), assigned haplogroups/haplotypes based on 21 previously described tag-SNPs and previously unknown haplotypes were defined through the addition of tag-SNP rs10015979. In this study African haplotypes A6 and A7 were added to haplogroup A, while haplogroup B was sub-divided into haplotypes B1 and B2, and haplogroup C was sub-divided into C and novel South African variant, C-South Africa (C-SA) (Baine et al., 2013).

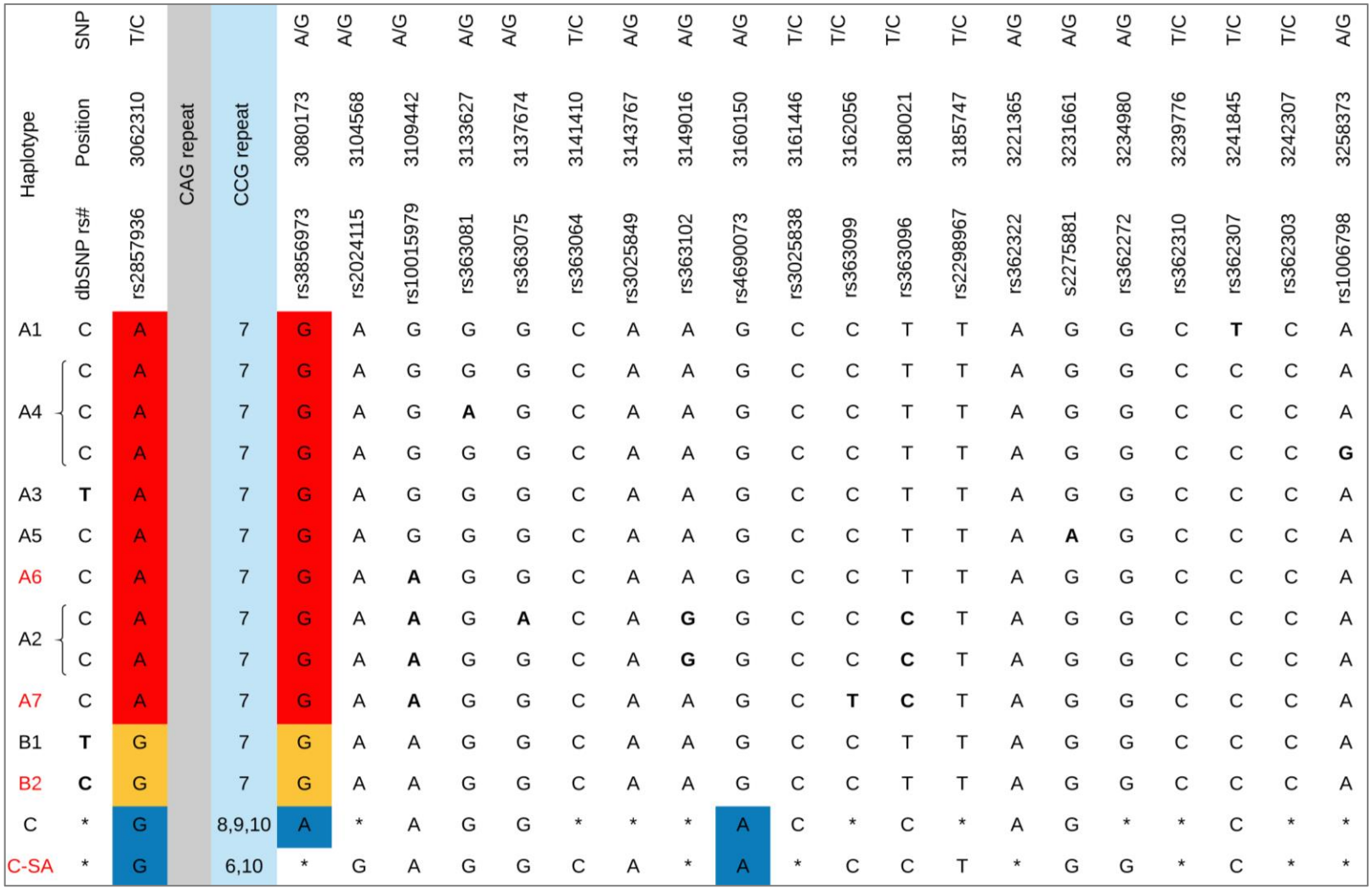


Figure 1.4. The 22 tag-SNPs spanning the *HTT* gene used to construct haplotypes. Figure adapted from (Baine et al., 2013).

The tag-SNP alleles indicated in red, yellow and blue define haplogroups A, B and C, respectively. The tag-SNP alleles indicated in **bold font** represent haplotype-specific tag-SNPs within haplogroups. Asterisks (*) represent variable tag-SNP alleles within haplogroup C and variant C-SA. The haplotypes indicated in red font are unique to the South African population.

1.3. Huntington disease-like 2

1.3.1. History of HDL2

Unlike the South African white population group, the *HTT* expansion in individuals with African ancestry does not account for all clinically diagnosed HD cases (Greenberg, Bardien and Carr, 2008). Up to one-third of black African patients presenting with an HD phenotype are affected by Huntington disease-like 2 (HDL2), a situation that has not been seen elsewhere in the world (Krause, 2015; Krause and Greenberg, 2008).

Huntington disease-like 2 was first described in 2000, in a large African-American pedigree (Pedigree W) from the South-eastern part of the United States (Margolis et al., 2001). The proband of pedigree W was chosen from patients of the Baltimore Huntington Disease Clinic, who had the clinical and pathological phenotype of HD but were negative for the *HTT* expansion. The family members of pedigree W were studied in detail with extensive clinical evaluations and molecular genotyping; for a subset, neuroimaging was performed, and the neuropathology examined. Affected family members were shown to present clinically within features consistent with the spectrum of HD. The neuroimaging was indistinguishable from an HD case of similar disease duration, and the neuropathology was consistent with HD (Margolis et al., 2001).

Despite the striking clinical similarities, the genotyping results of 10 affected family members revealed they were all negative for the *HTT* expansion. At least three affected family members were further tested for other known CAG repeat expansion disorders, DRPLA, SCAs 1, 2, 3, 6, 7, 8 and 12, the TATA-binding protein, *MAB21L1*, *SERS-1* and *Dir/ERDA1* genes; all of which were negative. However, a novel CAG repeat expansion, with between 50 and 60 repeats in affected individuals was identified (Margolis et al., 2001). Due to the similarity with HD, the authors in consultation with the Nomenclature Committee of Human Genome Organisation, termed the disease, Huntington disease like-2. Despite the identification and naming of the repeat expansion disorder, it was not until the following year that the associated gene and its location were discovered (Holmes et al., 2001).

The location of the repeat expansion was identified through the isolation of a 5.5 kb genomic clone from one of the HDL2-affected family members from pedigree W (Holmes et al., 2001). It was mapped by fluorescence *in situ* hybridisation to chromosome 16q23-24 and to an alternatively spliced exon of the Junctophilin-3 (*JPH3*) gene. On further analysis of the *JPH3* gene repeat expansion, HDL2 was found to rather result from an expanded CTG repeat (Holmes et al., 2001). No open reading frames were found in a CAG orientation, this orientation was rather only identified on the non-coding antisense strand (Holmes et al., 2001).

1.3.2. HDL2 molecular genetics

Huntington disease-like 2 is the result of a CTG repeat expansion mutation in the *JPH3* gene on chromosome 16q24.2 (Holmes et al., 2001). Within the *JPH3* gene the CTG repeat expansion is located on the sense strand of the variably spliced exon 2A. The full-length transcript of the gene does not include the CTG repeat as it rather contains exon 2B instead of exon 2A (Holmes et al., 2001; Margolis, Rudnicki and Holmes, 2005). Despite this, the CTG repeat may occur in three reading frames due to the various splice acceptor sites between exon 1 and 2A (Figure 1.5). The reading frames either encode for a polyalanine or a polyleucine tract or it lies in the 3' untranslated region (Holmes et al., 2001).

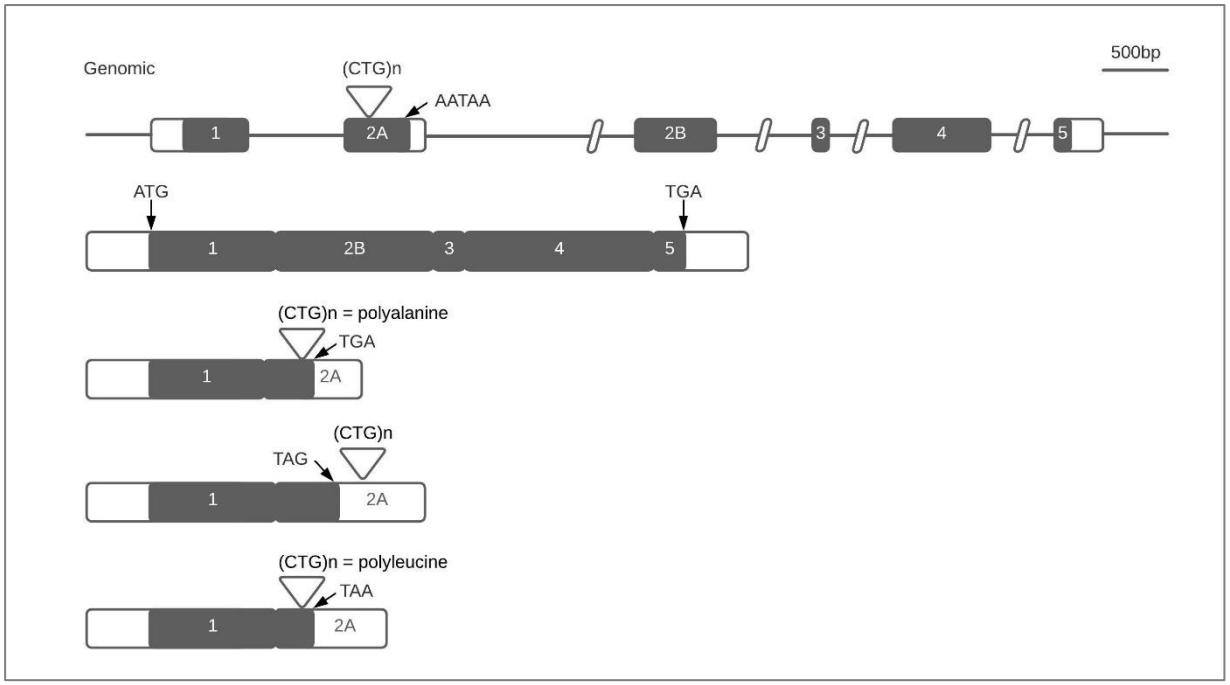


Figure 1.5.The *JPH3* gene structure, showing the full-length transcript cDNA and the three open reading frames containing the CTG repeats.
Figure adapted from (Margolis, Rudnicki and Holmes, 2005)

Like HD, HDL2 is an autosomal dominant late-onset disorder that typically presents in mid-adulthood (Holmes et al., 2001; Margolis, Rudnicki and Holmes, 2005). The most recent findings show an average age of onset of 41 years (Anderson et al., 2017). An inverse correlation between the CTG repeat size and age of onset is seen in HDL2, as individuals with larger repeat expansions have an earlier disease presentation (Margolis et al., 2004). The repeat is also unstable in the germline with anticipation observed, where an earlier onset of disease is seen with successive generations that can result in juvenile-onset as early as 12 years (Anderson et al., 2017).

The CTG repeat length for unaffected individuals is between 6 and 28 CTG repeats, whereas fully penetrant alleles in the disease range for affected individuals generally have 40 or more CTG repeats (Krause et al., 2015; Margolis, Rudnicki and Holmes, 2005). The largest reported expansion was in an individual with 60 CTG repeats (Anderson et al., 2019b). Unlike HD, no definitive intermediate or reduced penetrance range has been determined. However, this does not exclude the possibility that individuals in this uncertain CTG repeat range (29 to 39) will exhibit repeat instability (expansion/contraction) on transmission or very late onset of disease (Anderson, Krause and Margolis, 2004 [Updated 2019 Jun 27]). Repeat sizes within this range of uncertainty should thus be considered with caution as a few individuals have been identified that present with an HDL2-like phenotype (Margolis et al., 2006; Paradisi, Ikonomu and Arias, 2012).

Unlike the *HTT* locus, the *JPH3* repeat tract is only composed of the CTG repeats with no other sequence components considered to be part of the repeat tract. In the *JPH3* repeat tract, the human reference sequence is composed of a polyleucine encoding region (Ensembl: ENST00000301008.5), with a variable number of CTG repeats.

1.3.3. Mutant JPH3 protein and proposed pathogenesis

The first description of the three genes in the junctophilin family occurred in 2000, with different expression patterns noted for each protein (Nishi et al., 2000; Takeshima et al., 2000). Junctophilin-1 is expressed in the skeletal muscle, *JPH2* is expressed in cardiac and skeletal muscles and *JPH3* is expressed primarily in the brain, and to a lesser extent in the testes (Nishi et al., 2000; Takeshima et al., 2000). Members of the junctophilin protein family are composed of eight repeated N-terminal units of 14 amino acids known as MORN (membrane orientation and recognition nexus) motifs, that anchor the protein to the plasma membrane and a C-terminal endoplasmic/sarcoplasmic reticulum transmembrane domain (Nishi et al., 2000; Takeshima et al., 2000).

Based on their structure, the junctophilin proteins appear to establish junctional complexes between the plasma membrane and endoplasmic/sarcoplasmic reticulum. The functioning of these proteins is believed to act by linking plasma membrane voltage sensors with intracellular calcium channels (Margolis and Holmes, 2003). The *JPH3* transcript containing the CTG repeat is however shorter than the full-length transcript as it does not contain the endoplasmic reticulum insertion domain.

HDL2 was initially thought to be a polyglutamine disease due to the inheritance pattern and the fact that the clinical phenotype is nearly indistinguishable from HD (Margolis, et al., 2005). This

was further supported by immunohistochemistry studies that revealed the presence of intranuclear inclusions predominantly in the cerebral cortex. The inclusions were stained with IC2 antibody specific for expanded polyglutamine, ubiquitin, torsinA and TATA-binding protein (TBP) (Margolis et al., 2001; Rudnicki et al., 2008). The antisense strand of the *JPH3* gene containing the CAG has been demonstrated to be transcribed and thus bidirectional transcription does occur (Seixas et al., 2012). Although experimental findings did not reveal that the antisense CAG repeat expansion is expressed in the brain, this may be due to low levels that could not be detected (Seixas et al., 2012). Unlike the *HTT* gene in HD, the sense strand of the *JPH3* gene is made up of CTG repeats that encode polyleucine; the role of the polyglutamine in HDL2 thus remains unclear (Seixas et al., 2012).

The precise molecular mechanisms underlying HDL2 pathogenicity have not yet been elucidated; however, three mechanisms have been proposed. The first involves a loss of function of the full-length JPH3 protein as a result of the reduction of *JPH3* mRNA expression (Margolis and Rudnicki, 2016; Seixas et al., 2012). This is supported by *JPH3* knockout mouse models, where impaired motor coordination and selective neurodegenerative pathology was observed (Nishi et al., 2002; Wilburn et al., 2011).

The second mechanism involves a toxic gain of function of the RNA, resulting in neurotoxicity from the expanded *JPH3* transcript (Rudnicki et al., 2007). This mechanism is supported by cell experiments where cell toxicity was caused by the *JPH3* transcript formation of RNA foci that colocalized with Muscleblind-like protein 1 (MBNL1) (Rudnicki et al., 2007). These RNA foci resemble those formed in myotonic dystrophy type 1 (DM1) caused by CTG repeat expansion in the 3'-untranslated region of the DM1 protein kinase (*DMPK)* gene (Jiang et al., 2004).

The third proposed mechanism is a toxic gain of function resulting from the translation of the antisense CAG encoding a polyglutamine protein (Wilburn et al., 2011). However, this was only supported by the presence of the transcript in the CAG orientation in transgenic BAC-HDL2 mouse models and the detection of polyglutamine-positive intranuclear inclusions as expanded CAG repeats were not seen in the brain of HDL2 affected individuals (Margolis and Rudnicki, 2016; Wilburn et al., 2011).

1.3.4. Clinical phenotype and neuropathology of HDL2

The clinical phenotype is nearly indistinguishable from HD with the associated movement disorder, cognitive decline and psychiatric complications (Margolis et al., 2001; Walker et al., 2003). Affected individuals experience a progressive decline that leads to an early death 15 to 20 years after disease onset (Margolis and Holmes, 2003). The symptomatology is broad and differs depending on disease duration. Previous reports have suggested that HDL2 can present two distinct phenotypes which are associated with either an earlier or later onset. The earlier onset phenotype resembles the Westphal variant of HD and is associated with a Parkinsonian disorder with no oculomotor involvement, whereas the later-onset phenotype is more like typical HD (Margolis, Rudnicki and Holmes, 2005).

The early signs of HDL2 are behavioural changes and a deterioration of social interactions (Walker et al., 2003). The initial movement disorder is associated with hyperkinetic features, manifesting as prominent chorea, worsening involuntary movement and oculomotor abnormalities. However, as the disease progresses the later stages are associated with hypokinetic features characterised by rigidity and bradykinesia with worsening dystonia (Anderson et al., 2019b). The cognitive and psychiatric features associated with HD are also present in HDL2 patients. Although at the early stages of the disease the cognitive features are heterogeneous, as the disease progresses dementia is seen across all patients. Psychiatric features of depression, apathy and prominent irritability are also common (Ferreira-Correia et al., 2019).

A recent study which compared the clinical phenotype of HD and HDL2 has shown that HDL2 patients had a three-year earlier age of diagnosis and poorer test scores for dysarthria and dystonia. This suggests a more severe disease presentation in HDL2 patients despite the fact that the two patient groups could not be distinguished on clinical examination (Anderson et al., 2019b). Remarkably similar neurodegeneration patterns have been noted on brain magnetic resonance imaging (Anderson et al., 2019a; Greenstein et al., 2007; Rudnicki et al., 2008), with the typical features of atrophy of the caudate and cerebral cortex seen in HD, also present in HDL2 patients. The only exception was a greater loss of thalamic volumes noted in HDL2 patients by Anderson et al (2019a); previously reported to have no significant differences from HD (Margolis and Holmes, 2003).

1.3.5. HDL2 prevalence

Unlike HD, there are no precise worldwide estimates of HDL2. Approximately 1% of patients with symptoms suggestive of HD do not have an *HTT* repeat expansion and are said to be HD phenocopies (Margolis et al., 2004). The HD phenocopies include dentatorubral-pallidoluysian atrophy (DRPLA), spinocerebellar ataxias (SCA) 1, 2, 3, HD-like (HDL) 1, 2 and 4 (or SCA17) and the *C9orf72* gene expansion (Wild et al., 2008). Of these, the *C9orf72* gene expansions that are also the major cause of frontotemporal lobar degeneration (FTLD) and amyotrophic lateral sclerosis (ALS) are the most commonly occurring followed by SCA17 in European populations (Hensman Moss et al., 2014; Wild and Tabrizi, 2007b). However, when assessing potential phenocopies, ethnic descent should be used to direct additional testing as HDL2 is the most commonly occurring HD phenocopy in African populations (Margolis et al., 2004).

HDL2 is very rare with only approximately 69 pedigrees described worldwide (Anderson et al., 2017). The largest number of patients described are from South Africa, with 41 pedigrees (Anderson et al., 2017; Baine, Krause and Greenberg, 2016). However, cases have been reported in other countries such as the United States, Brazil, France, Mexico, United Kingdom, Venezuela and Botswana (Margolis et al., 2004; Mariani et al., 2016; Ocampo, Daimari and Oyekunle, 2018; Rodrigues Riccioppo et al., 2008). In Venezuela, HDL2 is estimated at around 1 per 2,000,000 individuals (Paradisi, Ikonomu and Arias, 2012). No cases of HDL2 have been reported in individuals with an HD phenotype from Japan, Germany, and Austria (Bauer et al., 2002; Margolis et al., 2004).

In South Africa, the minimum frequency estimates for HD and HDL2 combined were calculated at 5.10, 2.10 and 0.25 per 100,000 individuals in the white, coloured and black African population groups respectively (Baine, Krause and Greenberg, 2016). However, when examined by the HD: HDL2 ratio, no HDL2 affected individuals were identified in the white population group. The HD: HDL2 ratios were, 65.4:18.5, 34.6:20.6 and 60.9:0 in the black, coloured and white populations groups respectively (Baine, Krause and Greenberg, 2016). The presence of HDL2 in the coloured population is thought to be due to an African ancestral contribution (Krause and Greenberg, 2008; Krause et al., 2015). Similar findings were reported in another study of South African patients with an HD phenotype referred for molecular testing (Krause et al., 2015). Fifteen percent of black patients (20/130, 15.9%) were found to have HDL2, while none of the white patients (0/171, 0%) carried the *JPH3* expansion (Krause, 2015).

All HDL2-affected individuals have been shown to have African or probable African ancestry (Krause and Greenberg, 2008; Krause et al., 2015; Magazi et al., 2008). Santos et al (2008) published the first HDL2 patient with apparent European ancestry; a Brazilian man presenting

with an HD phenotype was confirmed to be affected by HDL2 (Santos et al., 2008). To determine the origin of the *JPH3* expansion, the authors conducted a detailed haplotype analysis which revealed that although the patient had an estimated likelihood of 97.4% of being of European ancestry, the repeat expansion was found on a haplotype that has only been identified in African individuals (Santos et al., 2008). Despite the patient not being aware of any African ancestry within the family, the presence of the African HDL2 haplotype suggests an African contribution.

1.3.6. HDL2 haplotypes

Haplotype analysis for the *JPH3* gene has been very limited by the small number of HDL2 affected individuals. Despite this, two studies have shown that HDL2-affected individuals share a common haplotype (Krause et al., 2015; Paradisi, Ikonomu and Arias, 2012) with an African origin. This has been further confirmed by ethnic marker analysis used to ascertain ethnicity in individuals where an African contribution may not have been apparent (Paradisi, Ikonomu and Arias, 2012; Santos et al., 2008).

Paradisi, Ikonomu et al (2012) used four polymorphisms; the African Duffy null allele (FY*0, at 1q21), the European (Caucasoid) Kell blood group allele K (KEL*K, at 7q33), the Amerindian/Asian Diego A blood group (SLC4A1*A, at 17q21) and the Amerindian specific allele 9 microsatellite D9S1120 (Paradisi, Ikonomu and Arias, 2012), to determine ethnic origin. In three of the four Venezuelan HDL2-affected families studied, an African ancestry origin was determined due to the presence of the African Duffy null marker allele. Although the African marker was not identified in the fourth family, the European and Amerindian/Asian markers were absent in all of the HDL2-affected families (Paradisi, Ikonomu and Arias, 2012).

In the same study, haplotypes were constructed using six intragenic SNPs: rs918368, rs3751725, rs34442559, rs9934222, rs1366520 and rs8053904. The first four SNPs were located in exon 2, the fifth located in intron 3 and the sixth located in intron 4 of the *JPH3* gene (Paradisi, Ikonomu and Arias, 2012). The results revealed a common disease haplotype among the Venezuelan HDL2-affected individuals with three SNPs differentiating African populations from the European and Asian population controls studied. The authors thus proposed a common origin, likely from sub-Saharan Africa, due to the similar SNP frequencies (Paradisi, Ikonomu and Arias, 2012).

In South Africa, Krause et al. (2015) investigated the possible origin of the *JPH3* mutation through haplotype construction using a different set of polymorphisms (Krause et al., 2015). These included three SNPs (rs2544635, rs3849254, rs918368) and three microsatellite markers

(D16S3074, D16S3048, D16S41). The relative location to the CTG repeat tract is shown in Figure 1.6. The analysis was conducted on 22 HDL2-affected individuals with African ancestry from South Africa and 23 HDL2-affected individuals of North American, Mexican, French West Indian and Moroccan origin. The results revealed that core haplotypes constructed using the three SNPs were shared across the patients except for one Moroccan individual in whom the third SNP differed. The authors suggest that this difference could be explained by a single cross-over event (Krause et al., 2015).

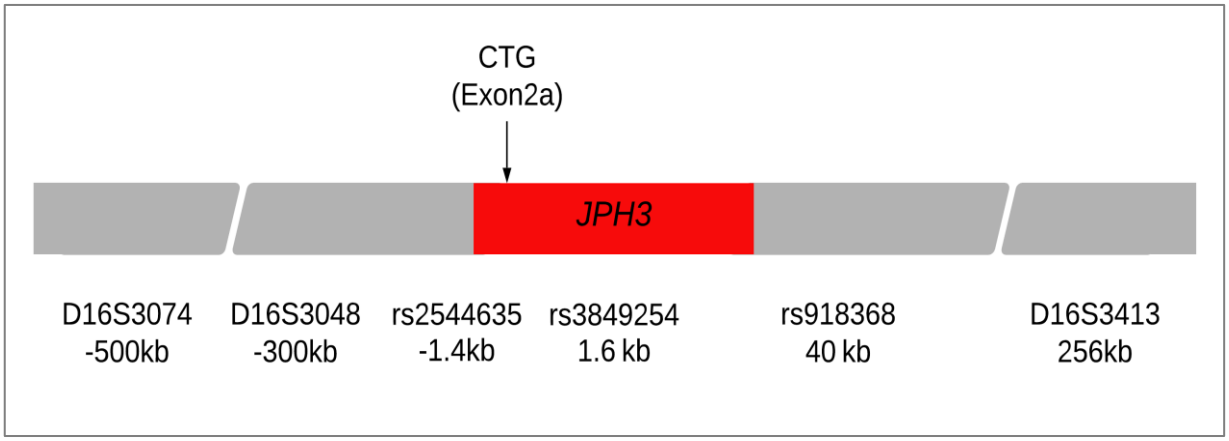


Figure 1.6. The *JPH3* haplotypes were constructed using the three SNPs and three microsatellites upstream and downstream of the *JPH3* gene.
Figure adapted from (Krause et al., 2015).

Like the previous study, the results of the haplotype analysis suggest a single founder origin for the *JPH3* expansion in sub-Saharan Africa (Krause et al., 2015). However, it is unknown whether the HDL2 haplotype acts via *cis*-acting mechanism predisposing the *JPH3* repeat tract to expansion, or the larger and/or expanded CTG repeats occur exclusively on this haplotype (Krause et al., 2015). Analysis of the repeat size distribution for the disease-associated haplotype in the general population may provide insights into the haplotype effect (Krause et al., 2015).

1.4. Potential modifiers of HD

Despite the underlying pathogenic mechanism in HD remaining unclear, various genetic modifiers have been proposed that affect the age of motor onset of HD. Although the main modifier is the repeat size, the remaining variation in the age of onset (AoO) has been shown to be at least partially heritable (Djousse et al., 2003).

1.4.1. CAG repeat

Like other TRDs, there is an inverse correlation between the length of the expanded CAG repeat and the age of onset; longer repeats are generally associated with earlier disease onset (The Huntington's Disease Collaborative Research Group, 1993). Although other components within the repeat tract have been shown to have an effect on the age of motor onset, the CAG repeat size is the primary determinant, accounting for approximately 60% of the variation in the age of onset (Lee et al., 2012). Recent studies have shown that the best predictor of HD severity is the number of pure CAG (Q^1) repeats rather than the total length of glutamines encoded ($Q^1 + Q^2$) along the repeat tract (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019).

The interaction of the unexpanded CAG repeat was also reported to influence the age of onset through an interaction with the expanded CAG repeat (Aziz et al., 2009). To test this, a rigorous statistical analysis of 4,068 individuals was conducted, which revealed only the expanded allele had a significant effect on the age of motor onset (Lee et al., 2012). In the case of compound heterozygotes, the effect of a second expanded CAG repeat was also tested to determine if the presence of a second expanded allele influenced the age of onset; however, only the largest expanded CAG repeat was shown to have an effect (Lee et al., 2012).

1.4.2. *HTT* repeat sequence

Previous research has shown that repeat sequences in the TRDs can contain interruptions which are likely to affect the repeat stability and disease severity (Chung et al., 1993; Eichler et al., 1994; Kraus-Perrotta and Lagalwar, 2016). This has been seen in spinocerebellar ataxia type 1, myotonic dystrophy type 1 and fragile X syndrome (Chung et al., 1993; Eichler et al., 1994; Kraus-Perrotta and Lagalwar, 2016; Liu et al., 2014; Musova et al., 2009). In spinocerebellar ataxia type 1 and fragile X syndrome the stabilising interruptions have been identified on unexpanded alleles, while they are absent or very rare on expanded alleles (Chung et al., 1993; Eichler et al., 1994). Mutations interrupting the repeat tracts in *HTT* may have a similar effect for HD. The entire *HTT* repeat tract sequence is reported to be a hotspot for single nucleotide variation (SNVs) (Gellera et al., 1996; Goldberg et al., 1995; Yu et al., 2000).

The identified mutations (or SNVs) alter the structure of the *HTT* reference sequence, resulting in the apparent duplication or loss of parts of the intervening CAACAG (Q^2) and CCGCCA (P^1) sequences and the downstream CCT (P^3) repeats (Goldberg et al., 1995; Yu et al., 2000). Variants within the repeat tract have been shown to lead to errors in CAG/CCG repeat sizing since they interfere with primer annealing and accurate polymerase chain reaction (PCR) amplification (Yu et al., 2000). However, the effect of this variation on the pathological mechanism and repeat stability was unknown (Yu et al., 2000).

Recent studies have however assessed the effect of the loss or duplication of the intervening CAACAG sequence that encodes for the polyglutamine region of the *HTT* repeat tract (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). These studies, conducted in European ancestry populations, revealed that the loss of the CAACAG sequence (Q^2=0) was associated with earlier age of onset and increased HD severity (Ciosi et al., 2019; Wright et al., 2019). The effect was thought to be mediated through an increased level of somatic expansion however, after correcting for the pure CAG length (Q^1) there was no longer a significant association between the loss of the CAACAG sequence and somatic expansion.

The polyproline region of the *HTT* repeat tract has also been shown to be involved in cytoplasmic localisation of HTT exon 1 protein (Httex 1p) and may be related to the stability and aggregation of huntingtin (Bhattacharyya et al., 2006). However, there is currently no detectable difference in the effect of the CCG length on disease severity.

1.4.3. *HTT* haplotypes

The HD haplotypes have also been studied to assess their effect on the age of onset of symptoms. A study by Nørremølle et al. (2009) conducted on Danish HD-affected individuals revealed that a 4p16.3 haplotype was associated with an 18.8% later age of onset (p = 0.001) (Nørremølle et al., 2009). This haplotype was later shown to be associated with the duplication of the intervening CAACAG sequence. However, other than the rare haplotypes that are associated with variants within the *HTT* repeat tract, there has been no conclusive data from genome-wide association studies that the haplotypes themselves affect the age of onset (GeM-HD Consortium, 2015; GeM-HD Consortium, 2019).

In the European population, haplogroup A was shown to be the most common disease haplogroup and therefore most associated with expanded alleles (Warby et al., 2009). Haplogroup A was also most commonly associated with intermediate alleles (CAG repeats: 27 to 35) in unaffected individuals, suggesting that alleles occurring on haplogroup A are unstable and predisposed to expansion (Warby et al., 2009). Haplotypes A1/A2, which are absent from

low HD prevalence East Asian and African Ancestry populations, were thus designated as high-risk variants (Warby et al., 2011).

1.4.4. Somatic expansion of CAG repeat

CAG repeat instability is observed in the germline and in somatic cells (Duyao et al., 1993; Kennedy et al., 2003). Somatic instability has been previously shown to be affected by various factors such as the length of the repeat, the age of the individual and the cell type (Kennedy et al., 2003; Telenius et al., 1994; Veitch et al., 2007). The highest somatic instability has been noted in the brain of HD-affected individuals, and the striatum is the tissue in which the largest somatic expansions have been observed (Kennedy et al., 2003; Shelbourne et al., 2007; Telenius et al., 1994) with somatic gains in CAG repeats of up to 1,000 detected (Kennedy et al., 2003; Shelbourne et al., 2007; Telenius et al., 1994). Recent data in European ancestry HD patients have confirmed somatic expansion of the *HTT* CAG repeat as a key driver of HD severity (Ciosi et al., 2019). With the level of somatic expansion in blood DNA of HD patients inversely associated with age at onset, and directly correlated with disease progression (Ciosi et al., 2019).

Recent genome-wide association studies (GWAS) have revealed the importance of DNA repair genes/pathways as modifiers of the age of onset (GeM-HD Consortium, 2015; Hensman Moss et al., 2017). Genome-wide significant signals were identified at *RRM2B/UBR5*, *FAN1*, *MSH3/DHFR*, *PMS1*, *PMS2*, *LIG1* and *MLH1* (Correia et al., 2015; GeM-HD Consortium, 2015; GeM-HD Consortium, 2019). Some of these had previously been shown to modify somatic expansion in mouse models (Pouladi et al., 2012), suggesting the effect on age of onset of HD symptoms may be mediated through the somatic expansion of CAG repeats (Ciosi et al., 2019; Wright et al., 2019). These modifier genes provide a potential therapeutic target, not only for HD, but for other repeat expansion disorders too (Ciosi et al., 2019; Flower et al., 2019).

1.5. Potential modifiers of HDL2

Like HD, the underlying pathogenic mechanism in HDL2 remains unclear, however even less is known about potential modifiers of the onset of HDL2. An inverse correlation has however been seen with the larger CTG repeats associated with an earlier age of onset of disease (Margolis et al., 2004). Within the *JPH3* repeat tract no variants have been identified within the expanded CTG repeat tract. However, a rare cytosine to thymine variant that interrupts the CTG repeat has been reported in unaffected European and Asian populations (Lek et al., 2016). All the HDL2 patients share a common haplotype, suggesting a single founder origin of the *JPH3* expansion (Krause et al., 2015). However, it is unknown whether the haplotype is predisposing to expansion through a *cis*-acting mechanism or rather if the larger unexpanded and/or expanded CTG repeats only occur on this haplotype (Krause et al., 2015). Somatic expansion of the CTG

repeats have also not been assessed in HDL2 patients thus it is unknown if a similar mechanism that is mediated by somatic expansion is occurring in HDL2.

1.6. Genetic testing approaches

The established diagnostic molecular genetic tests that genotype the triplet repeat length associated with HD, HDL2, and other expansion-based disorders are based on polymerase chain reaction (PCR) amplification, followed by capillary electrophoresis to estimate the repeat size (Losekoot et al., 2013). However, this standard approach does not examine the sequence itself and therefore cannot identify variants within the repeat tract. Variants within the repeat tract have been shown to lead to erroneous results in the molecular sizing of *HTT* alleles as they interfere with primer annealing and accurate amplification (Ciosi et al., 2019; GeM-HD Consortium, 2019; Yu et al., 2000).

Techniques that would be capable of accurately determining the DNA sequence such as Sanger sequencing, generally prove ineffective when applied to repeat loci and/or samples to investigate the prevalence and consequences of interruptions. The expanded repeats associated with HD/HDL2 are also difficult to amplify and sequence accurately due primarily to the confounding effects of Taq polymerase slippage along the repeat tract (Shinde, 2003).

Although next-generation sequencing (NGS) approaches are now widely used in molecular laboratories, the technology is generally unreliable for analysing repeat loci. The reasons include the length of the repeat tract that must be sequenced in its entirety to yield an accurate repeat count, number of reads, sequence mapping procedures and genotype calling algorithms. The collaborating laboratory at the Institute of Molecular, Cell and Systems Biology, University of Glasgow has developed a specialised MiSeq sequencing protocol to address the challenges associated with sequencing repeat loci (Ciosi et al., 2018), allowing for accurate CAG/CCG repeat sizing and identification of variation within the repeat tract. This sequencing approach also has sufficient read depth for the quantification of *HTT* CAG repeat length somatic expansion (Ciosi et al., 2018).

1.7. Potential treatments

Despite continued efforts there is currently no effective treatment to delay, halt or reverse the progression of HD and HDL2. Symptomatic treatment of the movement disorder and psychiatric symptoms only serves to moderately improve the quality of life of the patient. The only approved drug currently used for hyperkinesia and dystonia is tetrabenazine (Marshall, 2006). In 1996, the first transgenic mouse models with CAG repeats between 115 and 150, and exhibiting many of the clinical features of HD, were created (Mangiarini et al., 1996). They

provided insights into the causative disease factors and enabled exploration of potential disease treatments (Walker, 2007).

Several approaches have been tested and a number of clinical trials are currently underway. The Huntingtin-lowering trials which use antisense oligonucleotide (ASO) technology include; Ionis Pharmaceuticals and Roche/Genentech INOIS-HTT_{RX} trial (Ionis Pharmaceuticals, 2017) and Wave Life Sciences PRECISION-HD2 trial (Wave Life Sciences, 2019). Recently, the Wave Life Sciences PRECISION-HD2 trial has been discontinued as there was no significant lowering of the mutant huntingtin protein (Wave Life Sciences 2021). Despite this setback, WAVE Life Sciences has announced another clinical trial based on a different ASO, with a next-generation PN backbone chemistry to improve the drug's performance in the hope of achieving the desired allele specific huntingtin lowering (Wave Life Sciences 2021). Other approaches include the use of adeno-associated virus (AAV) by UniQure AMT-130 (UniQure Biopharma B.V., 2019) and Voyager theraputics VY-HTT01 (Voyager Therapeutics Inc, 2020); while PTC therapeutics uses a small molecule compound to lower production of mutant huntingtin (PTC Therapeutics Inc, 2018). With continued testing of various technologies, the promise for an HD treatment in the near future is a reasonable possibility.

Currently, there are no similar potential treatments or clinical trials underway for HDL2; which is likely a result of the rarity of the disease and scarcity of research studies. However, due to the striking clinical similarity with HD, a shared disease mechanism may be identified which could be targeted with similar approaches to those explored in HD.

1.8. Justification for the study

As with the clinical trials, most of the research conducted to determine the molecular mechanism behind HD has been performed in European ancestry populations. This has resulted in a significant gap in knowledge around HD in populations not of European ancestry. Small case studies in African ancestry individuals have provided mainly anecdotal reports. Far less research has been conducted in HDL2 and no studies have focused on the sequence diversity of the CTG repeat tract in the *JPH3* gene. Further work is therefore required to determine the sequence diversity of the *HTT* and *JPH3* genes in individuals of African ancestry and to characterise and gain further insights of variation in these populations on HD and HDL2 pathogenesis.

By studying the sequence diversity in the repeat tracts and haplotype background, associations can be identified between the variants and disease phenotypes. In addition, these studies will contribute novel data on African ancestry populations and inform current research efforts into the precise molecular aetiology underlying the HD phenotype, ultimately allowing for the

development of drugs that are specifically tailored to the unique genetic diversity. The majority of drug clinical trials are conducted on European ancestry patients and may thus prove ineffective when administered to individuals of African ancestry. Through the collaboration with the University of Glasgow and the unique access to a relatively large cohort of HD and HDL2 patients with African ancestry, this study allowed for the novel assessment of the Huntington disease loci, *HTT* and *JPH3*, in individuals of African ancestry.

In general, there is a greater amount of genetic diversity present in sub-Saharan African populations across the genome in comparison to European populations (Choudhury et al., 2017). It is therefore a reasonable expectation that there is more sequence diversity in the *HTT* gene of African individuals, that may not only provide insights for African HD patients but also aid in understanding the overall diversity associated with HD. This diversity may have a similar modifier effect on the HD phenotype as seen in European ancestry patients and could provide insights into disease modification in African ancestry patients. Due to the clinical similarities with HD, it is proposed that similar mechanisms may be implicated in HDL2 that will contribute to the currently very limited body of knowledge of this disease in an understudied population as there have been no such studies on the *JPH3* gene due to its limited distribution.

1.9. Aim and Objectives

The overall aim of the research was to investigate the CAG allele sequence diversity at the HD locus (*HTT*) and the CTG allele sequence diversity at the HDL2 locus (*JPH3*) in African ancestry individuals and to assess what effects the *HTT/JPH3* repeat tract sequence diversity, haplotype background and somatic expansion of the repeats have on disease modification.

Aim 1

Investigate the *HTT* repeat tract and haplotype diversity in African ancestry patients and unaffected controls.

Objectives

1. Library preparation and NGS sequencing of the *HTT* repeat tract in HD patients and unaffected controls.
2. Assessment of *HTT* allele structure sequence diversity through the identification of CAG/CCG repeat size and variants within the repeat tract to define *HTT* allele structures.
3. Genotyping of the tag-SNPs at the *HTT* locus in HD patients and controls.
4. Assessment of the haplotype background diversity through manual and statistical phasing of haplotypes.

Aim 2

Investigate the *JPH3* repeat tract and haplotype diversity in African ancestry patients and unaffected controls from various populations.

Objectives

1. Library preparation and NGS sequencing of the *JPH3* repeat tract in HDL2 patients and unaffected controls.
2. Assessment of the *JPH3* repeat tract sequence diversity through the identification of CTG repeat size and variants within the repeat tract to define the *JPH3* allele structures.
3. Genotyping of the tag-SNPs at the *JPH3* locus in HDL2 patients and controls.
4. Assessment of the haplotype background diversity through manual and statistical phasing of haplotypes.

Aim 3

Investigate potential modifiers of the Huntington disease phenotype (age of diagnosis) in patients with African ancestry.

Objectives

1. Assessment of the effect the individual components of the repeat tract and *HTT* disease alleles have on the HD phenotype.
2. Assessment of the effect the haplogroup and haplotype background have on the HD phenotype.
3. Quantification of the amount of somatic expansion of the CAG repeat.
4. Assessment of the effect the amount of somatic expansion has on the HD phenotype.

Aim 4

Investigate potential modifiers of Huntington disease-like 2 (age of onset) in patients with African ancestry.

Objectives

1. Assessment of the effect the *JPH3* repeat tract has on the HDL2 phenotype.
2. Assessment of the effect the haplogroup and haplotype background have on the HDL2 phenotype.
3. Quantification of the amount of somatic expansion of the CTG repeat.
4. Assessment of the effect the amount of somatic expansion has on the HDL2 phenotype.

Section 2

Chapter 2 describes the methodology undertaken to investigate the four aims of the research. A similar methodology was followed for the analysis of the *HTT* and *JPH3* loci and has therefore been combined into one chapter. The gene-specific components were however specified and details expanded on.

2. Chapter 2: Methodology

Ethical approval for this research was obtained on 17/06/2017 from the Wits Human Research Ethics Committee (Medical) (certificate number: M1704130, sub-study of M110443) and valid for five years from the date of approval as shown in Appendix 1.

2.1. Sample selection

The DNA samples from the South African, individuals of black African ancestry were sourced from the Division of Human Genetics, University of the Witwatersrand (Wits) and the National Health Laboratory Services (NHLS) in Johannesburg, South Africa. The samples for patients with HD or HDL2 were originally extracted from the blood of patients referred for molecular confirmation of a clinical diagnosis of either HD or HDL2 and were stored anonymously for research. The selection criteria for the study included confirmed HD or HDL2 mutation status and self-reported African ancestry. The unaffected control individuals were derived from a set of permanently de-identified archived black African individuals stored at the Division of Human Genetics.

2.1.1. HD patients and unaffected controls

The samples included in the *HTT* analysis comprised of 68 unrelated individuals affected with HD and 158 unrelated individuals unaffected with HD of black African ancestry (Table 2.1). The patient samples included 46 females and 29 males, with an age of diagnosis range between 12 and 77 years. Demographic information related to the individuals affected with HD is shown in Appendix 2.1.

Table 2.1. The African ancestry individuals affected and unaffected with HD.

Sample type	Number
1. HD affected	68
2. Unaffected control	158
Total	**226**

2.1.2. HDL2 patients and unaffected controls

The samples included in the *JPH3* analysis comprised of 32 individuals affected and 553 individuals unaffected with HDL2 (Table 2.2, page 31). The individuals affected with HDL2 were of black African ancestry while the unaffected controls were from various worldwide populations. The patient samples included 19 females and 14 males, with an age of diagnosis range between 38 and 74 years. Demographic information related to the individuals affected with HDL2 is shown in Appendix 2.2. The unaffected controls comprised 187 samples from

black African individuals sourced from the Division of Human Genetics, 281 samples from five different populations in the 1,000 Genomes Project: Luhya in Webuye, Kenya (LWK), Gambian in Western Division, The Gambia (GWD), Chinese Dai in Xishuangbanna, China (CDX), Japanese in Tokyo, Japan (JPT) and Bengali from Bangladesh (BEB) and 85 Scottish individuals from the Generation Scotland study.

Table 2.2. The individuals affected and unaffected with HDL2 from various worldwide populations.

Sample type: Country	Number
1. African HDL2 affected: South Africa	32
2. African unaffected: South Africa	187
3. African unaffected: Kenya	75
4. African unaffected: The Gambia	54
5. East Asian unaffected: China	40
6. East Asian unaffected: Japan	39
7. South Asian unaffected: Bangladesh	73
8. European unaffected: Scotland	85
Total	**585**

Unlike the *HTT* analysis, additional unaffected control samples were included in the *JPH3* analysis as the locus has not been studied in detail and no repeat tract sequence data has been previously generated across any populations. The 1,000 Genomes Project samples (two African, two East Asian and one South Asian population) and the Generation Scotland samples were available at the Glasgow laboratory.

2.2. Sample preparation

The DNA concentrations for the black African ancestry samples were measured using a Nanodrop spectrophotometer and diluted to 100 ng/µl. Depending on the amount of sample available, between 50 and 100 µl were transferred to 96-well plates which were heat-sealed and transported to the Institute of Molecular, Cell and Systems Biology, the University of Glasgow, United Kingdom. Prior to the first research visit in 2017 an export permit from the South African Department of Health was obtained, and a Material Transfer Agreement was signed between the two universities.

2.3. Library preparation and sequencing

The library preparation and NGS sequencing of the *HTT* and *JPH3* repeat tracts were conducted at the collaborating laboratory at the Institute of Molecular, Cell and Systems Biology at the University of Glasgow over two research visits in 2017 and 2018. The procedural steps carried out in the library preparation and sequencing of both genes included: PCR amplification of the repeat tract, clean-up and concentration of PCR products, assessment of the library quality and quantity, and sequencing using MiSeq sequencing technology.

2.3.1. *HTT* repeat tract PCR optimisation

The *HTT* repeat tract library preparation and sequencing protocol was established by the Glasgow laboratory (Ciosi et al., 2018), that followed the procedural steps mentioned in 2.3. No additional optimisation steps were necessary for the *HTT* gene although prior to the PCR amplification using the MiSeq-compatible PCR primers a small subset of the South African samples underwent PCR amplification with normal *HTT* primers and the amplicons were visualised on a NuSieve agarose gel to ensure amplification of the correct product size occurred and the samples were not contaminated during transportation.

2.3.2. *JPH3* repeat tract PCR optimisation

Unlike the *HTT* gene, a novel *JPH3* library preparation and sequencing protocol had to be developed and optimised. It was based on the methodology used for the *HTT* gene. The *JPH3* repeat tract primer pairs were designed by evaluating the regions upstream and downstream of the CTG repeat tract in the *JPH3* reference sequence (NG_009797.1) using various bioinformatics tools; Primer3, OligoAnalyzer, SNPCheck and Basic Local Alignment Search Tool (Primer-BLAST). These allowed for the most optimal primer pairs to be selected by checking for the presence of SNPs in the primer binding regions, assessing the primer pair properties and selecting optimal PCR conditions.

Optimisation experiments were conducted on control DNA to determine the ideal PCR conditions and reagent concentrations. A temperature gradient PCR (55°C to 65°C) based on the melting temperature (Tm) of the primers and the predictions from the bioinformatics tools were performed. Once amplified, the products were visualised on a 1.5% agarose gel. The optimal annealing temperature (58.5°C) was determined based on the presence of a distinct band of the correct product size and the absence of spurious bands. After the PCR optimisation of the *JPH3* repeat tract, the MiSeq-compatible primers were designed to encompass the *JPH3* repeat tract (locus-specific primers) and additional regions specific for sequencing (detailed in Section

2.3.3). The MiSeq-compatible PCR primers allowed for the direct sequencing of the PCR product. All primers were supplied by Sigma Aldrich in the United Kingdom.

2.3.3. PCR amplification of the *HTT* and *JPH3* repeat tracts

Amplification of the *HTT* and *JPH3* repeat tracts was carried out using MiSeq-compatible PCR primers flanking the respective repeat tracts using 10 ng of the sample DNA. The reactions were prepared in 96-well plates containing the DNA samples and two controls: a no-template control (NTC) and a positive control. The positive control used for the *HTT* sequencing was from an HD mouse model of known repeat size, while the control used in the *JPH3* sequencing was an HDL2 patient sample with a repeat size confirmed in the optimisation process. The optimised PCR protocols for both the *HTT* and *JPH3* repeat tracts are shown in Appendix 3. The reaction components were as follows: Taq polymerase, 10X custom buffer (Ciosi et al., 2018), the respective MiSeq-compatible primer pairs and double-distilled water (ddH_20) to make up 15 µl reactions. No enhancers or additives were required for successful amplification of the *JPH3* repeat tract. This contrasted with the *HTT* repeat tract amplification where 10% dimethyl sulfoxide (DMSO) was added; DMSO increases the efficiency of the PCR of GC rich regions by disrupting the formation of secondary structures (Jensen, Fukushima and Davis, 2010). The cycling temperatures used for the denaturation, annealing and extension steps were the same for both genes.

The MiSeq-compatible PCR primers included an adaptor, a Nextera XT Index Kit v2 barcode, a sequencing primer binding site and a locus-specific primer (Figure 2.1, page 34). The adaptors enable binding of the amplified products to the oligonucleotides on the flow cell of the Illumina MiSeq platform, which allows cluster generation (Ciosi et al., 2018). The Nextera XT Index Kit v2 barcodes are necessary for the recovery of individual sequencing reads associated with a particular sample from the larger sequence pool (Ciosi et al., 2018). The sequencing primer binding site and the locus-specific primers allow for the addition of fluorescently labelled nucleotides to precisely sequence the amplified products. The exact nucleotide composition of the respective MiSeq compatible primers for each gene are shown in Appendix 4.1 and 4.2. A total of 16 forward primers and 24 reverse primers were used; this allowed for the unique identification and sequencing of 384 individual samples (4 x 96-well plates) in each library preparation.

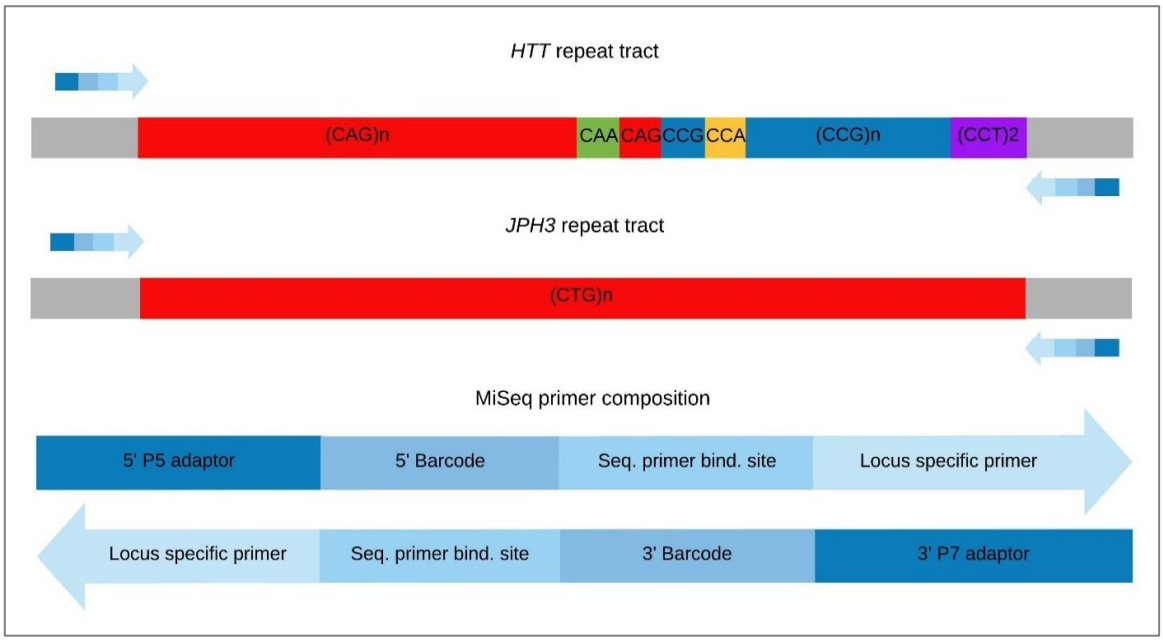


Figure 2.1. The *HTT* and *JPH3* repeat tracts showing the composition of the MiSeq-compatible primers used for library preparation and sequencing.

2.3.4. Cleaning and assessment of sequencing libraries

After amplification, the PCR products for each sample were pooled and distributed into LoBind Eppendorf tubes for purification using AMPure XP beads (Beckman Coulter, 2016). AMPure XP is the gold standard for bead-based clean-up of NGS libraries due to the efficient removal of contaminants, high recovery rate of amplified products greater than 100 bp, and predictable and consistent size selection (Beckman Coulter, 2016). The AMPure XP protocol uses a solid phase reversible immobilisation (SPRI) technology in which paramagnetic beads selectively bind to PCR products by type and size and allow for the removal of contaminants such as dNTPs, salts, primers and primer dimers introduced in PCR amplification (Beckman Coulter, 2016).

The AMPure protocol involved the introduction of the paramagnetic beads to the pooled library that allowed for nucleic acid immobilisation to the beads, leaving only the contaminants in solution. A magnetic field (Magnetic stand) was then introduced which pulled the magnetic beads and bound nucleic acids out of solution, allowing contaminants in solution to be removed. The bound nucleic acids then underwent an ethanol wash and drying step followed by elution of the purified nucleic acids in nuclease-free water (Beckman Coulter, 2016). The pooled library was subjected to two rounds of purification. The standard AMPure protocol was adapted (Ciosi et al., 2018) to use a ratio of 0.6:1 (beads:solution) in the first clean-up step, for the removal of primer-dimers and concentration of the sequencing library. A ratio of 1.4:1 was then used in the second clean-up step to purify the library.

The library was measured on the Qubit fluorometer using the dsDNA HS assay kit, according to the manufacturer's specifications (Life Technologies, 2015). The Qubit™ assays work through the emission of fluorescence from dyes that bind to specific targets (DNA, RNA or protein). The Qubit™ fluorometer allows the precise quantification of double-stranded DNA (ds-DNA) in the purified library over RNA and single-stranded DNA (ssDNA) (Life Technologies, 2015).

The quality and quantity of the purified library was then assessed on an Agilent 2100 Bioanalyzer using the Agilent High Sensitivity DNA Kit (Agilent Technologies, 2015). This ensured that the purified library was of the expected size, free of primer dimers and allowed for the estimation of the molarity. The Agilent 2100 Bioanalyzer functions through an on-chip electrophoresis and fluorescence-based quantification of DNA or RNA; nucleic acids intercalated with fluorescent dye molecules are electrophoretically separated and a fluorescent signal detected (Agilent Technologies, 2015). The fluorescent signal was then visualised as a gel image and an electropherogram that provided an average fragment size used to determine the molarity. An optimal purified library should have a volume of at least 30 µl and a molarity of 10 nM (Ciosi et al., 2018).

2.3.5. Quantification of library and MiSeq sequencing

The prepared libraries were then transferred to Glasgow Polyomics, a sequencing facility operated by the University of Glasgow. Although the Bioanalyzer provided a qualitative and quantitative assessment of the purified library, a final quantification of the libraries was required prior to sequencing on the MiSeq platform. Glasgow Polyomics performed quantitative PCR (qPCR) using the KAPA Library Quantification kit (Kapa Biosystems, 2016). This step accurately determines the library concentration as qPCR only analyses ds-DNA with adaptors (Illumina, 2016; Kapa Biosystems, 2016). The concentration of the library was then corrected for the fragment size estimated by the Bioanalyzer prior to sequencing.

Sequencing was conducted on an Illumina MiSeq sequencer using the MiSeq Reagent Kit v3 (600-cycles) with a cluster density of 1,000k cluster/mm^2 and a 5% PhiX spike-in (Ciosi et al., 2018). The PhiX spike-in was recommended for low diversity libraries such as PCR amplicons to increase the overall sequence diversity and act as a sequencing control (Illumina, 2016). The MiSeq uses a sequencing-by-synthesis (SBS) technology, where fluorescently labelled nucleotides are incorporated one base at a time to synthesise the DNA template in clusters on the flow cell (Ravi, Walton and Khosroheidari, 2018). Once the library is loaded onto the flow cell, cluster generation is carried out. This involves the hybridisation of the fragment strand to

bound oligonucleotides on the flow cell (Illumina, 2017), as the oligonucleotides were complementary to the MiSeq-compatible primers adaptor sequences (5' P5 and 3' P7 adaptor) on the fragment strand. Each bound fragment strand was then clonally amplified through bridge amplification and repeated to create millions of clonal clusters (Illumina, 2017).

Sequencing was then carried out through the extension of the sequencing primers by the introduction of fluorescently labelled nucleotides to create a read complementary to the fragment strand (Illumina, 2017). After the addition of each nucleotide, the flow cell is excited by a light source, which results in the emission of a characteristic fluorescent signal that is read at each cluster (Ravi, Walton and Khosroheidari, 2018). The emission wavelength and signal intensity are then used to produce a base call. The MiSeq reporter software was used to de-multiplex the reads corresponding to the barcoded primer pairs, which allowed for the separation of pooled sequencing libraries and the generation of two FASTQ files for each barcoded sample (R1: forward read; R2: reverse read) (Ciosi et al., 2018).

2.4. Sequencing analysis of repeat tracts

The *HTT* repeat tract analysis pipeline was previously established by the Glasgow laboratory (Ciosi et al., 2018). The *JPH3* data analysis pipeline was developed and optimized by the researcher following similar procedural steps; and were both were run on the University of Glasgow Galaxy server (Afgan et al., 2018). Galaxy is a web-based platform that allows for data analysis pipelines to be built from tools preloaded on the platform and data to then be run through the pipelines for bioinformatics analysis (Afgan et al., 2018). The schematic of the analysis pipeline followed for both genes is shown in Figure 2.2 (page 39).

2.4.1. Data analysis pipeline

The first procedural step followed in the analysis pipeline was the demultiplexing of the FASTQ files using the Cutadapt tool with a maximum error rate of 0.1 (Martin, 2011). This ensured that only the *HTT/JPH3* repeat tract reads would be analysed as multiple gene libraries were pooled together for sequencing. The adapter sequences were then trimmed from the raw reads, also using the Cutadapt tool with a maximum error rate of 0.39. The maximum error rate parameter specified the error tolerance when searching for the region of interest. The optimal maximum error rate was determined by comparing different rates and then assessing the quality of the reads after trimming. The Cutadapt tool worked by searching for the region of interest in all reads and once found the region was removed from the sequence (Martin, 2011).

The trimmed reads were then aligned to gene-specific synthetic reference sequences using the Burrows-Wheeler Alignment tool (BWA-MEM) with a gap extension of 4.4 (Li, 2013). The

BWA-MEM tool allows for the fast and accurate mapping of long-read sequences from 70 bp to 1 Mbp against a *HTT*/*JPH3* reference sequence (Li, 2013). These synthetic reference sequences were generated using the RefGeneratr tool developed by the Glasgow laboratory (Maxwell, 2016). For the *HTT* pipeline, RefGeneratr produced a total of 4,000 synthetic reference sequences which had a combination of 1 to 200 CAG repeats and 1 to 20 CCG repeats. For the *JPH3* pipeline, RefGeneratr was used to produce a total of 200 synthetic reference sequences with from 1 to 200 CTG repeats. The BWA-MEM tool outputted the aligned reads in a Binary Alignment Map (BAM) format which is not a human-readable format.

To produce a human-readable format, the aligned reads were converted from the BAM format into a Sequence Alignment Map (SAM) format using SAMtools (Li et al., 2009). The aligned reads in the SAM format were then filtered using the text manipulation in Galaxy to remove reads that were aligned to more than one reference by selecting reads with a map quality (MAPQ) of >0; a MAPQ of 0 indicates that the read maps to multiple references.

After each step in the pipeline, a quality check was conducted using the FastQC tool (Andrews, 2010). The FastQC tool outputted an HTML file which gave the read quality reports with various analysis modules. The modules included: basic statistics, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and kmer content. The results allowed for the assessment of the quality of the sequencing reads. If the read quality was not in an acceptable range the module would output a warning or QC failure (Andrews, 2010).

2.4.2. Repeat sizing and variant identification

The filtered reads were then analysed using an alignment visualiser, Tablet (Milne et al., 2013). Tablet enabled repeat sizing and the identification of variants along the *HTT/JPH3* repeat tracts, which allowed for the assignment of (typical and atypical) alleles in both affected and unaffected individuals. Typical alleles were characteristic of the reference gene with no variation along the repeat tract (Figure 2.1, page 34), whereas atypical alleles were defined by the presence of variations which differed from the reference sequence. In Tablet, the atypical alleles were identified by a mismatch from the aligned reference. To confirm atypical structures, a tailored reference was designed that was specific to the variant identified, using RefGeneratr, which allowed for the reads to align to the atypical reference without a mismatch.

For the *HTT* gene, the results obtained after the sequencing analysis were compared with the results of another bioinformatics pipeline developed by the Glasgow laboratory, ScaleHD (Maxwell, 2018). ScaleHD allowed for the automated genotyping of the *HTT* repeat tract which was the main difference to the Galaxy pipeline. This was achieved through the use of various steps which included; determination of CCG heterozygosity, detection of atypical structure, atypical re-alignment and distribution smoothing (Maxwell, 2018). For the *JPH3* gene, no comparative sequencing analysis was carried out as no equivalent bioinformatics pipeline had been previously developed.

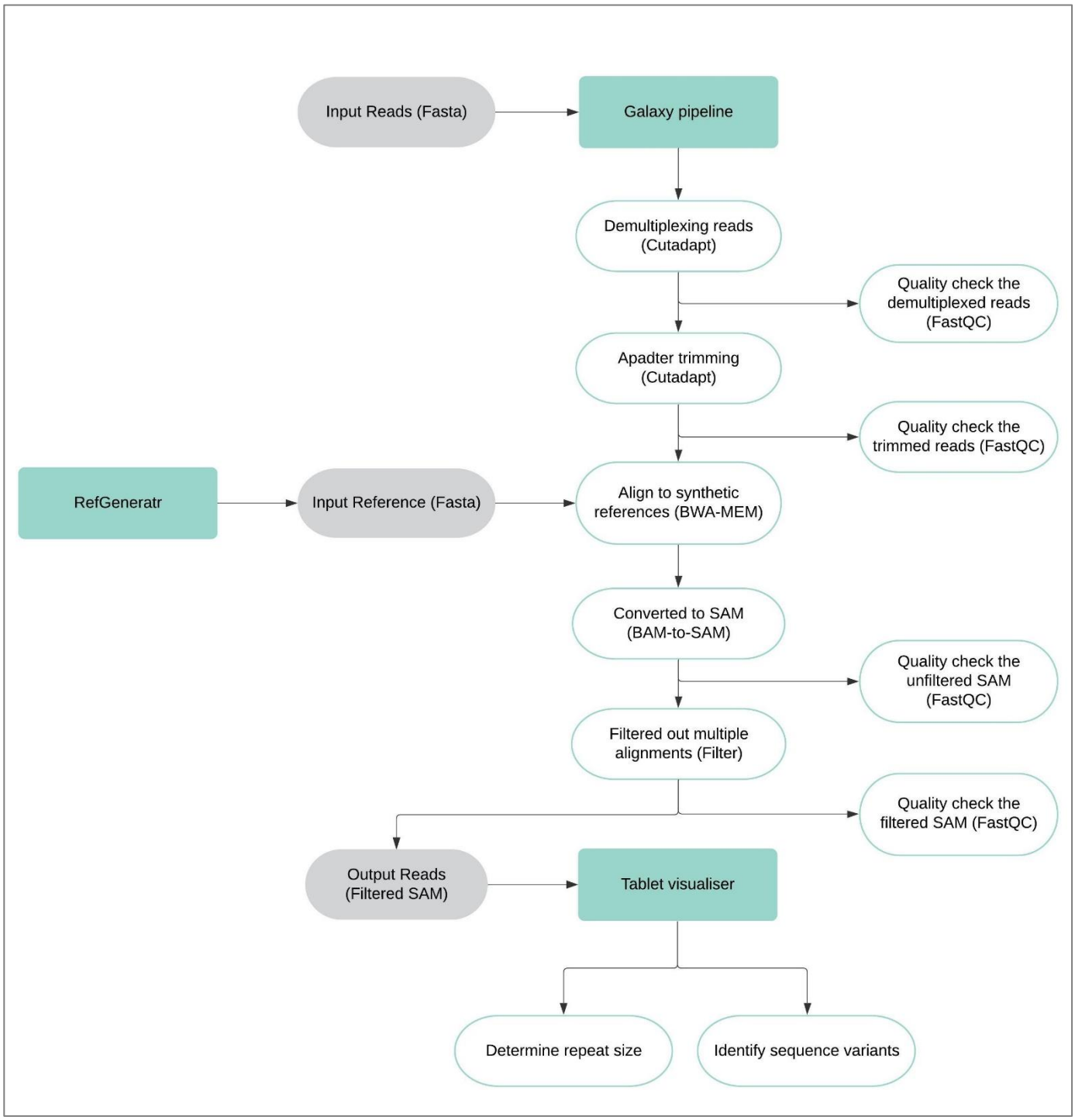


Figure 2.2. The procedural steps followed in the bioinformatics pipelines used for the sequencing analysis of the *HTT* and *JPH3* repeat tracts.

2.5. Haplotype analysis

The repeat tracts were further studied in the context of the haplotype background described for the *HTT* and *JPH3* loci. The haplotypes were constructed by genotyping the selected tag-single nucleotide polymorphisms (tag-SNPs) in the respective genes and then phasing the tag-SNPs genotypes to generate haplotypes.

2.5.1. Selection of *HTT* and *JPH3* tag-SNPs

For the *HTT* haplotype construction, a subset of 13 tag-SNPs from a previous study were selected (Baine et al., 2013). The 13 tag-SNPs were used to define the broader haplogroups A, B and C and the novel haplogroup variant C-South Africa. For the *JPH3* haplotype construction, three tag-SNPs were selected from a previous study (Krause et al., 2015); and a further four tag-SNPs selected from the gene region. The additional tag-SNPs were selected based on their distance from the *JPH3* repeat tract, population genetics/allele frequencies, heterozygosity and linkage disequilibrium. The 20 tag-SNPs used to construct the *HTT* and *JPH3* haplotypes are shown in Appendix 5.

2.5.2. Genotyping of tag-SNPs

The 20 tag-SNPs were genotyped using a MassARRAY® System from Agena Bioscience™, at Inqaba Biotec. The DNA samples were diluted to 100 ng/µl and transferred to 96-well plates, which were then heat-sealed for transportation to the genotyping facility. The MassARRAY® System technology combines a multiplexed primer extension chemistry with a highly sensitive matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry (Agena Bioscience, 2015).

The procedural workflow for genotyping was as follows: PCR amplification, iPLEX Extension Reaction and MALDI-TOF Mass Spectrometry Analysis (Agena Bioscience, 2015). The PCR used multiplex end-point PCR amplification of the target regions encompassing the tag-SNPs. Excess nucleotides produced in the reaction were then dephosphorylated by shrimp alkaline phosphatase (SAP). The PCR products then underwent the iPLEX Extension Reaction which involved a single base extension of mass-modified dideoxynucleoside terminators annealing upstream of the tag-SNPs. The extension products were then desalted and transferred to a SpectroChip where crystallisation occurred. The SpectroChip was placed in the MassARRAY® MALDI-TOF analyzer which was irradiated by a laser. This resulted in ionisation and the migration of the DNA ions through the vacuum (Agena Bioscience, 2016). The time of arrival at the detector was determined by the mass, which was nucleotide specific, and thus allowed for

the identification of the tag-SNP alleles by the MassARRAY® MALDI-TOF analyzer Typer software (Agena Bioscience, 2016).

In the *JPH3* analysis, the tag-SNPs for the individuals from various populations studied in the 1,000 Genomes Project were not genotyped as part of this study. This information was previously generated as part of the 1,000 Genomes Project thus for these individuals the tag-SNPs genotypes were accessed from the publicly available online platform Ensembl (www.ensembl.org/index.html).

2.5.3. Haplotype phasing

After genotyping, a two-step haplotype phasing analysis was used involving both manual and automated statistical phasing. Manual phasing was achieved through the construction of haplotypes based on homozygosity of genotyped tag-SNPs and associated repeat sizes shown in previous studies (Baine et al., 2013; Krause et al., 2015). In individuals where haplotypes were equivocal, haplotypes were inferred based on previously defined haplotypes.

The automated statistical phasing was achieved using PHASE (*v*2.1.1) software, which applies a Bayesian inference model to construct haplotypes (Stephens and Donnelly, 2003; Stephens, Smith and Donnelly, 2001). Before analysis, the genotype data was inputted into a specific format for the PHASE software which was run in an Ubuntu command line using an Oracle VirtualBox Manager. Once run, various output files were generated which included the haplotype frequencies, most likely pairs, recombination parameters, significance and the goodness of fit of an estimated haplotype. These outputs were interpreted together to assign haplotypes based on the most likely pairs of haplotypes and probabilities of the best haplotype estimates (Stephens and Donnelly, 2003; Stephens, Smith and Donnelly, 2001). The final haplotypes were then confirmed based on comparison and agreement between the manual and statistical phasing approaches for each sample.

2.6. Somatic expansion analysis

The method for the analysis of somatic expansion of the CAG repeats in the *HTT* gene was established by the Glasgow laboratory (Ciosi et al., 2018). The method was then also adapted to assess the CTG repeats in the *JPH3* gene. Somatic expansion of the disease allele was measured by counting the ratio of reads larger than the modal repeat (N+1 to 10); with the modal repeat (N) having the highest read count (*i.e.* progenitor repeat). The relative number of repeats larger than the modal repeat (*i.e*. somatic expansion) in the DNA sample was then quantified by counting the number of reads that aligned to synthetic reference sequences having an increased number of repeats (N+1 to 10).

$$Ratio\ of\ somatic\ expansion = \frac{(number\ of\ somatic\ expansion\ products)}{(number\ of progenitor\ allele\ products)}$$

The graphical representation of the *HTT* gene read count distribution of an HD patient is shown in Figure 2.3 (page 43). In this example, the HD patient was heterozygous at the *HTT* locus, with an unexpanded allele of 15 CAG and 10 CCG repeats and an expanded allele of 48 CAG and 7 CCG repeats. For the expanded allele, the reads to the left of the progenitor/modal repeat were considered PCR artefacts. Extensive analyses of small pool PCR by the Glasgow laboratory have determined anything shorter than the modal repeat to be an artefact resulting from slippage during amplification (Ciosi et al., 2019). The reads to the right of the modal repeat were considered somatic expansion due to repeat instability. For the unexpanded allele, there was no somatic expansion of the unexpanded repeats and the amount of PCR slippage was minimal.

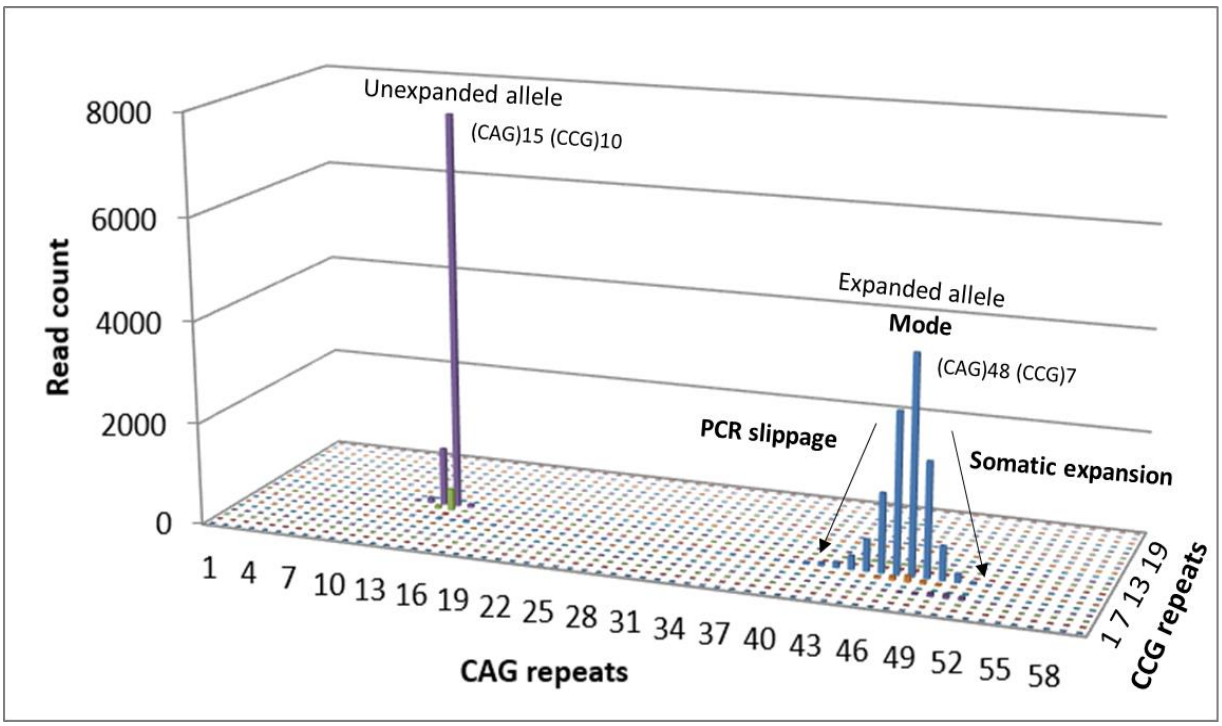


Figure 2.3. The sequencing read count distribution for the CAG/CCG repeat of an individual affected with HD.

Showing the unexpanded and expanded alleles with the respective CAG and CCG repeat sizes for both alleles, the expanded allele modal repeat with the largest read count (N), PCR slippage (N-1) and somatic expansion (N+1). Peaks identified in addition to the modal unexpanded and expanded alleles are a result of CAG/CCG PCR slippage and somatic expansion.

Once the ratio of somatic expansion was calculated, an expansion score was then determined from the residuals of a linear model of the ratio of somatic expansion. In the linear model the explanatory variables included the repeat size, the age at sampling and the interaction between the repeats and the age at sampling (Ciosi et al., 2019). In the HD and HDL2 analysis the linear models for the ratio of somatic expansion were log-transformed to allow for a better fit.

2.7. Analysis of potential modifiers of disease

Potential genetic modifiers of disease were investigated through their effect on the phenotype. When studying the HD phenotype, the age of onset (AoO) of motor symptoms has been used as a well-characterised measure of disease severity (Lee et al., 2012). However, less than 50% of our HD patients had AoO information available although age of diagnosis (AoD) information was available. Unlike the HD patients, approximately 70% of the HDL2 patients had AoO information and all of the patients had AoD information available. Due to the limitation of the available AoO information it was necessary to establish the most appropriate variable for the disease phenotype prior to the assessment of the potential modifiers. This was achieved through the examination of the AoD as an acceptable proxy for AoO.

The potential modifiers were statistically evaluated in R (*v*3.4.3) and Rstudio using an R-script specific for the HD and HDL2 analysis, respectively. The R-scripts were made up of commands to test the effect on the disease phenotype and create graphical representations of the assessment. The statistical analyses included linear regression analyses and multiple linear models in which the lm function was used to determine associations between the disease phenotype and various explanatory variables. The explanatory variables assessed included the components of the repeat tract, haplogroup/haplotype background and the ratio of somatic expansion. Once the estimates were outputted for each of the explanatory variables the effect in years was calculated. This was carried out by converting the estimate into an exponential, then determining the percentage change by multiplying by 100 and lastly determining the effect in years by multiplying by the mean of what was assessed (*i.e*. Disease phenotype: AoD for HD and AoO for HDL2).

2.7.1. Repeat tract analysis

The first potential modifier that was assessed for the effect on the disease phenotype was the repeat tract. The linear regression analysis and linear models that assessed the effect of the repeat tract was only conducted on a subset of HD and HDL2 patients carrying expanded repeat sizes between 39 and 52 repeats. Repeats greater than 53 have been previously shown to violate linear model assumptions in previously HD analyses (Lee et al., 2012). The disease phenotype used in the HD analysis was the AoD while in the HDL2 analysis it was the AoO.

In the HD analysis, a linear regression analysis was first constructed for the effect of only the CAG repeats on the AoD. This was followed by the two linear models that assessed the effect of the *HTT* repeat tract. The first linear model assessed the effect of the individual components of the *HTT* repeat tract (CAG, CAACAG, CCGCCA, CCG, CCT) and then the disease allele structures (all components of the *HTT* repeat tract). The reference used in the HD analysis was

the most common disease allele structure identified in HD patients of European ancestry. The estimated marginal mean or least square mean AoD in years for each of the allele structures was then established using the emmeans function in R (*v*3.4.3) (Russell, 2020). In this analysis, the AoD results were corrected for CAG repeat size to limit the bias across patients with larger CAG repeats. In the HDL2 assessment, a linear regression analysis was first constructed for the effect of only the CTG repeats on the AoO. This was followed by two linear models that assessed the effect of the *JPH3* repeat tract. Unlike the *HTT* analysis, all the HDL2 patients had the same allele structure and had no other sequence components. Thus, only the CTG repeats were assessed in the first linear model while the second model then also considered the effect of the non-disease associated CTG repeats and the sex of the HDL2 patients.

2.7.2. Haplogroup/haplotype analysis

The second potential modifier that was assessed for the effect on the disease phenotype was the haplogroup/haplotype background. Like the repeat tract analysis, the haplogroup/haplotypes background analysis was only conducted on HD/HDL2 patients carrying expanded repeat sizes between 39 and 52 repeats.

In the HD analysis, two linear models were constructed to assess the effect of the haplogroup background and then the haplotype background. The reference used in each of the linear models was the most common disease haplogroup (B) and haplotype (B2) identified in the African ancestry HD patients. The estimated marginal mean AoD in years for each of the haplotypes was then established. In the HD analysis, a goodness of fit test of the R-squared in 5,000 bootstrapped samples was also conducted, to determine if the variation in the HD phenotype was better explained by allele structure or the haplotype background.

In the HDL2 analysis, only one linear model was constructed to assess the effect of the background larger haplotype. Unlike the *HTT* analysis, only the larger haplotypes were assessed as all the HDL2 patients had the same haplogroup. The reference used in the linear models was the most common disease larger haplotype (larger haplotype 6) identified in the African ancestry HDL2 patients. The estimated marginal mean AoO in years for each of the larger haplotypes was then established.

2.7.3. Somatic expansion analysis

The third potential modifier that was assessed for the effect on the disease phenotype was the ratio of somatic expansion of the repeats. The somatic expansion analysis was only conducted on the HD/HDL2 patients carrying expanded alleles with less than 50 repeats. Repeats larger than 50 were excluded due to the difficulty in accurately identifying the modal repeat from a

somatically expanded repeat in large expansions. Once the ratio of somatic expansion was assessed, the effect it had on disease phenotype was considered through the calculation of an expansion score as outlined in 2.6. Somatic expansion analysis (page 43).

In the HD analysis, a linear model was constructed to assess the effect of the expansion score. In the assessment of the AoD the explanatory variables included the CAG repeats, the disease allele structures and the expansion score. The reference used in the HD analysis was the most common disease allele structure identified in the European ancestry HD patients. The estimated marginal mean expansion score for each of the disease allele structures was then established. In the HDL2 analysis, the linear model that assessed the effect of the AoO included the CTG repeats, the larger haplotypes and the expansion score as explanatory variables. The estimated marginal mean expansion score for each of the larger haplotypes was then also established.

Section 3

Section 3 combines the result and discussion chapters for the four research aims. Chapter 3 presents the results (3.1) and discussion (3.2) sections related to the investigation of the *HTT* repeat tract sequence diversity, haplotype diversity and potential modifiers of the HD phenotype. Chapter 4 comprises the results (4.1) and discussion (4.2) sections related to the investigation of the *JPH3* repeat tract sequence diversity, haplotype diversity and potential modifiers of HDL2 phenotype.

3. Chapter 3: *HTT* analysis

This chapter reports on the *HTT* repeat tract sequence diversity, haplogroup/haplotype background diversity and potential modifiers of the HD phenotype in African ancestry individuals affected with HD. The *HTT* repeat tract sequence diversity and haplogroup/haplotype background diversity was established in individuals affected and unaffected with HD. This was achieved through sequencing the *HTT* repeat tract to define allele structures, followed by genotyping and phasing of the *HTT* tag-SNPs to construct haplotypes. Potential modifiers of the HD phenotype were investigated in the HD patients using statistical methods to assess the effects of the *HTT* repeat tract sequence, haplogroup/haplotype background and ratio of somatic expansion of the CAG repeats.

3.1. Results of *HTT* analysis

3.1.1. *HTT* repeat tract sequence diversity

Across the African ancestry individuals, a large amount of sequence diversity was identified with 17 allele structures defined; three of which were novel.

The exact sequence composition of the *HTT* repeat tract has not been previously studied in African ancestry individuals. Research in European ancestry individuals has revealed significant sequence diversity within the *HTT* repeat tract and assessed how factors modify the Huntington disease phenotype. In order to explore/characterise the sequence diversity in individuals of African ancestry, high-throughput NGS sequencing of the *HTT* repeat tract was performed on a total of 68 unrelated individuals affected with HD and 158 individuals unaffected with HD of black African ancestry (68 disease alleles and 384 non-disease alleles).

The high-throughput sequencing results revealed a large amount of sequence diversity with 17 different allele structures identified (Table 3.1, page 50). In order to describe the allele structures, the Glasgow laboratory developed a nomenclature which visually represented each component of the *HTT* repeat tract (Ciosi et al., 2019). This nomenclature was based on the assignment of amino acid codon abbreviations for each of the components and numbering of the components based on their position in the repeat tract. The *HTT* repeat tract is made up of a polyglutamine region (Q = Glutamine) composed of the CAG repeats in the first position (Q^1 = CAG) and the intervening CAACAG sequence in the second position (Q^2 = CAACAG). These are followed by the polyproline region (P = Proline) composed of the intervening CCGCCA sequence (P^1 = CCGCCA), the CCG repeats (P^2= CCG) and lastly the CCT repeats (P^3= CCT). The allele structure names were thus written by the order of appearance in the repeat tract, separated by hyphens (Q^1-Q^2-P^1-P^2-P^3). The allele structures were defined as either typical or atypical based on a comparison with the reference allele structure (LRG_763). The typical

allele structures were made up of a variable number of CAG and CCG repeats, indicated by the abbreviations Q^1 and P^2; while the other components of the repeat tract were not variable, thus the abbreviations Q^2, P^1 and P^3 were replaced by numbers that represented the sum of the triplet nucleotides in each component (*e.g.* Q^1-2-2-P^2-2). In this way, the composition of the *HTT* repeat tract was more easily illustrated and referred to.

Of the 17 allele structures identified, eight were typical, defined based on the variability in the number of CAG (Q^1) and CCG (P^2) repeats only. In the typical allele structures, the CAG repeat sizes were either in the non-disease range from 11 to 32 repeats or within the pathogenic range from 39 to 58 repeats, while the CCG repeats ranged from 6 to 13 repeats across both the non-disease and disease alleles. The remaining nine structures were defined as atypical based on variability across all the components in the *HTT* repeat tract, not just in the number of CAG (Q^1) and CCG (P^2) repeats. These variations either resulted in an apparent loss or duplication of the intervening sequences CAACAG (Q^2 = 0 or 4) and CCGCCA (P^1 = 0 or 4) and/or accompanied by an additional downstream CCT (P^3 = 3) repeat. These atypical allele structures thus did not follow the previously defined reference allele structure.

Three allele structures were unique to this study as they had not been previously identified in any other studies or populations (* asterisk in Table 3.1, page 50). One of the novel alleles was typical while the other two had an atypical allele structure. The typical structure Q^1-2-2-13-2, seen on one non-disease allele was novel due to a CCG repeat size of 13 not previously identified. The two atypical alleles Q^1-4-2-10-2 seen on four non-disease alleles and Q^1-4-2-4-3 seen on one disease allele, had not been previously identified. These allele structures were novel due to the duplication of the intervening sequence CAACAG (Q^2= 4), accompanied by a CCG repeat size of 10 and four, respectively. For the Q^1-4-2-4-3 allele structure, this duplication with four CCG repeats was accompanied by an additional downstream CCT (P^3= 3).

To ascertain the significance of the sequence diversity findings, a statistical comparison of the frequency of each structure on disease and non-disease alleles was undertaken. The Fisher exact test *p*-values are shown with the statistically significant frequencies in bold in Table 3.1. The results revealed that only one typical allele Q^1-2-2-10-2 was statistically more frequent in the non-disease alleles, while four atypical alleles differed significantly between the disease and non-disease alleles. Three structures Q^1-2-2-6-3, Q^1-2-0-9-2 and Q^1-0-0-9-2 were more frequent in disease alleles, while one structure Q^1-2-2-9-3 was more frequent in non-disease alleles.

Table 3.1. A summary of the African ancestry *HTT* disease and non-disease allele structures.

Allele types	Allele structure nomenclature	*HTT* repeat tract						Allele occurrence				Fisher exact test
		Q^1		Q^2	P^1	P^2	P^3	Non-disease N=384		Disease N=68		
		CAG		CAACAG	CCGCCA	CCG	CCT	n	%	n	%	*p*-values
Typical alleles	Q^1-2-2-6-2	14–17	-	2	2	6	2	11	2.9	–	–	$3.836x10^{-1}$
	Q^1-2-2-7-2	15–28	41–55	2	2	7	2	99	25.8	10	14.7	$6.386x10^{-2}$
	Q^1-2-2-8-2	17	–	2	2	8	2	5	1.3	–	–	1
	Q^1-2-2-9-2	15–28	40	2	2	9	2	29	7.6	1	1.5	$6.622x10^{-2}$
	Q^1-2-2-10-2	11–20	40–54	2	2	10	2	**71**	**18.5**	20	29.4	**$4.831x10^{-2}$**
	Q^1-2-2-11-2	12–21	–	2	2	11	2	18	4.7	–	–	$8.909x10^{-2}$
	Q^1-2-2-12-2	17	–	2	2	12	2	1	0.3	–	–	1
	*Q^1-2-2-13-2	17	–	2	2	13	2	1	0.3	–	–	1
Typical alleles subtotal								235	61.2%	31	45.6%	
Atypical alleles	Q^1-2-2-4-3	23	–	2	2	4	3	1	0.3	–	–	1
	Q^1-2-2-6-3	15–23	42–44	2	2	6	3	2	0.5	**4**	**5.9**	**$5.587x10^{-3}$**
	Q^1-2-2-9-3	12–21	–	2	2	9	3	**92**	**24.0**	–	–	**$9.142x10^{-8}$**
	Q^1-2-2-10-3	16	–	2	2	10	3	1	0.3	–	–	1
	*Q^1-4-2-4-3	–	42	4	2	4	3	–	–	1	1.5	$1.540x10^{-1}$
	Q^1-4-2-7-3	14–19	–	4	1	7	3	22	5.7	–	–	$5.878x10^{-2}$
	*Q^1-4-2-10-2	16–19	–	4	1	10	2	4	1.0	–	–	1
	Q^1-2-0-9-2	16–32	40–58	2	0	9	2	27	7.0	**30**	**44.1**	**$3.119x10^{-13}$**
	Q^1-0-0-9-2	–	39–46	0	0	9	2	–	–	**2**	**2.9**	**$2.235x10^{-2}$**
Atypical alleles subtotal								149	38.8%	37	54.4%	

The novel allele structures unique to this study are indicated by an asterisk (*). The most common non-disease allele structure is indicated in **bold green font** and the most common disease allele structure is indicated in **bold red font**. The statistically significant differences between the non-disease and disease alleles are shown in **bold font**.

The schematics of the allele structures are shown in Figure 3.1 (page 51) for the disease and non-disease alleles, respectively. Coloured bars along the *HTT* repeat tract were used to distinguish each of the components. In the schematics, the typical allele structures were grouped together as Q^1-2-2-P^2-2 with repeat size ranges indicated (*e.g.* (CAG)40-55). The atypical allele structures were shown individually for deviations from the reference allele structure to be clearly demonstrated. The first atypical allele structure shown in the figure was Q^1-2-0-9-2 and was atypical due to the loss of the intervening CCGCCA sequence. The nomenclature Q^1-2-0-9-2 was used due to the variable number of CAG repeats (Q^1), followed by the presence of the intervening CAACAG (Q^2 = 2) sequence, the loss of the intervening CCGCCA (P^1=0), the nine CCG (P^2 = 9) repeats, and lastly two downstream CCT (P^3 = 2) repeats.

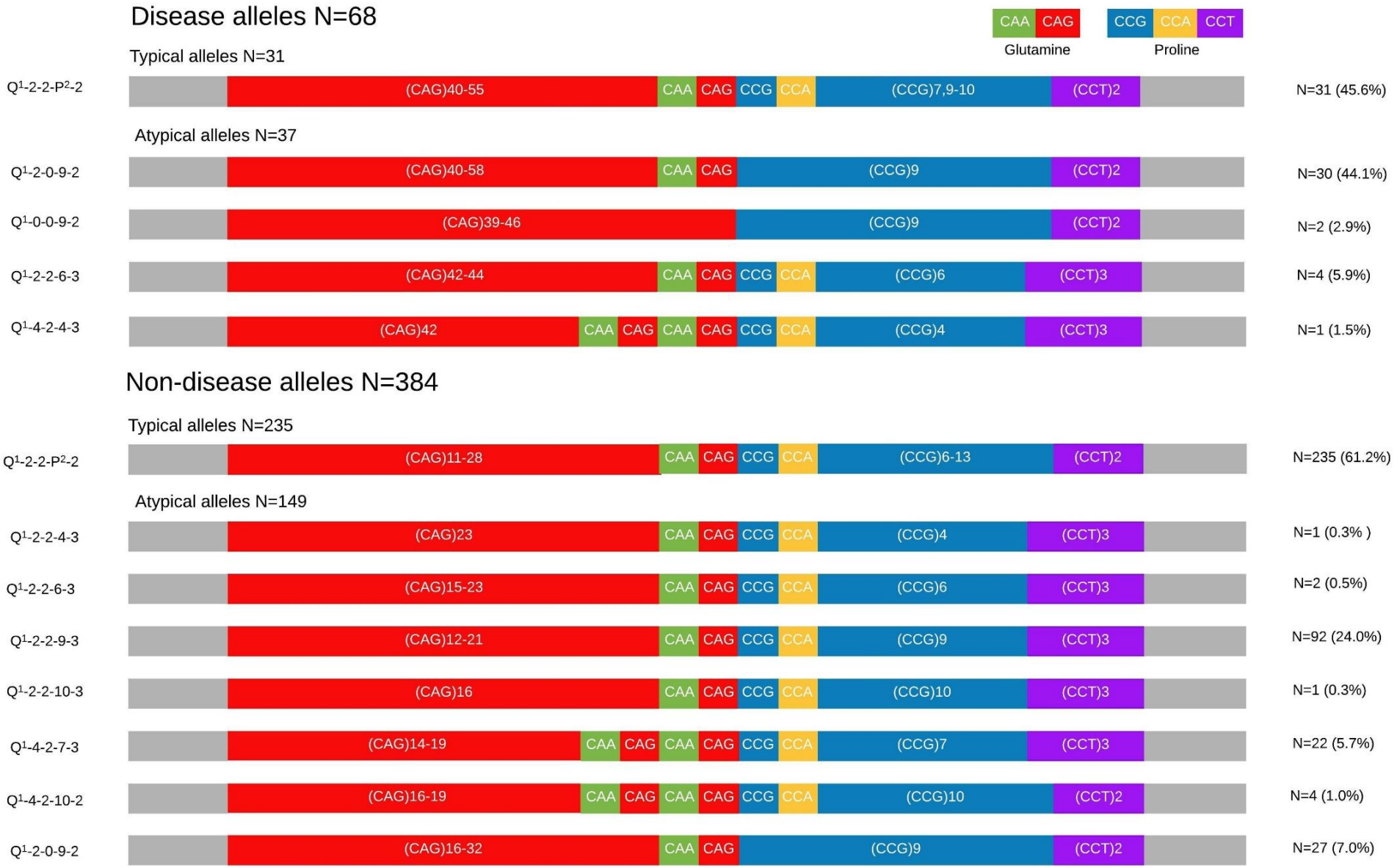


Figure 3.1. Schematic representation of the *HTT* disease and non-disease allele structures defined for this study.

The *HTT* repeat tract is composed of a region that codes for polyglutamine with a variable number of CAG repeats (Q^1) and a CAACAG (Q^2) sequence. This is followed by a region coding for polyproline which includes a CCGCCA (P^1) sequence, a variable number of CCG (P^2) repeats and downstream CCT repeats (P^3). Typical alleles were grouped and defined as such if the allele structures were characteristic of the *HTT* reference gene (locus reference genomic: LRG_763). The atypical alleles were defined based on deviation from the reference gene.

The most common African ancestry disease allele structure Q^1-2-0-9-2 was atypical and characterized by the loss of the CCGCCA sequence. This differed significantly in frequency from the most common African ancestry non-disease allele and the European ancestry alleles.

The comparison of the most common allele structures revealed differences between the non-disease and disease alleles. The most common non-disease allele structure Q^1-2-2-7-2 (99/384, 25.8%) was typical in structure, with variability occurring only in the length of the CAG repeat. Although this allele structure was present in disease alleles, it occurred in a smaller percentage of alleles (10/68, 14.7%). The most common disease allele structure Q^1-2-0-9-2 (30/68, 44.1%) was atypical in structure due to the loss of the CCGCCA sequence (P^1= 0). Although this allele structure was also present in the non-disease alleles, it occurred only in a small proportion of alleles (27/384, 7.0%).

The results were then compared with previously described European ancestry alleles (Ciosi et al., 2019). The European ancestry alleles were composed of 807 individuals carrying HD disease alleles that have been studied by the Glasgow laboratory (Ciosi et al., 2019). Among these, 92.7% of the non-disease and 97.3% of disease alleles had a typical allele structure; (Ciosi et al., 2019) compared to the African ancestry *HTT* alleles where only 61.2% of non-disease and 45.6% of disease alleles had a typical allele structure. The comparison revealed statistically significant allele frequency differences (typical African vs European non-disease: 235/384 vs 748/807 $p < 2.2 \times 10^{-16}$ and typical African vs European disease: 31/68 vs 785/807 $p < 2.2 \times 10^{-16}$). Despite the statistically significant frequency differences, the most common allele structure Q^1-2-2-7-2 in the European ancestry non-disease and disease alleles, was also the most common in African ancestry non-disease alleles.

However, in the African ancestry disease alleles the most common allele structure was the atypical Q^1-2-0-9-2 (30/68, 44.1%). Although this allele structure was present in the European ancestry individuals, it was found at a low frequency in non-disease alleles (32/807, 4%) and was extremely rare in disease alleles (1/807, 0.1%) (Ciosi et al., 2019). These findings were supported by a genome-wide study conducted in over 9,000 individuals affected with HD, which showed comparable results with low frequencies in the non-disease alleles (412/9,463, 4.4%) and disease alleles (6/973, 0.6%) (GeM-HD Consortium, 2019). The differences not only occurred between the individuals affected and individuals unaffected with HD from the same ancestry, but also among individuals affected with HD from different ancestries.

3.1.2. *HTT* haplogroup/haplotype diversity

Across African ancestry individuals, a large amount of haplotype diversity was identified with 16 haplotype structures defined with distribution differences between disease and non-disease alleles.

The *HTT* locus was further investigated in the context of haplogroup and haplotype background that occur along the gene spanning ~196,063 kb of chromosome 4. The haplotypes are made up of closely linked tag-SNPs that are inherited together, and haplogroups are groups of similar haplotypes. The haplogroup/haplotypes were constructed using 13 previously studied tag-SNPs (Baine et al., 2013) that defined haplogroups A, B and C, and the South African specific haplogroup variant C-South Africa (C-SA) and the haplotypes that occur within them. The haplogroup/haplotype background was constructed for a total of 224 samples from individuals of African ancestry through the tag-SNP genotyping and phasing of haplotypes in 67 individuals affected and 157 unaffected with HD.

The results revealed 16 different haplotypes across the African ancestry individuals, shown in Table 3.2 (page 54) alongside the corresponding allele structures. Haplotypes were identified across the four previously defined haplogroups (A, B, C and C-SA) as well as an "other" haplogroup. The addition of the "other" haplogroup was necessary when the composition of tag-SNP did not fall into any of the previously defined haplogroups/haplotypes.

For the non-disease alleles, the largest proportion occurred on haplogroup C (159/381, 41.7%), indicated in bold green font. Within haplogroup C, the most common haplotype was C5 (94/381, 24.8%). The allele structures that occurred on haplotype C5 were predominantly typical in structure however, the following atypical structures Q^1-2-2-10-3, Q^1-2-0-9-2 and Q^1-4-2-10-2 also occurred on the C5 haplotype. A proportion (39/381, 10.2%) of non-disease alleles could not be defined based on the previously identified haplotypes and were thus designated into the "other" haplogroup.

For the disease alleles, the largest proportion of alleles occurred on haplogroup B (33/67, 49.3%) as indicated by red bold in Table 3.2 (page 54). Within haplogroup B, the most common haplotype was B2 (32/67, 47.8%) and the associated allele structures were atypical; Q^1-0-0-9-2, Q^1-4-2-4-3 and Q^1-2-0-9-2. These allele structures were characterised by either a loss of the intervening sequence/s, duplication of the glutamine intervening sequence and/or the addition to the downstream CCT repeat. The most common allele structure in the disease alleles, Q^1-2-0-9-2 was found exclusively on haplotype B2, shown in bold red font. The most common allele structure Q^1-2-2-7-2 in the non-disease alleles was found on six different haplotypes; A2b, A4a,

A4b, A6, C9 and the "other" haplotype. Although the allele structure (Q^1-2-2-7-2) occurred on various haplotypes it was predominantly associated with haplogroup A (56/99, 56.6%) and only seven of the alleles (7/99, 7.1%) occurred on the South African specific haplogroup variant C-South Africa (C-SA).

Two European 'high-risk' haplogroup A variants, A2a and A2b were identified in this study. These 'high-risk' haplogroup A variants were previously thought to be absent in African ancestry individuals (Warby et al., 2011); however, A2a was identified in two disease alleles (2/67= 3%) and A2b in thirteen non-disease alleles (13/384= 3.4%).

Table 3.2. A summary of the *HTT* haplogroups/haplotypes and associated allele structures.

Haplogroups	Haplotypes	Allele structures	Non-disease		Disease	
			n	%	n	%
A	*A2a	Q^1-2-2-7-2	-	-	1	1.5
	*A2b	Q^1-2-2-7-2	13	3.4	1	1.5
	A4a	Q^1-2-2-7-2	5	1.3	3	4.5
		Q^1-2-2-12-2	1	0.3	-	-
		Q^1-2-2-13-2	1	0.3	-	-
	A4b	Q^1-2-2-7-2	2	0.5	5	7.5
	A6	Q^1-2-2-7-2	34	8.9	-	-
B	B1	Q^1-2-2-9-2	1	0.3	1	1.5
	B2	**Q^1-2-0-9-2**	25	6.6	**29**	**43.3**
		Q^1-0-0-9-2	-	-	2	3.0
		Q^1-4-2-4-3	-	-	1	1.5
C	C2	Q^1-4-2-7-3	21	5.5	-	-
		Q^1-2-2-8-2	5	1.3	-	-
	C4	Q^1-2-2-9-2	21	5.5	1	1.5
	C4c	Q^1-2-2-6-2	11	2.9	-	-
	C5	Q^1-2-2-10-2	69	18.1	19	28.4
		Q^1-2-2-11-2	18	4.7	-	-
		Q^1-2-2-10-3	1	0.3	-	-
		Q^1-2-0-9-2	2	0.5	-	-
		Q^1-4-2-10-2	4	1.0	-	-
	C8	Q^1-2-2-9-2	7	1.8	-	-
C-SA	C3	Q^1-2-2-10-2	1	0.3	-	-
		Q^1-2-2-9-3	90	23.6	-	-
	C9	Q^1-2-2-6-3	2	0.5	4	6.0
		Q^1-2-2-7-2	7	1.8	-	-
	C10	Q^1-2-2-4-3	1	0.3	-	-
"Other"	O	Q^1-2-2-7-2	36	9.4	-	-
		Q^1-2-2-9-3	2	0.5	-	-
		Q^1-2-2-10-2	1	0.3	-	-
Total			381		67	

The most common non-disease haplogroup/haplotype are indicated in **bold green font** and the most common disease haplogroup/haplotypes are indicated in **bold red font**. The most common disease associated allele structure Q^1-2-0-9-2 (29/67= 43.3 %) is indicated in **bold font**. The two haplotypes that had not been previously identified in African ancestry individuals are indicated by an asterisk (*).

The most common disease allele structure Q^1-2-0-9-2 occurred exclusively on haplotype B2 that appeared to be Africa-specific.

Given that the most common allele structure in the disease alleles was found exclusively on haplotype B2, it was further assessed to determine if the haplotype was African-specific. This was achieved using the online LDhap tool, which calculated the frequencies of the four tag-SNPs that define haplotype B2 in publicly available population data from the 1,000 Genomes Project. This analysis could only be conducted in individuals assumed to be unaffected with HD and thus provided frequencies of the occurrence of haplotype B2 in apparent non-disease alleles in various populations. The geographic distributions of the haplotype B2 frequencies among the non-disease alleles from the 1,000 Genomes Project data are shown in Figure 3.2 (page 56) and tabulated in Appendix 6. The distributions were broadly grouped by continent with African, American, European, South Asian and East Asian populations. The African grouping also included two populations outside of Africa - Americans of African Ancestry in SW USA and African Caribbeans in Barbados.

The result revealed increased haplotype B2 frequencies in the seven African and African ancestry populations, ranging between 6.6% and 9.9%. The highest frequency (9.9%) was identified in the African Caribbeans in Barbados and the lowest frequency (6.6%) in the Americans of African Ancestry in SW USA. Among the non-disease alleles included in the present study, a comparable frequency (6.6%) was identified for haplotype B2 (Table 3.2, page 54). Outside of the continental African populations, Puerto Rico (American), had the highest frequency of haplotype B2 (3.4%), followed by the five East Asian populations (range from 0.5 to 1.0%). The Columbian (American), Utah Residents (European) and Sri Lankan (South Asian) populations had low frequencies (0.5%); and B2 was not detected in the rest of the populations analysed.

Haplotype B2 was shown to have the highest frequencies among the African and African ancestry populations, with reduced frequencies identified in the other populations. The results were comparable with the frequency of B2 in the African ancestry non-disease alleles included in this study. This indicates that although this analysis was only conducted in non-disease alleles, haplotype B2 may be of an African origin and an African-specific haplotype. The Africa HD mutations may have arisen independently on this haplotype.

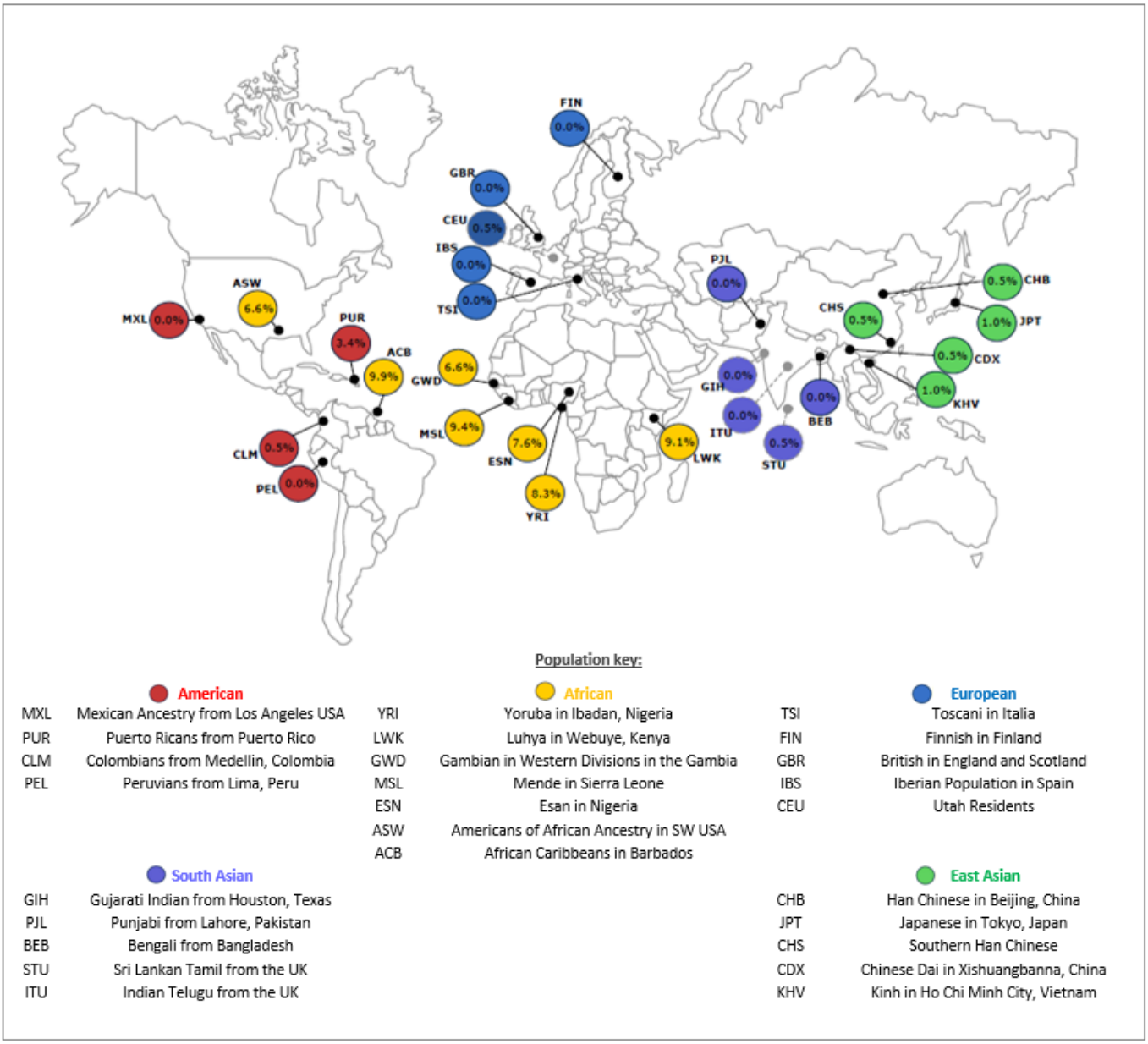


Figure 3.2. Haplotype B2 estimated frequencies for populations in the 1,000 Genomes Project, using the LDhap tool.

3.1.3. Potential modifiers of the HD phenotype

3.1.3.1. HD phenotype assessment

Assessment of the HD phenotype revealed the AoD was an acceptable proxy for the AoO of HD motor symptoms.

The final analysis of the *HTT* locus assessed potential genetic modifiers of the HD phenotype. Unlike the sequence and haplotype diversity analyses, potential modifiers were only assessed in the individuals affected with HD. Prior to assessing the effect of the potential disease modifiers, it was necessary to examine whether age of diagnosis (AoD) could be an acceptable proxy for age of onset (AoO) that is the most characterised measure of disease severity in most studies but was only available in a small number of patients. A linear regression analysis was performed to confirm that AoD could be used as a proxy for AoO in this group of patients; the graphical representation of the linear regression analysis is shown in Figure 3.3. A highly significant correlation ($p = 8.22 \times 10^{-7}$) was shown between the log transformed AoD and AoO, with an R-square value of 0.68. These results confirm that AoD was an acceptable proxy for the HD phenotype as 68% of the variation in the AoD could be attributed to the AoO.

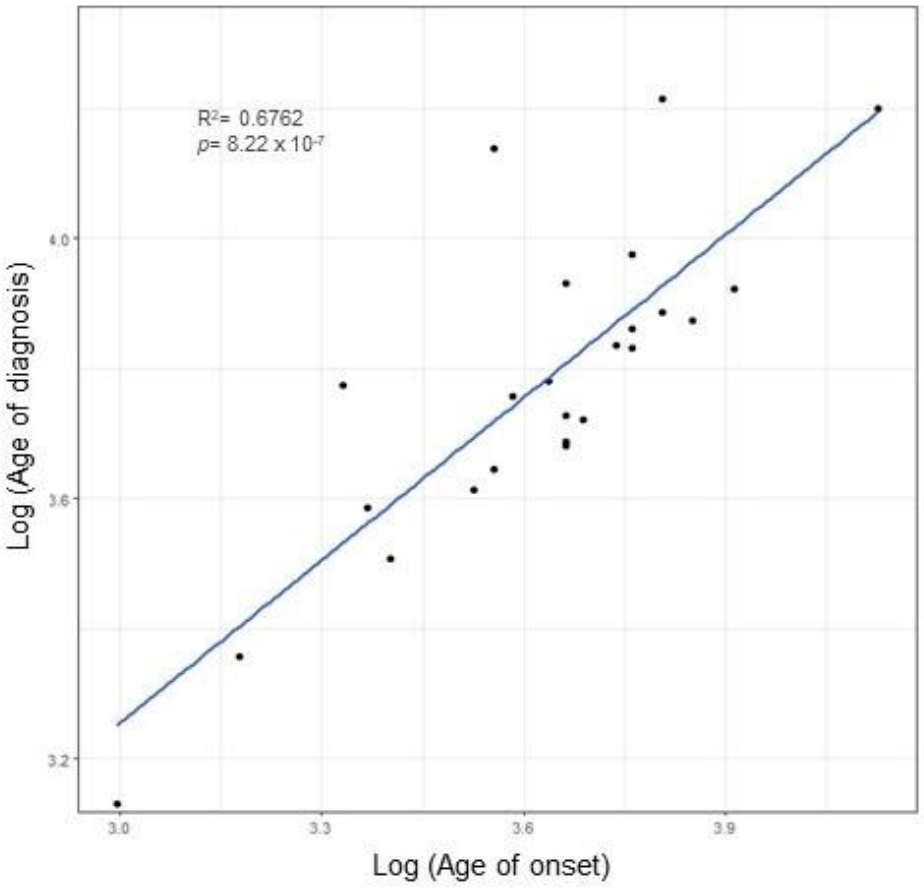


Figure 3.3. Linear regression analysis testing the association between the log transformed AoD and AoO.

The R-square and *p*-values show a significant association ($r^2 = 0.68$ and $p = 8.22 \times 10^{-7}$). The AoD is an acceptable proxy for the AoO in studying the HD phenotype.

3.1.3.2. *HTT* repeat tract modification

The assessment of the effect of the HTT repeat tract revealed that the increased CAG repeat length and the loss of the CCGCCA sequence were significantly associated with an earlier AoD.

The first potential modifier of the HD phenotype (AoD) that was assessed was the effect of the *HTT* repeat tract. This was divided into two parts: the effect of the individual components of the repeat tract (*i.e*. CAG, CAACAG, CCGCCA, CCG, CCT), and the effect of the disease allele structure (*i.e*. all the components of the repeat tract together). Each of these assessments was conducted through the construction of a linear regression analysis that tested the association with the AoD, and a linear model that estimated the effect in years on the AoD.

The CAG repeat was assessed individually as repeats have previously been shown to be the main driver of HD severity (Lee et al., 2012). The graphical representation of the linear regression analysis shown in Figure 3.4, revealed a highly significant association (p = 2.545 x 10^{-14}) with an R-square value of 0.59. These results confirmed the previous findings as the CAG repeat was shown to account for 59% of the variation in the AoD and thus confirmed as the major driver of the HD phenotype.

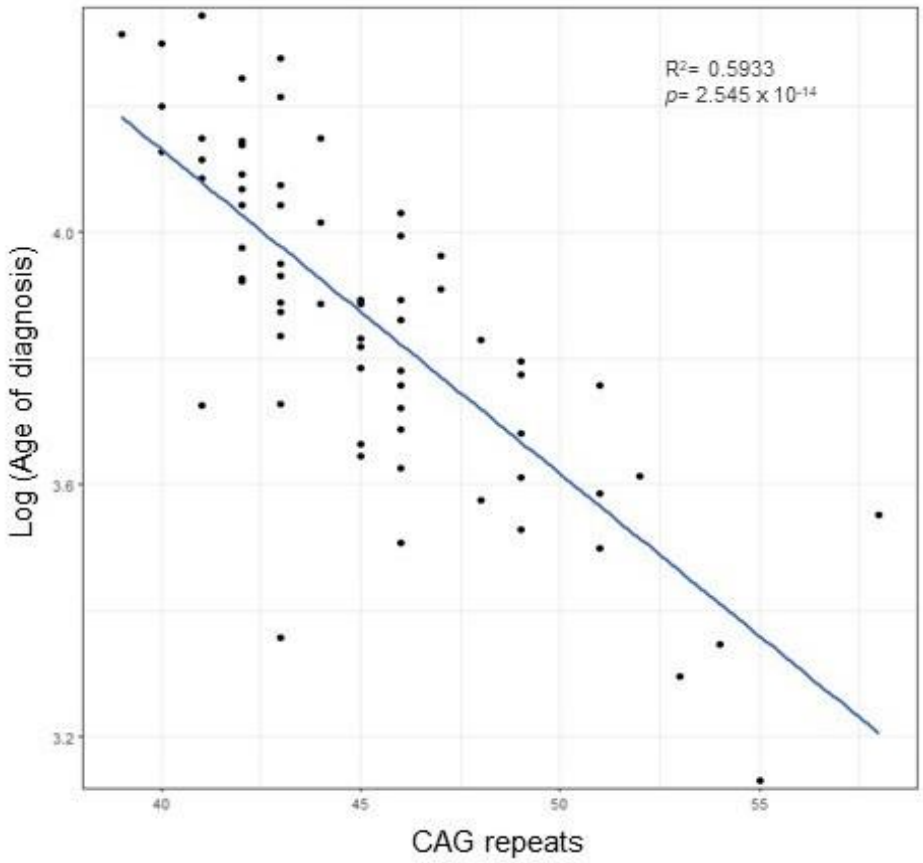


Figure 3.4. Linear regression analysis testing the association between the log transformed AoD and the disease associated CAG repeats.

The R-square and p-values show a significant association (r^2 = 0.59 and p= 2.545 x 10^{-14}). The CAG repeat accounts for the majority of the variation in the HD phenotype. The CAG repeat length was the contiguous number of CAG repeats.

The other components of the *HTT* repeat tract were then assessed individually for their effects, in years, on the AoD. This was done using a linear model to determine associations between the disease phenotype and the explanatory variables (CAG, CAACAG, CCGCCA, CCG, CCT). The result of this linear model shown in Table 3.3, demonstrated a highly significant correlation ($p = 1.44 \times 10^{-10}$) and an R-square value of 0.61 for the overall model. However, only two of the explanatory variables in the model had a significant individual effect on the HD phenotype; the CAG repeat (in line with the findings of the linear regression analysis in Figure 3.4, page 58) and the CCGCCA sequence. The effect in years revealed that every CAG repeat increase resulted in an earlier AoD by 2.9 years ($p = 6.11 \times 10^{-12}$); while the loss of the CCGCCA sequence resulted in an earlier AoD by 4.0 years ($p = 7.34 \times 10^{-4}$).

Table 3.3. Linear model testing the effect of the individual components of the *HTT* repeat tract on the AoD.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
Ln (AoD)~ CAG + CAACAG + CCGCCA + CCG + CCT	0.609	1.44×10^{-10}	64	**CAG**	**-2.902**	**6.11×10^{-12}**
			0=2, 1=61, 2=1	CAACAG	-1.554	4.98×10^{-1}
			0=31, 1=33	**CCGCCA**	**4.029**	**7.34×10^{-4}**
			7/9/10=59, 4/6=5	CCG	-0.258	7.99×10^{-1}
			2=59, 3=5	CCT	2.175	6.80×10^{-1}

The individual components of the *HTT* repeat tract were assessed to separately determine the effect of each on the AoD. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.61$ and $p = 1.44 \times 10^{-10}$). The CAG repeat and CCGCCA sequence were individually significant and are indicated in **bold font**. In the sample size column, the numbers represent how many patients have a particular sequence composition for each component in the *HTT* repeat tract (0=2 represents that two patients have a loss of the CAACAG sequence).

The effect of each allele structure (all components of the repeat tract together) on the AoD was then assessed. The graphical representation of the linear regression analysis is shown in Figure 3.5 (page 60). A linear model was then constructed to determine the effect in years of the CAG repeat and the disease allele structures relative to a reference. The reference used was the grouped typical allele structure Q^1-2-2-P^2-2. The results shown in Figure 3.5 (page 60) and Table 3.4 (page 61), demonstrated a highly significant correlation ($p = 1.36 \times 10^{-10}$) and an R-square value of 0.61. The allele structures composed of all the components of the *HTT* repeat tract accounted for 61% of the variation in the AoD.

The comparison of the R-square value of the linear regression analysis that assessed the AoD and disease allele structures (Figure 3.5) demonstrated an increase by 0.2 compared to the linear regression analysis that assessed the AoD and the CAG repeat (Figure 3.4, page 58). Indicating that the CAG repeat remained the major driver of the HD phenotype, with only a relatively small effect attributable to the other significant explanatory variable, the CCGCCA sequence.

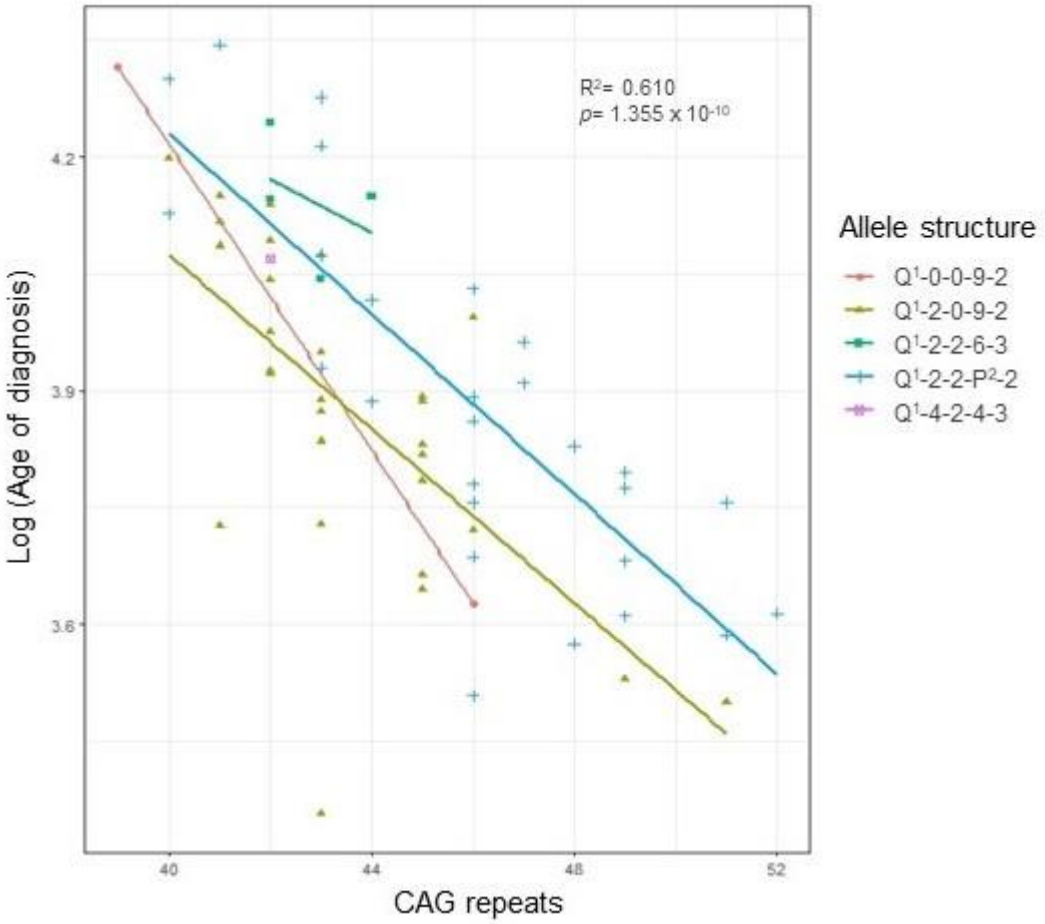


Figure 3.5. Linear regression analysis testing the association between the log transformed AoD and the CAG repeats for each disease allele structure.

The R-square and *p*-values show a significant association ($r^2 = 0.61$ and $p = 1.33 \times 10^{-10}$). The Q^1-0-0-9-2 and Q^1-2-0-9-2 disease allele structures had the earliest AoD.

Although the linear model in Table 3.4 (page 61) demonstrated a significant association, only two of the explanatory variables had a significant effect on the HD phenotype, the CAG repeat and the atypical allele structure Q^1-2-0-9-2. The effect in years revealed that every CAG repeat increase resulted in an earlier AoD by 2.9 years ($p = 5.71 \times 10^{-12}$), in line with the CAG repeat result for the model in Table 3.3 (page 59). The Q^1-2-0-9-2 disease allele structure resulted in an earlier AoD by 7.1 years ($p = 8.15 \times 10^{-4}$) when compared with the reference (or typical) allele structure Q^1-2-2-P^2-2. Thus, patients carrying the commonest African Q^1-2-0-9-2 disease allele structure had a significantly earlier AoD than patients carrying the commonest European Q^1-2-2-P^2-2 allele structure.

Table 3.4. Linear model testing the effect of the allele structures on the AoD.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
Ln (AoD)~ CAG + Allele structures	0.610	1.36×10^{-10}	64	**CAG**	**-2.910**	**5.71×10^{-12}**
			2	Q^1-0-0-9-2	-5.681	2.88×10^{-1}
			29	**Q^1-2-0-9-2**	**-7.133**	**8.15×10^{-4}**
			4	Q^1-2-2-6-3	3.691	3.96×10^{-1}
			1	Q^1-4-2-4-3	-2.489	7.43×10^{-1}

All the components of the *HTT* repeat tract were assessed together as disease allele structures, relative to the reference of the grouped typical allele structure Q^1-2-2-P^2-2. The R-square and *p*-values of the overall model show a significant association (r^2 = 0.61 and $p = 1.36 \times 10^{-10}$). The CAG repeat and Q^1-2-0-9-2 disease allele structure had a significant effect and are indicated in **bold font**.

Although Figure 3.5 (page 60) showed the graphical representation of AoD for each of the allele structures, the analysis did not provide information on the mean AoD of each of the allele structures. The estimated marginal mean AoD in years for each of the allele structures was established using the emmeans function in R (*v*3.4.3). In this analysis the AoD results were corrected for CAG repeat size to limit the bias across patients with larger CAG repeats. The results of the estimated marginal mean AoD across the allele structures are shown in Figure 3.6 (page 62). This analysis confirmed that the Q^1-2-0-9-2 allele structure had the earliest AoD of 45.5 years while the most delayed AoD was Q^1-2-2-6-3 at 56.9 years. Thus, patients with the commonest African Q^1-2-0-9-2 allele structure were confirmed to have the earliest mean AoD while patients with the Q^1-2-2-6-3 allele structure had the most delayed mean AoD.

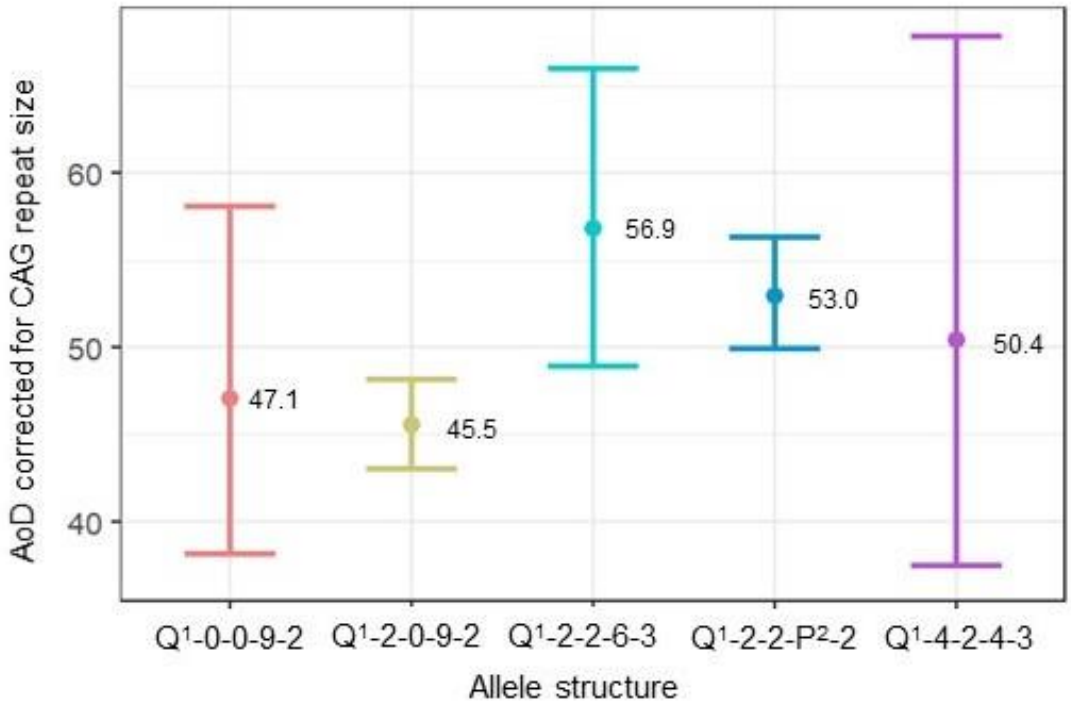


Figure 3.6.The estimated marginal mean AoD for the disease allele structures, corrected for repeat size.

The estimated marginal mean AoD for the allele structures was established using the emmeans function in R (*v*3.4.3). The earliest mean AoD was identified for the Q^1-2-0-9-2 allele structure. The estimated marginal mean AoD for each of the disease alleles were as follows: Q^1-2-0-9-2 (n= 30, 45.5 years: 95% CI = 43.0 to 48.2), Q^1-0-0-9-2 (n= 2, 47.1 years: 95% CI = 38.1 to 58.1), Q^1-4-2-4-3 (n=1, 50.4 years: 95% CI = 37.4 to 67.9), Q^1-2-2-P^2-2 (n=31, 53.0 years: 95% CI = 49.9 to 56.3) and Q^1-2-2-6-3 (n=4, 56.9 years: 95% CI = 48.9 to 66.0.

3.1.3.3. *HTT* haplogroup/haplotype modification

Patients with the allele structure Q^1-2-0-9-2 that occurred exclusively on haplotype B2 had a significantly earlier AoD.

The haplogroup and haplotype background was assessed as potential modifiers of the HD phenotype (AoD). The assessment was divided into two parts; the effect of the haplogroup, followed by the effect of the haplotype. Each of these analyses was conducted through the construction of a linear model that revealed the effect in years.

In the first linear model the disease haplogroups and the CAG repeat were assessed as explanatory variables relative to a reference, haplogroup B. Haplogroup B was used as the reference as it was the most common haplotype among the African ancestry HD patients. The result shown in Table 3.5, revealed a highly significant correlation ($p = 2.054 \times 10^{-10}$) and an R-square value of 0.59. All the explanatory variables, the CAG repeats and haplogroups A, C and haplogroup variant C-SA had a significant effect on the AoD when compared with the reference, haplogroup B. The effect in years revealed that every CAG repeat increase resulted in an earlier AoD by 2.9 years; while haplogroup B resulted in an earlier AoD when compared with haplogroup A by 8.6 years, haplogroup C by 6.2 years and haplogroup variant C-SA by 11.8 years. Thus, patients with allele structures that occurred on haplogroup B had a significantly earlier AoD.

Table 3.5. Linear model testing the effect of the haplogroup background on the AoD.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
Ln (AoD)~ CAG + Haplogroups	0.587	2.054×10^{-10}	64	**CAG**	**-2.903**	**2.09×10^{-11}**
			9	**A**	**8.551**	**1.37×10^{-2}**
			18	**C**	**6.202**	**2.24×10^{-2}**
			4	**C-SA**	**11.752**	**1.24×10^{-2}**

The background haplogroups were assessed relative to a reference, the most common haplogroup B. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.587$ and $p = 2.054 \times 10^{-10}$). Every CAG repeat increase resulted in an earlier AoD by 2.9 years ($p = 2.09 \times 10^{-11}$); haplogroup B resulted in an earlier AoD when compared with haplogroup A by 8.6 years ($p = 2.24 \times 10^{-2}$), haplogroup C by 6.2 years ($p = 1.37 \times 10^{-2}$) and haplogroup variant C-SA by 11.8 years ($p = 1.24 \times 10^{-2}$) as indicated in **bold font**.

The haplotype background was then assessed in a second model examining the individual disease haplotypes and the CAG repeat as explanatory variables relative to haplotype B2 as the reference. The results of the linear model are shown in Table 3.6 (page 64) and revealed a highly significant correlation ($p = 5.412 \times 10^{-9}$) and an R-square value of 0.64. The R-square value of 0.64 indicates an increase in the overall effect on the AoD when only the CAG repeat or the

CAG and allele structure was assessed, thus showing that haplotypes increased the effect on the AoD. The results revealed that the CAG repeats and haplotypes A4a, A4b, B1, C5 and C9 had a significant effect on the AoD when compared with the reference haplotype B2. The effect in years revealed that every CAG repeat increase resulted in earlier AoD by 3.1 years (p = 3.19 x 10^{-11}), which was an increase from the 2.9-year effect seen across all the previous linear models. Haplotype B2 resulted in an earlier AoD when compared with haplotype A4a by 16.8 years, haplotype A4b by 9.1 years, haplotype B1 by 22.3 years, haplotype C5 by 7.8 years and haplotype C9 by 12.3 years. Thus, patients with allele structures that occurred on haplotype B2 had a significantly earlier AoD.

Table 3.6. Linear model testing the effect of the haplotype background on the AoD.

Model	r^2	p-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in years	p-value for explanatory variable
Ln (AoD)~ CAG + Haplotypes	0.643	$5.412x10^{-9}$	64	**CAG**	**-3.069**	**$3.19x10^{-11}$**
			1	A2a	1.864	$8.11x10^{-1}$
			1	A2b	9.225	$2.75x10^{-1}$
			2	**A4a**	**16.826**	**$1.77x10^{-2}$**
			5	**A4b**	**9.086**	**$2.90x10^{-2}$**
			1	**B1**	**22.293**	**$1.88x10^{-2}$**
			1	C4	1.419	$8.57x10^{-1}$
			17	**C5**	**7.771**	**$6.76x10^{-3}$**
			4	**C9**	**12.323**	**$7.85x10^{-3}$**

The haplotype background was assessed relative to a reference, the most common haplotype B2. The R-square and p-values of the overall model show a significant association (r^2 = 0.643 and p = 5.412 x 10^{-9}). Every CAG repeat increase resulted in earlier AoD by 3.1 years (p = $3.19x10^{-11}$); haplotype B2 resulted in an earlier AoD when compared with haplotype A4a by 16.8 years (p = 1.77 x 10^{-2}), haplotype A4b by 9.1 years (p = 2.90 x 10^{-2}), haplotype B1 by 22.3 years (p = 1.88 x 10^{-2}), haplotype C5 by 7.8 years (p = 6.76 x 10^{-3}) and haplotype C9 by 12.3 years (p = 7.85 x 10^{-3}) as indicated in **bold font**.

Although the linear model in Table 3.6 showed the AoD effect in years for each of the haplotypes, the analysis did not provide information on the mean AoD of each of the haplotypes. The estimated marginal mean AoD in years was established as shown in Figure 3.7 (page 65). The results confirmed that haplotype B2 had the earliest AoD at 45.5 years followed by C4 at 46.8 years, A2a at 47.2 years, C5 at 52.5 years, A4b at 53.6 years, A2b at 53.8 years, C9 at 56.5 years, A4a at 60.6 years and lastly B1 at 65.5 years. The earliest mean AoD of 45.5 years was the same for patients with the most common allele structure Q^1-2-0-9-2 as these alleles occurred exclusively on the haplotype background B2.

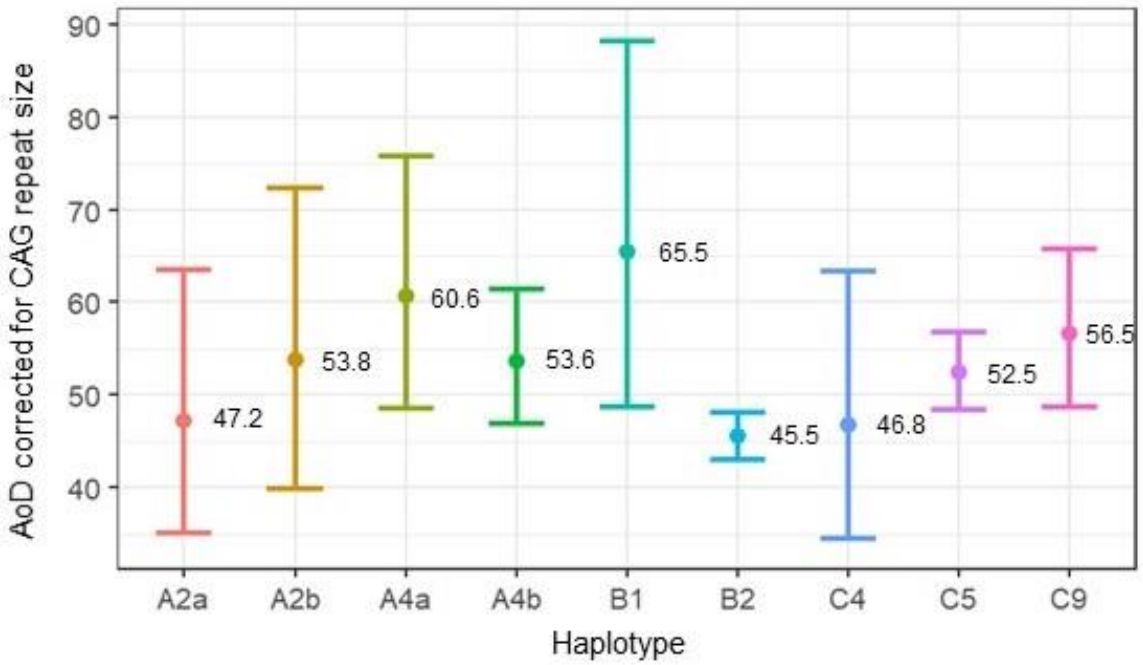


Figure 3.7. The estimated marginal mean AoD for the haplotypes, corrected for CAG repeat size.

The estimated marginal mean AoD for the haplotypes were established using the emmeans function in R (*v*3.4.3). The earliest mean AoD was identified for the haplotype B2. The estimated marginal mean AoD for each haplotypes were as follows: A2a (n=1, 47.2 years: 95% CI = 35.0 to 63.4), A2b (n=1, 53.8 years: 95% CI = 39.9 to 72.4), A4a (n=3, 60.6 years: 95% CI = 48.5 to 75.7), A4b (n=5, 53.6 years: 95% CI = 46.9 to 61.3), B1 (n=1,65.5 years: 95% CI = 48.6 to 88.2), B2 (n=29, 45.5 years: 95% CI = 43.0 to 48.1), C4 (n=1,46.8 years: 95% CI = 34.5 to 63.4), C5 (n=19, 52.5 years: 95% CI = 48.5-56.8) and C9 (n=4, 56.5 years: 95% CI = 48.6 to 65.7.

Neither the Q^1-2-0-9-2 allele structure nor the B2 haplotype was identified to have more of an effect to the AoD.

The previous linear models that assessed the effect of the allele structures (Table 3.4, page 61) and the haplotype background (Table 3.6, page 64) revealed that patients with the Q^1-2-0-9-2 allele structure and haplotype B2 had the same effect on the AoD. A goodness of fit test on 5,000 bootstrapped samples was conducted to determine whether this effect was more likely attributed to the allele structure Q^1-0-2-9-2 or the haplotype background B2. The results of the frequency distribution of the R-square difference between the two models are shown in Figure 3.8. Although the frequency distribution was skewed towards the haplotype having a greater effect on the AoD, there was no statistical indication that the haplotype B2 was better associated with the AoD than the Q^1-2-0-9-2 allele structure.

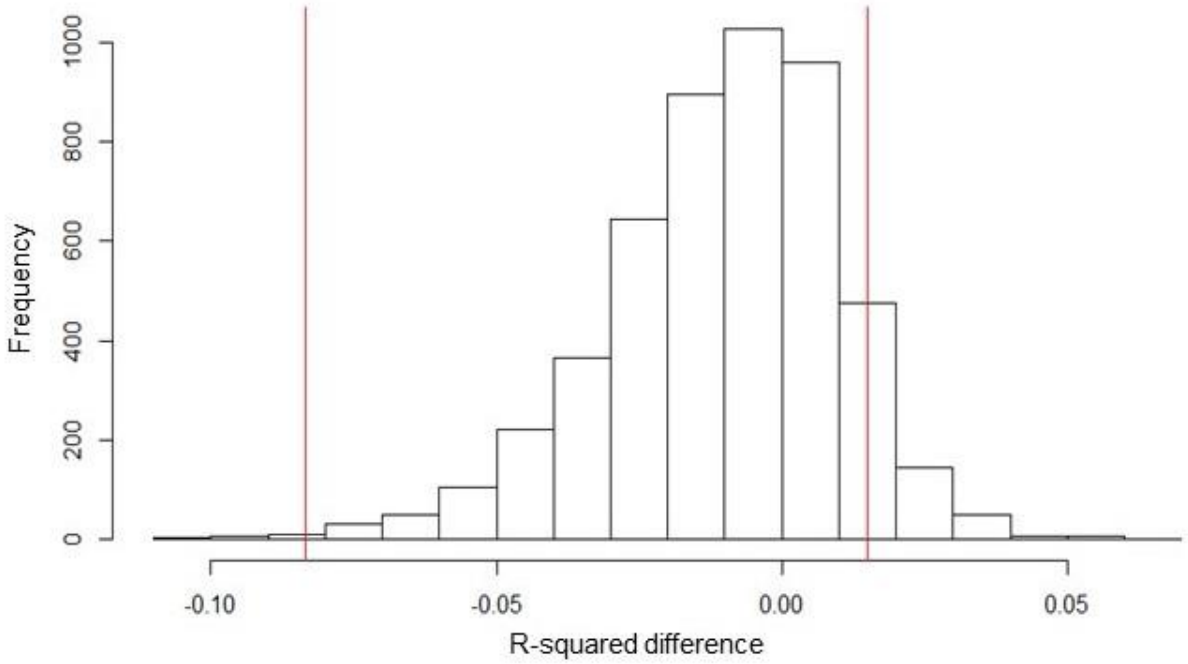


Figure 3.8. The frequency distribution of the R-squared difference between the Q^1-2-0-9-2 allele structure and haplotype B2 models.

The goodness of fit test was conducted on 5,000 bootstrapped samples in R (*v*3.4.3). The effect of the AoD could not be separated out as the 95% confidence interval (red lines) of the R-square difference between the allele structure and haplotype models was around zero.

3.1.3.4. CAG somatic expansion modification

The CAG repeat size was shown to account for majority of the variation in the ratio of somatic expansion.

The last modifier that was assessed for an effect on the AoD was the ratio of somatic expansion due to the instability of the CAG repeat. Prior to the construction of this linear model, an expansion score was established that was then used to assess the effect of the ratio of somatic expansion on the AoD. The ratio of somatic expansion was evaluated through the construction of a linear regression analysis for the log transformed somatic expansion and CAG repeats grouped by allele structure. The results of the linear regression shown in Figure 3.9, revealed a highly significant association ($p < 2.2 \times 10^{-16}$) with an R-square value of 0.77. The CAG repeats therefore accounted for 77% of the variation in the ratio of somatic expansion.

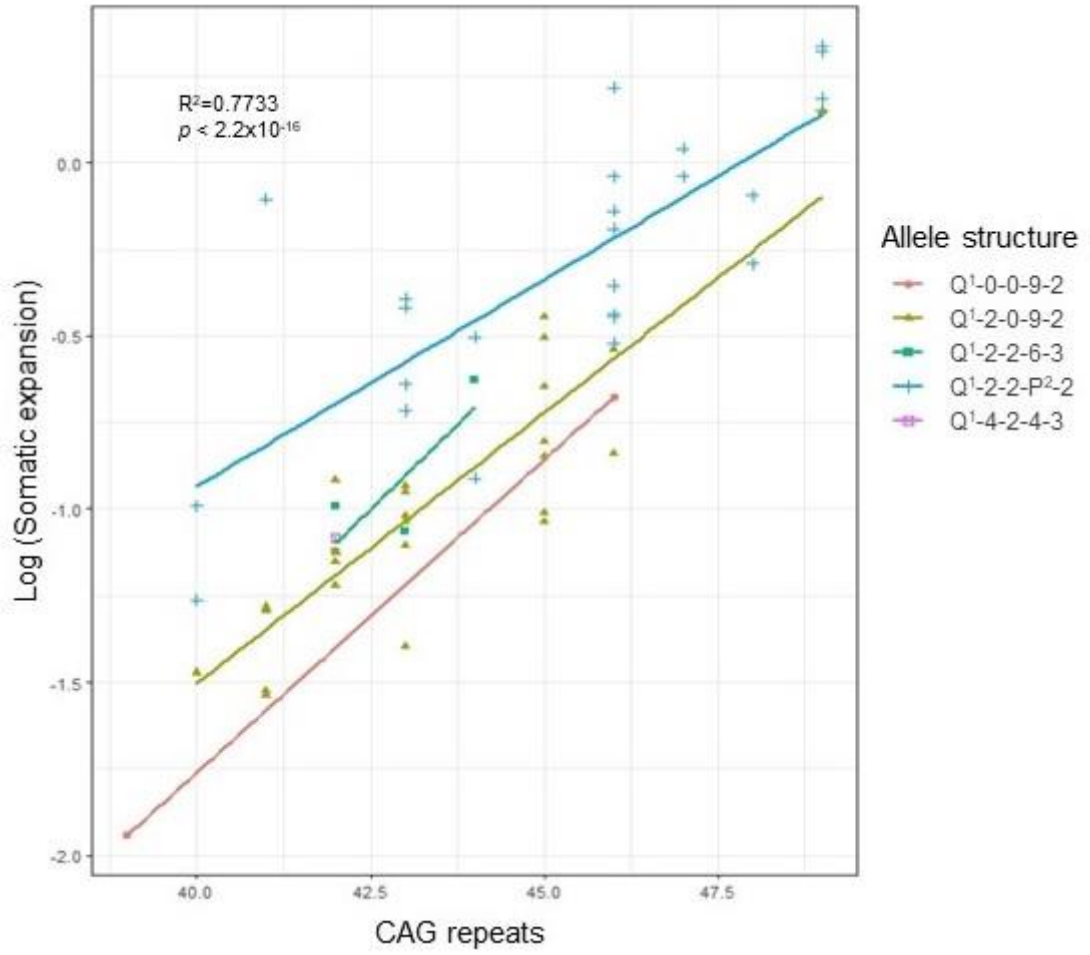


Figure 3.9. Linear regression analysis testing the association between the log transformed ratio of somatic expansion and CAG repeats for each disease allele structure.

The R-square and *p*-values show a significant association ($r^2 = 0.77$ and $p < 2.2 \times 10^{-16}$). The Q^1-0-0-9-2 and Q^1-2-0-9-2 disease allele structures showed the lowest ratio of somatic expansion.

A linear model was then constructed to assess the effect on the ratio of somatic expansion. The explanatory variables used the model included; the CAG repeat, the age at sampling and the interaction between the CAG repeat and the age at sampling. The age of sampling was incorporated in the analysis due to the increase in instability of the CAG repeat with increasing age. It was thus also necessary to include the interaction between the CAG repeats and the age at sampling. The result of the linear model shown in Table 3.7, revealed a highly significant correlation ($p < 2.2 \times 10^{-16}$) and an R-square value of 0.76. The inclusion of the age at sampling in this assessment reduced the overall variation in the ratio of somatic expansion by 0.1% compared to the linear regression analysis in Figure 3.9 (page 67) that only considered the CAG repeats.

The CAG repeat size ($p = 7.81 \times 10^{-16}$) and the age at sampling ($p = 1.79 \times 10^{-3}$) were shown to have a significant effect on the ratio of somatic expansion; however, the interaction between the age at sampling and CAG repeat size was not significant ($p = 7.64 \times 10^{-1}$). The effect on the ratio of somatic expansion revealed that every additional CAG repeat increased the ratio of somatic expansion by 0.13. While every year delay in the age at sampling increased the ratio of somatic expansion by 0.008. In line with the result of the linear regression analysis in Figure 3.9 (page 67), the CAG repeat was shown to be the driver of the ratio of somatic expansion.

Table 3.7. Linear model testing of the effect of the CAG repeat and age at sampling on the ratio of somatic expansion.

Model	r^2	*p*-value for model		Parameter values		
			Sample size	Explanatory variable	Effect to ratio of SE	*p*-value for explanatory variable
Ln (SE)~ CAG + Age at sampling + CAG*Age at sampling	0.758	2.2×10^{-16}	60	**CAG**	**0.131**	**7.81×10^{-16}**
				Age at sampling	**0.008**	**1.79×10^{-3}**
				CAG*Age at sampling	0.000	7.64×10^{-1}

The CAG repeat, age at sampling and the interaction between the two were assessed to determine the effect on the AoD. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.79$ and $p = 2.2 \times 10^{-16}$). The CAG repeat and age at sampling had a significant effect and are indicated in **bold font**.

The effect on the ratio of somatic expansion was further assessed through the construction of a second linear model that also included the allele structures. In the second model the explanatory variables were the CAG repeat, the age at sampling, age at sampling interaction with CAG repeat size and the disease allele structures relative to a reference, the Q^1-2-2-P^2-2 allele structure. The results shown in Table 3.8 (page 69), revealed a highly significant correlation ($p < 2.2 \times 10^{-16}$) and an R-square value of 0.85. The inclusion of the allele structures increased the R-square value by 9% when compared with the model in Table 3.7. In the model, the CAG

repeat and the disease allele structures Q^1-0-0-9-2, Q^1-2-0-9-2 and Q^1-2-2-6-3, were shown to have a significant effect on the ratio of somatic expansion when compared with the grouped typical allele structure. Every CAG repeat increase resulted in an increase by 0.11 ($p = 2.59x10^{-3}$). However, specific allele structures resulted in a decrease in somatic expansion: Q^1-0-0-9-2 by 0.26 ($p = 7.66 \times 10^{-4}$), Q^1-2-0-9-2 by 0.17 ($p = 1.02 \times 10^{-5}$) and Q^1-2-2-6-3 by 0.15 ($p = 1.14 \times 10^{-2}$), when compared with the grouped typical allele structure Q^1-2-2-P^2-2. Thus, patients with the typical allele structure Q^1-2-2-P^2-2 had a significantly higher ratio of somatic expansion overall.

Table 3.8. Linear model testing of the effect of the CAG repeat, age at sampling and the allele structures on the ratio of somatic expansion.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect to ratio of SE	*p*-value for explanatory variable
Ln (SE)~ CAG + Age at sampling + CAG*Age at sampling + Allele structures	0.851	$2.2x10^{-16}$	60	**CAG**	**0.114**	**$2.59x10^{-3}$**
				Age at sampling	0.018	$5.21x10^{-1}$
			2	**Q^1-0-0-9-2**	**-0.257**	**$7.66x10^{-4}$**
			30	**Q^1-2-0-9-2**	**-0.168**	**$1.02x10^{-5}$**
			4	**Q^1-2-2-6-3**	**-0.151**	**$1.14x10^{-2}$**
			1	Q^1-4-2-4-3	0.196	$1.80x10^{-1}$
				CAG*Age at sampling	0.000	$6.13x10^{-1}$

The CAG repeat, age at sampling, interaction between the two and the allele structures relative to a reference, Q^1-2-2-P^2-2 allele structure were assessed to determine the effect on the AoD. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.85$ and $p = 2.2 \times 10^{-16}$). The CAG repeat and the allele structures Q^1-0-0-9-2, Q^1-2-0-9-2 and Q^1-2-2-6-3 had a significant effect and are indicated in **bold font.**

Although somatic expansion had a significant effect on AoD, the earlier AoD seen in patients with the Q^1-2-0-9-2 allele structure could not be attributed to somatic expansion.

Once the ratio of somatic expansion was assessed, the effect it had on the AoD was then considered. This was achieved through the construction of a linear model that included the effect of the expansion score. The expansion score was taken from the residuals of the ratio of somatic expansion, linear model in Table 3.7 (page 68). In the linear model, the explanatory variables were the CAG repeat, expansion score and the disease allele structures relative to a reference, the Q^1-2-2-P^2-2 allele structure. The results of the linear model shown in Table 3.9 (page 70), revealed a highly significant correlation ($p = 1.296 \times 10^{-9}$) and an R-square value of 0.63. The CAG repeat, disease allele structures Q^1-0-0-9-2 and Q^1-2-0-9-2 and the expansion score were shown to have a significant effect on the AoD. Every CAG repeat increase resulted in an earlier AoD by 3.5 years; while the allele structures Q^1-2-0-9-2 and Q^1-0-0-9-2 resulted in an earlier AoD by 10.2 years ($p = 4.20 \times 10^{-5}$) and 11.5 years ($p = 3.43 \times 10^{-2}$) respectively, when

compared with the grouped typical allele structure Q^1-2-2-P^2-2. Lastly, every unit increase in the expansion score resulted in an earlier AoD by 10.6 years (p = 1.20 x 10^{-2}).

Table 3.9. Linear model testing of the effect of the CAG repeats, allele structure and the expansion score on the AoD.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
Ln (AoD)~ CAG + Allele structures + Expansion score	0.625	$1.296x10^{-9}$	60	**CAG**	**-3.504**	**$1.56x10^{-11}$**
			2	**Q^1-0-0-9-2**	**-11.491**	**$3.43x10^{-2}$**
			30	**Q^1-2-0-9-2**	**-10.180**	**$4.20x10^{-5}$**
			4	Q^1-2-2-6-3	-0.840	$8.46x10^{-1}$
			1	Q^1-4-2-4-3	-5.903	$4.11x10^{-1}$
				Expansion score	**-10.600**	**$1.20x10^{-2}$**

The CAG repeat, expansion score and the allele structures relative to a reference, the grouped typical allele structure Q^1-2-2-P^2-2 were assessed to determine the effect on the AoD. The R-square and *p*-values of the overall model show a significant association (r^2 = 0.63 and p = 1.296 x 10^{-9}). The CAG repeat, the allele structures Q^1-0-0-9-2 and Q^1-2-0-9-2 and the expansion score had a significant effect and are indicated in **bold font.**

Although the linear model in Table 3.9 showed the AoD effect in years for the expansion score, the analysis did not provide information on the mean expansion score for each of the allele structures. The estimated marginal mean expansion score for each of the allele structures was established as shown in Figure 3.10 (page 71). The results confirmed that the largest mean expansion score was 0.60, identified in the grouped typical allele structure Q^1-2-2-P^2-2; while the lowest expansion scores were associated with the atypical allele structures Q^1-2-0-9-2 at 0.42 and Q^1-0-0-9-2 at 0.32, both of which are characterised by loss of the CCGCCA sequence. The other two disease allele structures; Q^1-2-2-6-3 and Q^1-4-2-4-3, had an estimated marginal mean expansion score at 0.44. Thus, although somatic expansion was shown to have a significant effect on the AoD, in general, patients with the commonest African Q^1-2-0-9-2 allele structure that had the earliest AoD also had one of the lowest expansion scores. The earlier AoD seen in these patients could not be attributed to somatic expansion.

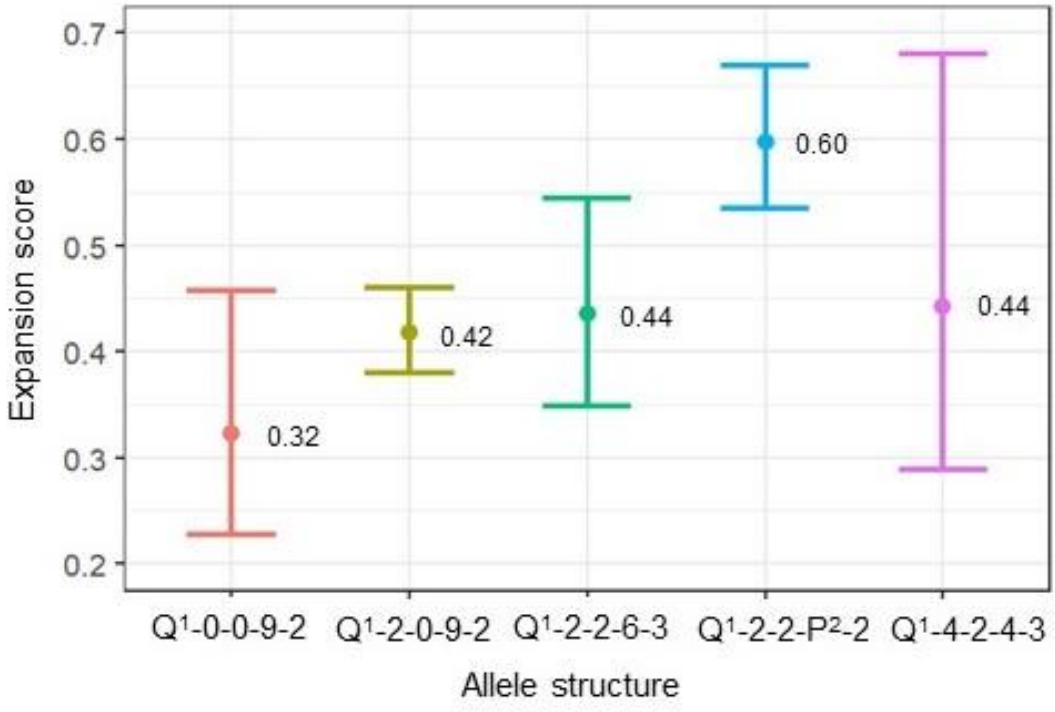


Figure 3.10. The estimated marginal mean expansion score for the allele structures, corrected for CAG repeat size and age at sampling.

The estimated marginal mean Expansion score for the allele structures were established using the emmeans function in R (*v*3.4.3). The lowest mean expansion score was identified for the Q^1-0-0-9-2 and Q^1-2-0-9-2 allele structures. The estimated marginal mean expansion score for the allele structures were as follows: Q^1-0-0-9-2 (n=2, 0.32: 95% CI = 0.227 to 0.458), Q^1-2-0-9-2 (n=30, 0.42: 95% CI = 0.380 to 0.460), Q^1-2-2-6-3 (n=4, 0.44: 95% CI = 0.348 to 0.545), Q^1-2-2-P^2-2 (n=31, 0.60: 95% CI = 0.535 to 0.669) and Q^1-4-2-4-3 (n=1, 0.44: 95% CI = 0.288 to 0.680).

3.2. Discussion of *HTT* analysis

3.2.1. *HTT* repeat tract sequence diversity

A large amount of sequence diversity is identified across the *HTT* repeat tract in both the non-disease and disease alleles in individuals of African ancestry in this study. Of the 17 different allele structures defined, eight are classified as typical (variation only in the number of CAG/CCG repeats); while nine are atypical (variation present throughout the *HTT* repeat tract). The comparison of the non-disease and disease alleles shows that the typical allele structures are more frequent in non-disease alleles, while the atypical allele structures are more frequently associated with the disease alleles. This increased frequency of atypical allele structures occurring in association with the disease alleles may be attributable to expanded CAG repeats being highly unstable and thus prone to mutation; over and above the *HTT* repeat tract in general being a hotspot for mutations (Gellera et al., 1996; Goldberg et al., 1995; Yu et al., 2000).

Within the non-disease alleles, the typical allele structure Q^1-2-2-7-2 is the most commonly occurring. In previous studies conducted on European ancestry individuals, this allele structure is the most common in both non-disease and disease alleles (Ciosi et al., 2019; GeM-HD Consortium, 2019). However, this is not seen in the African ancestry disease alleles in this study, where the most commonly occurring allele structure Q^1-2-0-9-2 is atypical, characterised by loss of the intervening CCGCCA sequence. Although previous research by Ciosi et al. (2019) and the GeM-HD Consortium (2019) has identified this allele structure in European ancestry individuals, it is very rare, especially in disease alleles (Ciosi et al., 2019; GeM-HD Consortium, 2019). This may indicate an African origin for the Q^1-2-0-9-2 allele structure, or support it being an African specific disease allele.

The high proportion of atypical allele structures identified in the African ancestry individuals indicates that the current *HTT* reference gene (locus reference genomic: LRG_763) is not characteristic of the most common genetic sequence (Q^1-2-0-9-2) in this African population. Although the ancestry of the current compiled reference sequence is not known, the *HTT* repeat tract composition of the reference sequence is that of the most common European ancestry allele structure (Q^1-2-2-7-2). Using this sequence as the reference is the reason most common disease allele structure (Q^1-2-0-9-2) in this study is considered atypical, despite being the most frequently identified in African ancestry individuals affected with HD. This further emphasises the need for population-specific reference sequences to be made available, especially for African populations where greater genetic diversity is present (Choudhury et al., 2017).

Three novel allele structures are identified in the African ancestry disease and non-disease alleles and thus unique to this study. One of the three structures is typical in structure, Q^1-2-2-13-2, and two are atypical, Q^1-4-2-10-2 and Q^1-4-2-4-3. These allele structures are novel due to the unique composition of the *HTT* repeat tract occurring in the CCG repeats, CAACAG sequence and/or the CCT repeats. Despite the small cohort size in this study, novel alleles are identified that have not been previously seen in genome-wide studies that included approximately a thousand HD affected individuals of European ancestry (GeM-HD Consortium, 2019). These allele structures are therefore likely to be African-specific. Alternatively, they may occur at very low frequencies and are thus rare in previously studied populations.

The assessment of the *HTT* repeat tract in individuals of African ancestry has allowed for novel and atypical allele structures to be defined, despite the limited sample size, thus expanding the knowledge of the sequence diversity of the *HTT* repeat tract. Considerable frequency differences were identified in the occurrence of the atypical allele structures among African alleles (Non-disease: ~39%; Disease: ~54%) and recently reported European alleles (Non-disease: ~3%; Disease: ~8%) (Ciosi et al., 2019). Further highlighting the genetic sequence differences between populations affected by the same disease, and the importance of research in diverse populations.

The sequence diversity findings presented here are significant as they have important implications for allele-specific therapeutic approaches for African ancestry individuals affected with HD. Allele-specific approaches can be applied through targeting the *HTT* repeat tract itself, the expanded CAG repeat or through targeting SNPs that are in linkage disequilibrium with the expanded repeat (Kay et al., 2014). All of which require a tailored strategy based on the patient genotype. Some of the allele-specific approaches are exploring gene expression modification using antisense oligonucleotides to lower mutant huntingtin (Marxreiter, Stemick and Kohl, 2020). Although currently the selected targets are prevalent in North American and European populations and may not be applicable in non-European populations (Ionis Pharmaceuticals, 2017; Wave Life Sciences, 2019). Thus, unless alternate allele-specific approaches are developed on a population-specific basis, only a portion of the worldwide individuals affected with HD will benefit from potential treatment opportunities (Marxreiter, Stemick and Kohl, 2020).

3.2.2. *HTT* haplogroup/haplotype diversity

The assessment of the broader *HTT* locus in individuals of African ancestry has revealed a large amount of diversity extends to the haplogroup/haplotype background. Sixteen different haplotypes across five haplogroups were observed associated with the non-disease and disease alleles, shown in Table 3.1. The largest proportion of non-disease alleles occur on the haplotype background C5, which is part of haplogroup C; while in the disease alleles, the largest proportion occurs on the haplotype background B2, which is part of haplogroup B.

The comparisons to previously studied haplotypes in European ancestry individuals, reveal the largest proportion of non-disease alleles also occur on haplogroup C, whilst the largest proportion of disease alleles occur on haplogroup A (Warby et al., 2011). The population specific difference in the disease alleles are in line with what has been previously reported and presumed to occur as a result of multiple HD mutations (Andrew et al., 1993). While the similarity seen in the non-disease haplogroups between the African ancestry individuals in this study and the European ancestry individuals studied previously, suggests that the majority of the *HTT* haplotypes migrated out of Africa although at different frequencies.

The European 'high-risk' haplogroup A variants, A2a and A2b, are identified in two African ancestry individuals affected with HD in this study. These variants are defined as 'high-risk' in European ancestry individuals due to their high mutation rate and association with both the expanded disease alleles and intermediate alleles (Warby et al., 2009). Prior to this, A2a and A2b were thought to be absent from East Asian and African ancestry populations (Warby et al., 2011). The results of this study suggest that these haplogroup variants are present in African HD alleles, albeit at very low frequencies (1/67, 1.5%, respectively). The low frequencies may indicate the possibility of admixture in the South African population.

Differences in the allele structures that occur in association with the haplotype background in the non-disease and disease alleles have also been identified. The results show that different allele structures may occur on the same haplotype background. The most common haplotype in the non-disease alleles (C5) is associated with five different allele structures, two typical and three atypical. The most common haplotype in the disease alleles (B2) is associated with three different atypical allele structures. Notably, the most common of these atypical allele structures, Q^1-2-0-9-2, occurs exclusively on haplotype B2 which is more common in Africa and suggests an African origin for most HD in Africa.

Although haplotype B2 has been identified in European ancestry individuals, it is rare and differs from the African B2 haplotype at a single SNP, rs762855 (Kay et al., 2018). Previous research by Baine et al. (2013) describes the African B2 haplotype with a G allele at rs762855, found exclusively in black African ancestry or admixed individuals. A study by Kay et al. (2018) and email correspondence (Collins, 2020, personal communication; Hayden, 2020, personal communication and Wright, 2020, personal communication) confirm the description of the European haplotype B2 with an A allele at rs762855, found in European ancestry individuals. Although this SNP is not included in the subset of tag-SNPs used to define haplotypes in this study, the haplotype B2 defined in this study is likely the African haplotype B2.

The assessment of haplotype B2 as an African specific haplotype has been achieved using the frequencies of the tag-SNPs that define the haplotype in populations studied in the 1,000 Genomes Project. The results reveal an increased frequency of haplotype B2 in the seven African and African ancestry populations, with the highest frequency observed in African Caribbeans in Barbados; a population that is ancestrally African although no longer on the Africa continent due to forced relocation to Barbados where there have subsequently been relatively low levels of non-African admixture (Benn-Torres et al., 2008). Apart from the African populations, the highest frequency is reported in the Puerto Rican population; an observation that is likely attributable to the African ancestry contribution within the population. A study by Via et al. (2011) revealed that 21.2% of the Puerto Rican population are of Sub-Saharan African ancestry (Via et al., 2011).

These findings reveal that although haplotype B2 was not absent from other populations, the frequencies were considerably reduced when compared with the African populations, and those with African ancestry. This supports an African origin and an African-specific haplotype, with the occurrence in other non-African ancestry populations likely resulting from low-level admixture between populations.

3.2.3. Potential modifiers of the HD phenotype

3.2.3.1. HD phenotype assessment

Before the assessment of the potential modifiers of the HD phenotype in the African ancestry individuals with HD, the HD phenotype is established. The most characterised measure of the HD phenotype is the onset of motor symptoms (Lee et al., 2012), in this study it is referred to as the AoO. Despite this information being absent for most of the individuals affected with HD in this study, AoD information is available. The comparison between the two shows that the AoD is a suitable proxy in the absence of AoO, with a highly significant association identified.

Extensive research conducted in European ancestry individuals affected with HD extends to both *cis*-acting modifiers at the *HTT* locus and *trans*-acting modifiers focusing on the DNA repair or maintenance genes identified in the HD GWAS studies (GeM-HD Consortium, 2015; GeM-HD Consortium, 2019). The potential modifiers assessed in this study evaluated various components of the *HTT* locus, the *cis*-acting modifiers. This included the assessment of the effect of the *HTT* repeat tract, the broader haplogroup/haplotype background and lastly the contribution of the somatic expansion of the CAG repeat.

3.2.3.2. *HTT* repeat tract modification

The *HTT* repeat tract was the first modifier assessed for an effect on the AoD as research on European ancestry individuals affected with HD show the CAG repeats to be the primary determinant of the HD phenotype, the age of onset of motor symptoms (Lee et al., 2012). Significant effects have also been demonstrated for sequence variants within the repeat (Wright et al., 2019). The assessment of the *HTT* repeat tract is divided into two parts, the evaluation of the effect of the individual components of the repeat tract (CAG, CAACAG, CCGCCA, CCG, CCT) and then the effect of the disease allele structures (all components of the *HTT* repeat tract).

The assessment of the individual components in this study reveals the CAG repeat and the intervening CCGCCA sequence have a significant effect on the AoD. In line with previous research, the CAG repeat accounts for approximately 60% of the variation in the AoD and the effect of the CAG repeat shows that every repeat increase results in an earlier AoD of 2.8 years. While the effect of the intervening CCGCCA sequence shows that the loss of the CCGCCA sequence is associated with an earlier AoD by 4.0 years. This CCGCCA effect has not been previously identified in European ancestry individuals affected with HD. It is rather the loss of the intervening CAACAG sequence that has been reported to have a

significant effect on the AoO (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). Duplications of this CAACAG sequence are associated with a delayed AoO, while the loss is associated with an earlier AoO, however after adjusting for the pure CAG repeat size the CAACAG duplication does not appear to have a residual effect on the HD phenotype (Ciosi et al., 2019). In this study after adjusting for the pure CAG repeat size, the loss of the CCGCCA sequence still has a residual effect on the HD phenotype. Thus, it is not just the effect of the CAG repeats which is causing the effect to the AoD despite it being the primary determinant of the HD phenotype, the CCGCCA sequence also has a significant contribution. This effect of the CCGCCA sequence has not been implicated in previous studies.

The assessment of the effect of all the components of the *HTT* repeat tract considered together reveals that is has a significant effect on the HD phenotype, accounting for approximately 61% of the variation in the AoD. However, the CAG repeat remains the primary determinant driving this effect in all populations, including the African ancestry population in this study. In this assessment, the most common disease allele structure Q^1-2-0-9-2 that is characterised by the loss of the CCGCCA sequence is identified to have a significant effect on the AoD. Patients of African ancestry with the Q^1-2-0-9-2 allele structure have an earlier AoD by 7.1 years compared with patients with the reference allele structure Q^1-2-2-P^2-2. Considerable differences are identified between different allele structures, with the earliest mean AoD in Q^1-2-0-9-2 at 45.5 years, while the most delayed AoD is in Q^1-2-2-6-3 at 56.9 years although only identified in four HD patients.

The Q^1-2-0-9-2 allele structure characterised by the loss of the intervening CCGCCA sequence is thus proposed as a *cis*-acting modifier of the HD phenotype in the African ancestry individuals affected with HD. However, at this stage it cannot be distinguished as to whether this is a direct effect or the effect of a *cis*/*trans*-acting modifier in linkage disequilibrium. Although the Q^1-2-0-9-2 allele structure has been reported in European HD patients this modifier effect to the HD phenotype has not been seen in European ancestry individuals with HD; however, minimal variation is reported in the proline sequence downstream of the CAG repeat (Ciosi et al., 2019; GeM-HD Consortium, 2019). The identification of Q^1-2-0-9-2 as a *cis*-acting modifier provides a potential target for an allele-specific approach that could be applied in an African ancestry context as it is not only the most common disease allele structure but also results in the earliest age of diagnosis.

These findings also have clinical implications for patients that may have a considerably earlier AoD, translating to an earlier AoO depending on the allele structure. The current diagnostic molecular genetic test for HD does not provide any information on the allele structure as it uses capillary electrophoresis which can only estimate the CAG/CCG repeat size (Losekoot et al., 2013). In African ancestry individuals, where most patients have an atypical structure, a future sequencing-based approach may have more benefit. That is capable of accurate CAG repeat sizing and the exact allele structure, which have been shown in this study to have a considerable effect on the AoD. Although there is currently no effective treatment to delay disease, halt or reverse the progression of HD, better earlier diagnostic approaches could potentially provide more effective symptomatic treatment of the movement disorder and psychiatric symptoms that may improve the quality of life of the patient.

3.2.3.3. *HTT* haplogroup/haplotype modification

The haplogroup/haplotype background is the second modifier assessed for an effect on the AoD. This broader component of the *HTT* locus is assessed due to the association of certain haplotypes with disease and the hypothesis of 'high-risk' haplotypes having *cis*-elements that predispose to repeat expansion (Warby et al., 2009; Warby et al., 2011). The assessment is divided into two parts, first the effect of the haplogroup background and the haplotype background. The haplogroup assessment reveals patients with the most common haplogroup B have a significantly earlier AoD between 6.2 to 11.8 years in comparison to the other haplogroups. Despite the significant association with the AoD, the inclusion of the haplogroups results in a slight reduction (R-squared value) to the variation to AoD. This may suggest that the variation in the AoD is not better explained through the inclusion of the haplogroups.

This contrasts with the assessment of the effect of the haplotypes that reveals an increase in the overall association with the AoD. Haplotype B2 has a significant negative effect on the AoD in comparison to the other significant haplotypes. Thus, patients with disease alleles on haplotype B2 had an earlier AoD that ranges from 7.8 to 22.3 years. Patients with alleles on haplotype B2 have the earliest mean AoD of 45.5 years while the most delayed AoD is in haplotype B1 at 65.5 years although only identified in one HD patient. The result for haplotype B1 and the other haplotypes with small sample sizes should be considered with caution. The earliest mean AoD of 45.5 years is also seen in patients with the most common disease allele structure Q^1-2-0-9-2 as these alleles occur exclusively on the haplotype background B2. The designation of a 'high-risk' haplotype for haplotype B2 in individuals

of African ancestry is thus proposed. The repeat size distribution of the Q^1-2-0-9-2 allele structure on haplotype B2 would however have to be assessed and compared to the high-risk haplogroup variants A1/A2 in Europeans where there is thought to be more intermediate alleles and instability (Warby et al., 2009).

To determine if the driver of the *cis*-effect causing the earlier AoD is the Q^1-2-0-9-2 disease allele structure or haplotype B2 these components have been assessed further. The findings reveal that neither the allele structure nor the haplotype had more of a significant effect on the AoD. Thus, at this stage the effect could not be separated out as there is no statistical indication that the earlier AoD exhibited in these patients is better explained by the Q^1-2-0-9-2 allele structure or haplotype B2, especially as they co-occur. It may be possible that there is another component on the *HTT* locus that is in linkage disequilibrium that is causing the variation in AoD.

3.2.3.4. CAG somatic expansion modification

In previous studies on European ancestry individuals affected with HD, somatic expansion significantly associates with an earlier AoO and poorer HD outcomes (Ciosi et al., 2019; Wright et al., 2019). Somatic expansion is affected by various factors that include the CAG repeat size, the age of the individual and the cell type (Kennedy et al., 2003; Telenius et al., 1994; Veitch et al., 2007). Recently, *trans*-acting modifiers have also been shown to be associated with somatic expansion and to have a significant effect on the HD phenotype (Ciosi et al., 2019; GeM-HD Consortium, 2019). In this study, somatic expansion is assessed through the association with *cis* components, to produce an expansion score. This expansion score is used to determine the effect of the ratio of somatic expansion on the AoD.

The assessment of the ratio of somatic expansion reveals the CAG repeat size to have a significant effect, accounting for 76% of the variation in somatic expansion, with every additional CAG repeat resulting in an increase in the ratio of somatic expansion by 0.131. The residuals of this assessment are used to establish the expansion score, which is then used to determine the effect on the AoD. This assessment shows that the expansion score has a significant effect on the AoD, with every unit increase in the expansion score resulting in an earlier AoD by 10.6 years.

Although somatic expansion is significantly associated with an earlier AoD, differences in the ratio of somatic expansion are identified among the allele structures. The most common African ancestry disease allele structure Q^1-2-0-9-2 has one of the lowest expansion scores while the highest occurs in the combined typical allele structures Q^1-2-2-P^2-2. Prior to these

findings, it was hypothesised that the most common disease allele structure Q^1-2-0-9-2, is associated with increased somatic expansion that contributes to the earlier AoD. Previous studies of European ancestry patients have shown increased somatic expansion is associated with an earlier onset of HD (Ciosi et al., 2019; GeM-HD Consortium, 2019). In this study, however the findings paradoxically reveal that patients with the Q^1-2-0-9-2 allele structure have the lowest somatic expansion, suggesting that the reducing effect on the AoD in these individuals is not mediated through the effects of somatic expansion in the blood. The findings therefore suggest that there is a different disease modification mechanism occurring in the African ancestry individuals affected with HD that have the Q^1-2-0-9-2 allele structure and that differs from the proposed mechanism that is thought to occur in European ancestry individuals affected with HD. Although this may include effects on the *HTT* transcript by altered folding, splicing, canonical or RAN translation (Bañez-Coronel et al., 2015; Busan and Weeks, 2013; Krauss et al., 2013; Neueder et al., 2017). Alternatively, it is possible there is an effect mediated by a linked variant (Becanović et al., 2015).

The results of the *HTT* analysis revealed a high level of sequence diversity was observed in African ancestry individuals with the identification of three novel allele structures. The most common disease-associated allele structure Q^1-2-0-9-2 was defined as atypical, due to the loss of the CCGCCA sequence. Although the CAG repeat was confirmed to be the main driver of disease severity, as it is in the European population, the Q^1-2-0-9-2 allele structure was associated with an earlier age of diagnosis of approximately 7.1 years and occurred exclusively on haplotype B2. Although the somatic expansion was associated with an earlier age of diagnosis, the Q^1-2-0-9-2 allele structure displayed reduced somatic expansion. Thus, the Q^1-2-0-9-2 allele structure occurring on haplotype B2 was proposed as an African *cis*-acting modifier that appears to modify the age of diagnosis, through a mechanism that was not driven by somatic expansion. That appears to reduce the age of diagnosis in African ancestry individuals affected with HD, through a mechanism that is not driven by somatic instability, although the underlying mechanism is currently undetermined.

4. Chapter 4: *JPH3* analysis

This chapter reports on the *JPH3* repeat tract sequence diversity, haplogroup/haplotype background diversity and potential modifiers of the HDL2 phenotype in African individuals affected with HDL2. The *JPH3* repeat tract sequence diversity and haplogroup/haplotype background diversity was established in individuals affected and unaffected with HDL2. This was achieved through sequencing of the *JPH3* repeat tract, to define allele structures. This was followed by the genotyping and phasing of the *JPH3* tag-SNPs to construct haplogroup/haplotype background. In contrast to the *HTT* analysis, additional unaffected individuals from six worldwide populations were included in the diversity analyses to establish a comparison to other populations with different ancestries. Potential modifiers of the HDL2 phenotype were investigated in the HDL2 patients using statistical methods to assess the effects of the *JPH3* repeat tract, haplogroup/haplotype background and ratio of somatic expansion of the CTG repeats. Unlike *HTT*, the *JPH3* locus has not been well studied with no previous assessment of the sequence diversity of the *JPH3* repeat tract or the ratio of somatic expansion. This study thus allowed for the novel assessment of the *JPH3* locus in individuals affected and unaffected with HDL2.

4.1. Results of *JPH3* analysis*JPH3* repeat tract sequence diversity

Limited sequence diversity was identified in the JPH3 repeat tract across all individuals. In the HDL2 disease alleles, no sequence variants were identified in stark contrast to the diversity seen at the HTT locus.

The *JPH3* repeat tract sequence diversity was assessed as the exact sequence composition of the repeat tract has not been widely studied previously. Like the *HTT* locus, diagnostic testing only assesses the size of the CTG repeat containing fragments and research conducted in HDL2 is limited due to the rarity of the disease and the fact that it only affects individuals of African ancestry. Due to the clinical similarity with HD and the triplet repeat expansion present in both conditions, it was hypothesised that the similarities may extend to other components of the *JPH3* locus. To explore the scenario occurring in the *JPH3* repeat tract, the sequence diversity was established for a total of 586 samples from South African, African ancestry individuals and various other populations from different ancestries. This was achieved through the high-throughput sequencing of the *JPH3* repeat tract of 32 individuals affected with HDL2 and 554 unaffected individuals that resulted in 32 disease alleles and 1,140 non-disease alleles.

The sequence diversity was established based on the comparison of the structure of the *JPH3* repeat tract with the reference sequence. The typical alleles were defined if the allele structure was characteristic of the *JPH3* reference gene (ENST00000301008.5), made up of a polyleucine

encoding CTG repeat tract comprising a variable number of CTG repeats. Atypical alleles were defined based on variants within the CTG repeat tract and the 5'- and 3'-flanking regions. Unlike the *HTT* repeat tract, no other components in and around the repeat tract are considered part of the *JPH3* repeat tract. The upstream or downstream flanking regions of the CTG repeat were also considered in this study.

The high-throughput sequencing results revealed a limited amount of sequence diversity across the non-disease and disease allele with only six different allele structures identified. The occurrence and distribution of the allele structures identified is shown in Table 4.1 (page 83) and the schematics of the composition of allele structures is shown in Figure 4.1 (page 84). Of the six allele structures identified, one typical allele structure was identified. Unlike the *HTT* locus, only one typical allele was identified where variation was present in the CTG repeat only. Five atypical alleles were defined based on interruptions in the CTG repeat tract or variants in the upstream or downstream flanking regions.

The five atypical allele structures were designated numerical identifiers, atypical 1 to atypical 5. Atypical 1 described a 5' flanking variant in which an A to T missense variant (rs182521282) resulted in a glutamine (CAG) to leucine (CTG) amino acid change 29 bp upstream from the CTG repeat. Atypical 2 described another 5'-flanking variant in which a C to G missense variant (rs539856127) resulted in a proline (CCA) to arginine (CGA) amino acid change 11 bp upstream of the CTG repeat. Atypical 3 described an interruption in the CTG repeat in which a C to T variant (rs564806887) occurred at CTG 4 with no amino acid change, leucine (CTG and TTG). Atypical 4 described another interruption in the CTG repeat in which a C to A missense variant (rs532306398) resulted in a leucine (CTG) to methionine (ATG) amino acid change at CTG 16. Atypical 5 described a 3'-flanking variant in which an A to G missense variant (rs115916470) resulted in an asparagine (AAC) to serine (AGC) amino acid change 23 bp downstream from the CTG repeat. A detailed schematic of the *JPH3* repeat tract sequence composition with the respective variants indicated is shown in Appendix 7.

All the disease alleles were defined as typical allele structures (32/32, 100%), with the only variation in the number of CTG repeats that ranged from 42 to 53 repeats. This was in stark contrast to the *HTT* disease allele where most of the HD patients had an atypical allele structure on their expanded allele. In the non-disease alleles, the typical allele structure (1,056/1,140, 92.6%) occurred most commonly with CTG repeats that ranged from 5 to 33 repeats. Unlike the disease alleles, five atypical allele structures were also identified (84/1,140, 7.4%). The most commonly occurring atypical allele structure was Atypical 5 (73/1,140, 6.4%) that occurred on

CTG repeats ranging from 12 to 19 repeats. This was followed by Atypical 4 (7/1,140, 0.6%) that occurred on CTG repeat ranging from 15 to 16 repeats, Atypical 1 (2/1,140, 0.2%) that occurred on CTG repeat 16 and Atypical 2 and 3 (1/1,140, 0.1%) seen in only one non-disease allele each and occurred on CTG repeat 13 and 10, respectively. To ascertain the significance of the sequence diversity findings, a statistical analysis of the comparison of the frequency of the disease and non-disease allele structures was undertaken. However, no significant frequency differences were identified between the non-disease and disease alleles, the results of the Fisher exact test results are shown in Table 4.1.

Table 4.1. A summary of the *JPH3* disease and non-disease associated allele structures.

Allele type	Variant	CTG repeats		Allele occurrence				Fisher exact test
		Non-disease range	Disease range	Non-disease 1140		Disease 32		
				n	%	n	%	*p*-value
Typical	-	5 – 33	42 – 53	1,056	92.6	32	100.0	$6.36x10^{-1}$
Typical alleles subtotal				**1,056**	**92.6**	**32**	**100.0**	
Atypical 1	5'-flank 1: CTG	16	-	2	0.2	-	-	1
Atypical 2	5'-flank 2: CGA	13	-	1	0.1	-	-	1
Atypical 3	CTG interruption 1: TTG	10	-	1	0.1	-	-	1
Atypical 4	CTG interruption 2: ATG	15 – 16	-	7	0.6	-	-	1
Atypical 5	3'-flank: AGC	12 – 19	-	73	6.4	-	-	$2.571x10^{-1}$
Atypical alleles subtotal				**84**	**7.4**	**0**	**0**	

The nucleotides indicated in red font show the location of the variant that resulted in the atypical allele structure. The Fisher exact test results are shown above however there were no statistically significant differences between the non-disease and disease alleles.



Figure 4.1. Schematic representation of the *JPH3* repeat tract allele structures in the non-disease and disease

The *JPH3* repeat tract is composed of a region that codes for leucine with a variable number of CTG repeats. The CTG repeats are surrounded by 5'/3' flanking regions, indicated in the green. Typical alleles were grouped and defined as such if the allele structures were characteristic of the *JPH3* reference gene (Ensembl: ENST00000301008.5). The atypical alleles were defined based on deviations from the reference gene, due to variants within the CTG repeat and the 5'/3' flanking regions.

Four of the five atypical allele structures were identified in the African ancestry populations.

The non-disease alleles were then assessed based on the breakdown of allele structures identified in the seven populations studied as shown in Table 4.2. This allowed for the establishment of the diversity across various populations, with differences identified in the occurrence of atypical alleles among populations. In the three African ancestry populations from South Africa, The Gambia and Kenya, four of the five atypical allele structures (Atypical 1, 2, 4 and 5) were identified. Of these four atypical alleles structures, three (Atypical 2, 4 and 5: 4/108, 4.6%) were identified in the non-disease alleles from The Gambia. While two were identified in the non-disease alleles from Kenya (Atypical 1 and 5: 22/150, 14.6%) and South Africa (Atypical 4 and 5: 56/408, 13.7%), respectively. The remaining atypical allele structure (Atypical 3: 1/146, 0.7%) was identified in a South Asian ancestry non-disease allele from Bangladesh. No atypical allele structures were identified in the East Asian ancestry non-disease alleles from Japan and China and in the European ancestry non-disease alleles from Scotland as all the alleles had a typical structure.

Table 4.2. A summary of the *JPH3* non-disease allele structures in the various populations.

Allele type	Allele occurrence per population														**Allele total**
	RSA		**GWD**		**LWK**		**BEB**		**JPT**		**CDX**		**SCT**		
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	
Typical	352	86.3	103	95.4	128	85.3	145	99.3	78	100.0	80	100.0	170	100.0	**1,056**
Atypical 1	-	-	-	-	2	1.3	-	-	-	-	-	-	-		**2**
Atypical 2	-	-	1	0.9	-	-	-	-	-	-	-	-		-	**1**
Atypical 3	-	-	-	-	-	-	1	0.7	-	-	-	-	-		**1**
Atypical 4	4	1.0	3	2.8	-	-	-	-	-	-	-	-	-	-	**7**
Atypical 5	52	12.7	1	0.9	20	13.3	-	-	-		-	-	-	-	**73**
Population total	**408**		**108**		**150**		**146**		**78**		**80**		**170**		**1,140**

Population key: RSA: South Africa, GWD: The Gambia, LWK: Kenya, BEB: Bangladesh, JPT: Japan, CDX: China and SCT: Scotland.

The assessment of available data for the atypical allele structure variants revealed both similarities and differences in the frequencies and population occurrence within the populations studied for the atypical alleles. For atypical 1, the allele frequency in gnomAD was 1.19×10^{-4}, however was only identified in African ancestry or Latin or admixed American individuals. For atypical 2, the allele frequency in the 1,000 Genomes Project was 1.00×10^{-3}, however it was only identified in African ancestry individuals from The Gambia. For atypical 3, the allele frequency in gnomAD was 5.04×10^{-4}, across various populations although the highest frequency was identified in South Asian individuals. For atypical 4 the allele frequency in gnomAD was 3.20×10^{-5}, however it was only identified in one African ancestry individual. For

atypical 5, the allele frequency in gnomAD was 3.29×10^{-3}, across various populations with the highest frequency in African ancestry individuals. For atypical 1 to 5 there was no clinical significance identified and the annotation prediction tools identified that the most severe consequence of the variants respectively was a non-coding transcript variant.

The African ancestry populations generally had larger mean CTG repeats and wider repeat distributions.

The assessment of the CTG repeat length in these populations revealed that the difference identified in the sequence diversity extended to the mean CTG repeat size and repeat distributions across the populations. The largest estimated mean repeat size was identified in the non-disease alleles from The Gambia at 16.8, while the smallest was identified in the non-disease alleles from China at 14.1. The mean CTG repeat size and CTG repeat distributions for all of the populations are shown in Figure 4.2 (page 87). The CTG repeat distributions showed considerable differences in distribution among the three African ancestry populations compared to the European ancestry and Asian ancestry populations. The widest repeat distribution was observed in the African populations, The Gambia (CTG range: 8 to 33), followed by Kenya (CTG range: 8 to 28) and South Africa (CTG range: 5 to 27). The narrowest repeat distributions were observed in the Asian populations, Japan (CTG range: 13 to 16), Bangladesh (CTG range: 10 to 17) and China (CTG range: 11 to 17) and European population, Scotland (CTG range: 9 to 18). These findings have shown population specific differences despite the assessment being conducted only on non-disease alleles. The repeat distributions in the African populations were wider and generally had larger estimated mean CTG repeat size when compared with the European and Asian populations.

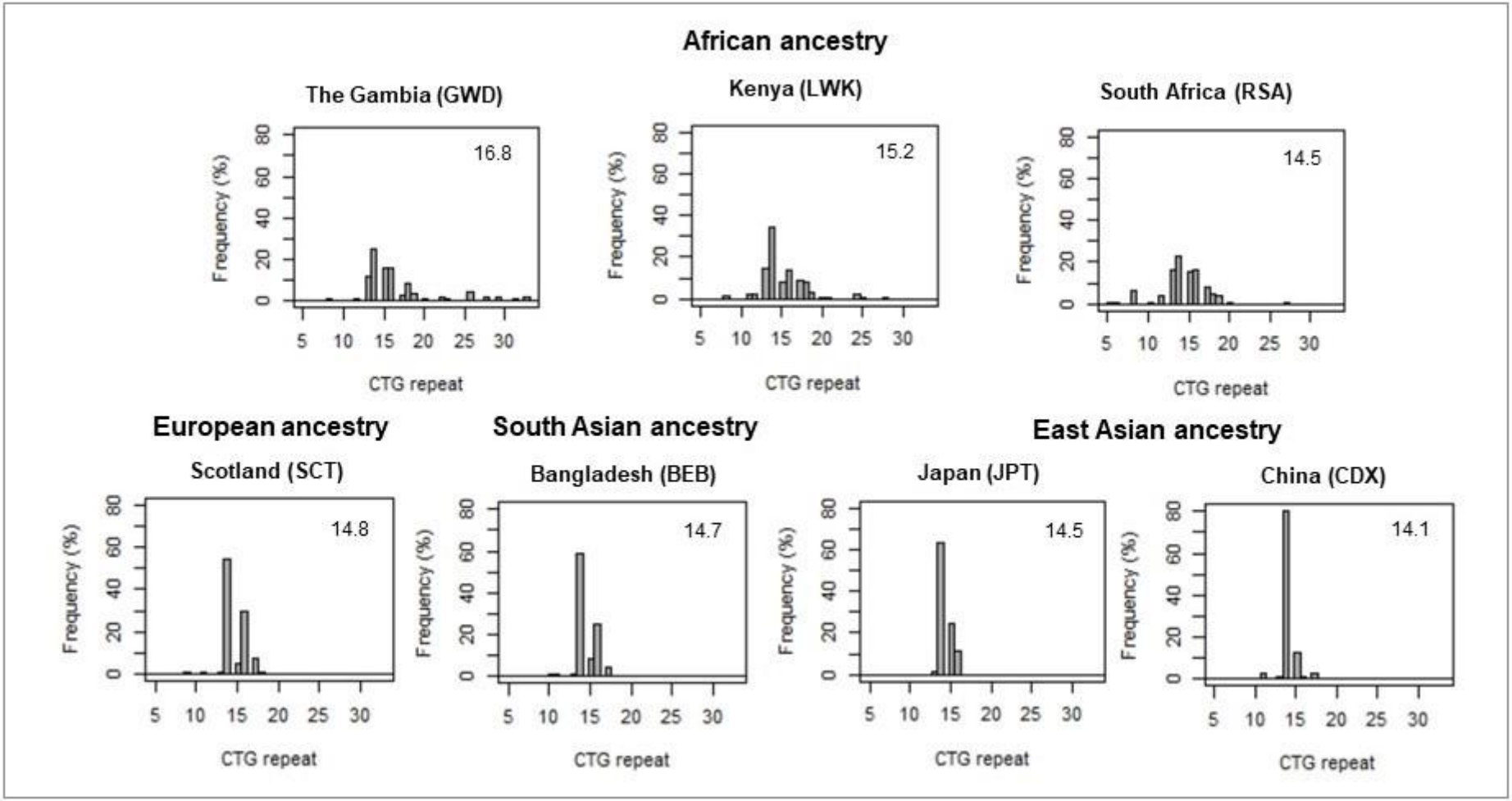


Figure 4.2. The CTG repeat distributions and estimated mean repeat size in the non-disease alleles from various Africa, European and Asian populations.

4.1.2. *JPH3* haplogroup/haplotype diversity

Like the sequence diversity analysis, a limited amount of haplogroup/haplotype diversity was identified at the JPH3 locus. All the disease alleles occurred on the same haplogroup with the only variation identified when the larger haplotypes were assessed.

The *JPH3* locus was further investigated in the context of haplogroup and haplotype background that occurred along the gene spanning ~ 664,470 kb of chromosome 16. The haplogroup/haplotype background was constructed using seven tag-SNPs, three previously studied (Krause et al., 2015) and an additional four identified for this study. Unlike the *HTT* analysis, the *JPH3* haplotypes have not been extensively characterised with only the haplogroups previously defined and therefore there was no standard nomenclature for the haplotypes or haplotypes that could be used as a reference. This added to the complexity in the assignment of haplogroups and haplotypes, thus for the *JPH3* analysis the haplogroups and haplotypes were assessed in stages through the inclusion of certain tag-SNPs for determination of the most informative selection of tag-SNPs for the analysis.

The haplogroup/haplotype background was constructed for 495 samples from 32 African ancestry individuals affected with HDL2 and 463 unaffected individuals from various populations. This resulted in the construction of haplogroups/haplotypes for a total of 32 disease alleles and 958 non-disease alleles. The tag-SNPs for the individuals from various populations studied in the 1,000 Genomes Project were not genotyped as part of this study. The tag-SNPs genotypes were accessed from a publicly available platform and then underwent haplotype phasing. A total of 91 unaffected individuals included in the sequence diversity analysis were excluded from the haplogroup/haplotype diversity analysis (South Africa: 2; Scotland: 85, Kenya: 3 and The Gambia: 1). These individuals were excluded due to either the unsuccessful genotyping of the South African individuals or the tag-SNP results not being published for the 1,000 Genomes Project individuals. The Generation Scotland individuals were excluded in this analysis as the access to the samples was only limited to the sequence diversity analysis.

The haplogroup background were first analysed using the three previously identified tag-SNPs (JPH3 3: rs2544635, JPH3 4: rs3849254 and JPH3 6: rs918368) (Krause et al., 2015). The summary of the haplogroups identified in the disease and non-disease alleles is shown in Table 4.3 (page 89). In the table the haplogroups were given unique colour codes that were extended to the haplotype analysis and allowed for the easier distinction of different haplogroups. The haplogroup/haplotype nomenclature used in *JPH3* analysis was based on a

previous haplotype study (Krause et al., 2015), that named haplogroups/haplotypes based on the genotype composition of the three-core tag-SNPs.

The results show that all the disease alleles occurred on the same haplogroup, haplogroup 1: CAC (32/32, 100%). For the non-disease alleles, seven different haplogroups were identified at varying frequencies. The highest frequency was identified for haplogroup 1: CAC (379/958, 39.6%), while the lowest was identified for haplogroup 7: AGC (1/958, 0.1%). Despite an increased diversity identified in the non-disease alleles, the largest proportion of alleles also occurred on haplogroup 1: CAC. Looking specifically at the five different atypical allele structures identified in the repeat tract sequence diversity analysis, these non-disease allele structures occurred on three of the haplogroups. Atypical 1 occurred on haplogroup 4: CGT (1/1, 100%), Atypical 2 occurred on haplogroup 2: CGC (1/1, 100%), Atypical 3 occurred on haplogroup 3: CAT (1/1, 100%), Atypical 4 occurred on haplogroup 2: CGC (6/7, 85.7%) and haplogroup 4: CGT (1/7, 14.3%), while Atypical 5 occurred on haplogroup 2: CGC (65/71, 91.5%), haplogroup 4: CGT (5/71, 7.0%) and haplogroup 3: CAT (1/71, 1.4%).

Table 4.3. A summary of the *JPH3* haplogroups in the non-disease and disease alleles.

Haplogroup	Non-disease			Disease		
	CTG	n	%	CTG	n	%
1: CAC	**11 – 33**	**379**	**39.6**	**42 – 53**	**32**	**100.0**
2: CGC	6 – 26	236	24.6	–	–	–
3: CAT	6 – 29	212	22.1	–	–	–
4: CGT	12 – 33	45	4.7	–	–	–
5: AAC	13 – 31	54	5.6	–	–	–
6: AAT	8 – 20	31	3.2	–	–	–
7: AGC	17	1	0.1	–	–	–
Total	958			32		

The haplogroups were uniquely colour coded to distinguish the haplogroups from one another. The most common non-disease haplogroup is in **bold green font** and the most common disease haplogroup is in **bold red font.**

The larger haplotypes were then constructed using all seven of the genotyped tag-SNPs located in and around the *JPH3* locus. When all the tag-SNPs were considered additional haplotype diversity was identified with a total of 83 different larger haplotypes as shown in Appendix 8. All 83 larger haplotypes were identified in the non-disease alleles while only four were identified in the disease alleles. The nomenclature used in the haplogroups were extended to the haplotypes with the only distinction being the genotypes that represent the haplogroups bolded in the name. In the disease associated alleles, the largest proportion of alleles occurred on the larger haplotype 6: CC**CA**T**C**G (10/32 = 31.31%). Whereas in the non-disease alleles the largest proportion of alleles occurred on the larger haplotype 12: CT**CA**T**C**G (52/958= 5.4%).

To assess if the increased diversity identified in the larger haplotypes was a result of the inclusion of less informative tag-SNPs the haplotype background was then assessed without the two-outer tag-SNPs (JPH3 1: rs3849253 and JPH3 7: rs2303771). The summary of the haplotype diversity identified in the disease and non-disease alleles was shown in Table 4.4. The results of the analysis revealed that all the disease alleles occurred exclusively on haplotype 2: C**CATC** (32/32, 100.0%), that occurred on haplogroup 1: CAC. In the non-disease alleles, a greater amount of haplotype diversity was identified with 25 different haplotypes identified across the seven haplogroups. However, the largest proportion of non-disease alleles occurred on haplotype 4: T**CATC** (147/958, 15.3%), that also occurred on haplogroup 1: CAC. Although haplotype 2: C**CATC** was identified as the disease haplotype in all the individuals affected with HDL2, it was also present in the non-disease alleles, but was statistically more frequent in the disease alleles (83/958 < 32/32: p <2.2 x 10^{-16}).

Table 4.4. A summary of the *JPH3* haplotypes in the non-disease and disease alleles.

Haplogroup	Haplotype	Non-disease			Disease		
		CTG	n	%	CTG	n	%
1: CAC	1: C**CACC**	8 – 22	58	6.1	–	–	–
	2: C**CATC**	8 – 33	83	8.7	**42–53**	**32**	**100**
	3: T**CACC**	11 – 27	91	9.5	–	–	–
	4: T**CATC**	**8 – 26**	**147**	**15.3**	–	–	–
2: CAT	5: C**CATT**	8 – 26	41	4.3	–	–	–
	6: T**CATT**	8 – 28	137	14.3	–	–	–
	7: T**CACT**	14 – 18	28	2.9	–	–	–
	8: C**CACT**	14 – 29	7	0.7	–	–	–
3: CGC	9: C**CGCC**	8 – 22	41	4.3	–	–	–
	10: C**CGTC**	8 – 33	106	11.1	–	–	–
	11: T**CGCC**	11 – 27	21	2.2	–	–	–
	12: T**CGTC**	8 – 26	67	7.0	–	–	–
4: CGT	13: C**CGCT**	12 – 33	9	0.9	–	–	–
	14: C**CGTT**	13 – 16	15	1.6	–	–	–
	15: T**CGTT**	12 – 18	20	2.1	–	–	–
	16: T**CGCT**	15	1	0.1	–	–	–
5: AAC	17: C**AATC**	14 – 31	5	0.5	–	–	–
	18: T**AACC**	13 – 28	18	1.9	–	–	–
	19: T**AATC**	14 – 17	30	3.1	–	–	–
	20: C**AACC**	16	1	0.1	–	–	–
6: AAT	21: T**AATT**	14 – 17	25	2.6	–	–	–
	22: C**AACT**	16	1	0.1	–	–	–
	23: C**AATT**	14 – 20	3	0.3	–	–	–
	24: T**AACT**	8 – 16	2	0.2	–	–	–
7: AGC	25: T**AGCC**	17	1	0.1	–	–	–
Total			958			32	

The haplogroups were uniquely colour coded to distinguish the haplogroups within which each of the haplotypes fall. The most common non-disease haplotype is indicated in **bold green font** and the most common disease haplogroup is indicated in **bold red font**. The CTG repeat ranges, number and percentages were shown for each of the haplotypes.

The breakdown of the non-disease haplotypes for each of the populations revealed differences in the occurrence that suggest population specific haplotypes. The breakdown of the haplotypes across the various populations is shown in Table 4.5 (page 92). The majority of the 25 haplotypes identified occurred in populations with African ancestry, particularly in the Gambian individuals (21/25, 84.0%), where increased diversity was identified. Over half of the haplotypes (13/25, 52.0%) were only identified in the African ancestry individuals. In the South African and Kenyan individuals, the most common non-disease haplotype was haplotype 10: C**C**G**T**C. In the Gambian individuals it was haplotype 2: C**CA**T**C**. While in the Asian individuals the haplotype 4: T**CA**T**C** was the most common.

Larger non-disease CTG repeats (27-35), which would be considered in an intermediate range in HD were identified on disease associated haplotype, haplotype 2: C**CA**T**C**. This would be expected as this was the only disease associated haplotype and thus the CTG expansion into the pathogenic range could be expected. However, there were also larger non-disease CTG repeats on another seven haplotypes, which could potentially expand into the pathological range. The majority of these larger CTG repeats were however identified in the Gambian non-disease alleles. No Gambian HDL2 disease alleles have been identified or studied.

The findings of the haplotype analysis suggest that the *JPH3* haplotype background may be better constructed without the inclusion of the outer two tag-SNPs, as they were shown to be less informative especially when assessed in the non-disease alleles. An increased amount of recombinant larger haplotypes were identified when compared to the haplotype constructed with only the five tag-SNPs. In the disease alleles the assessment of the haplotype constructed with only the five tag-SNPs were shown to be in linkage disequilibrium, suggesting a single origin for the mutation in the South African HDL2 patients which appears to be African specific.

Table 4.5. A summary of the *JPH3* haplotypes in the non-disease alleles for each population.

Haplogroup	Haplotype	CTG	RSA	GWD	LWK	BEB	JPT	CDX	Total
1: CAC	1: C**CA**C**C**	8 – 22	33	12	11	2	-	-	58
	2: C**CA**T**C**	8 – **33**	50	**14**	19	-	-	-	83
	3: T**CA**C**C**	11 – **27**	12	5	10	32	20	12	91
	4: T**CA**T**C**	8 – 26	17	10	15	**35**	**34**	**36**	147
2: CAT	5: C**CA**T**T**	8 – 26	23	9	9	-	-	-	41
	6: T**CA**T**T**	8 – **28**	59	7	19	26	11	15	137
	7: T**CA**C**T**	14 – 18	1	3	1	9	4	10	28
	8: C**CA**C**T**	14 – **29**	-	5	2	-	-	-	7
3: CGC	9: C**CG**C**C**	8 – 22	29	6	6	-	-	-	41
	10: C**CG**T**C**	8 – **33**	**66**	9	**31**	-	-	-	106
	11: T**CG**C**C**	11 – **27**	17	1	1	-	1	1	21
	12: T**CG**T**C**	8 – 26	57	3	5	2	-	-	67
4: CGT	13: C**CG**C**T**	12 – **33**	5	3	1	-	-	-	9
	14: C**CG**T**T**	13 – 16	3	8	4	-	-	-	15
	15: T**CG**T**T**	12 – 18	13	-	3	2	-	2	20
	16: T**CG**C**T**	15	-	-	-	-	-	1	1
5: AAC	17: C**AA**T**C**	14 – **31**	1	3	1	-	-	-	5
	18: T**AA**C**C**	13 – **28**	11	-	2	3	2	-	18
	19: T**AA**T**C**	14 – 17	4	2	1	20	1	2	30
	20: C**AA**C**C**	16	-	1	-	-	-	-	1
6: AAT	21: T**AA**T**T**	14 – 17	2	1	1	15	5	1	25
	22: C**AA**C**T**	16	-	1	-	-	-	-	1
	23: C**AA**T**T**	14 – 20	-	1	2	-	-	-	3
	24: T**AA**C**T**	8 – 16	-	2	-	-	-	-	2
7: AGC	25: T**AG**C**C**	17	1	-	-	-	-	-	1
Total			404	106	144	146	78	80	985

Population key: RSA: South Africa, GWD: The Gambia, LWK: Kenya, BEB: Bangladesh, JPT: Japan and CDX: China. The most common non-disease haplotype for each population is indicated in **bold font**. For the CTG repeat ranges, the larger non-disease CTG repeats in the uncertain range are indicated in **red bold font**.

4.1.3. Potential modifiers of the HDL2 phenotype

4.1.3.1. HDL2 phenotype assessment

The age of onset was the most appropriate measure of the HDL2 phenotype.

The final analysis of the *JPH3* locus assessed potential genetic modifiers of the HDL2 phenotype. Unlike the sequence and haplogroup/haplotype diversity analyses, potential modifiers were only assessed in the individuals affected with HDL2. Prior to assessing the effect of the potential modifiers, the age of diagnosis and age of onset were assessed to confirm the most appropriate measure of the HDL2 phenotype. Unlike the HD patients, age of onset information was available for the majority of the HDL2 patients and thus could be assessed directly without the AoD used as a proxy. The AoO of motor symptoms is the most well characterised measure of disease in HD and can also be transferred to HDL2 (Lee et al., 2012). The graphical representation of the linear regression analysis for the log transformed AoO and the CTG repeat shown in Figure 4.3 (page 94), revealed a significant association ($p = 3.45 \times 10^{-3}$) with an R-square value of 0.39. The CTG accounted for approximately 39% of the variation in the age of onset of HDL2. This association was considerably lower and less significant than the comparative result in the HD phenotype assessment where a highly significant association ($p = 2.55 \times 10^{-14}$) with an R-square value of 0.59 was identified.

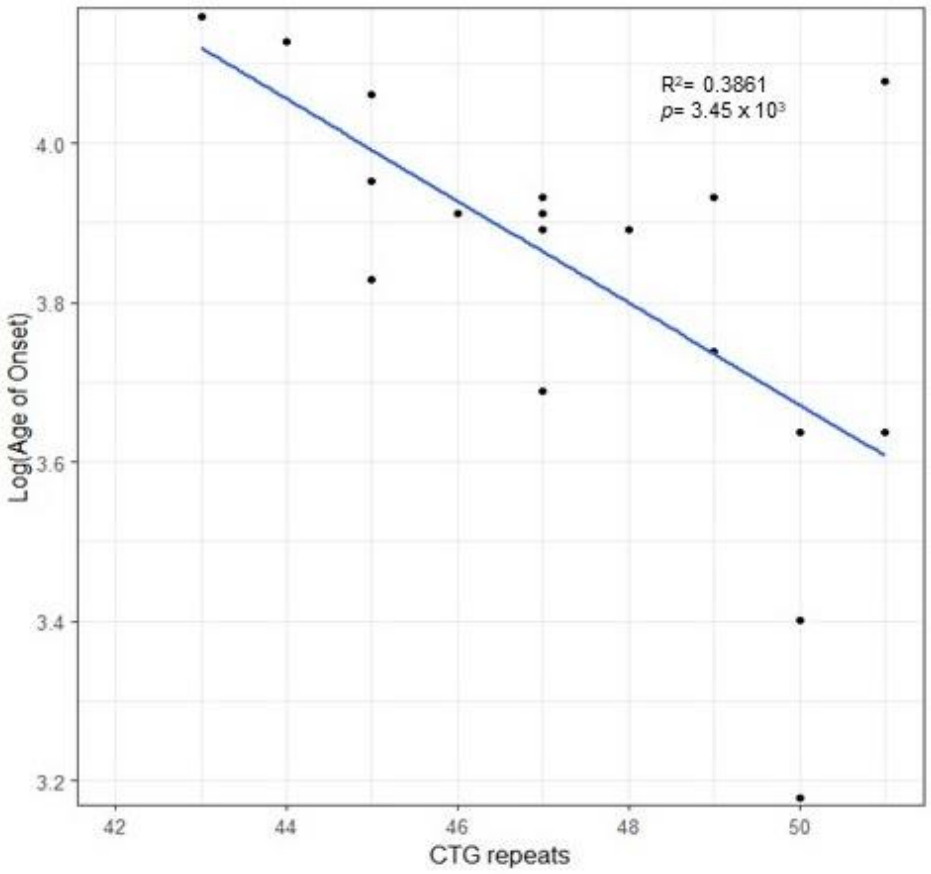


Figure 4.3. Linear regression analysis testing the association between the disease associated CTG repeat and the log transformed AoO.

The R-square and *p*-values show a significant association ($r^2 = 0.39$ and $p = 3.861 \times 10^{-3}$). The CTG repeat does not account for the majority of the variation in the HDL2 phenotype.

Due to the identification of a considerably reduced AoO association when compared to the HD analysis, the AoD was then assessed. To determine if the AoD was a more appropriate measure of the HDL2 phenotype. The graphical representation of the linear regression shown in Figure 4.4 (page 95), revealed a significant association ($p = 6.73 \times 10^{-3}$) with an R-square value of 0.23. The CTG repeats thus only accounted for approximately 23% of the variation in the age of diagnosis of HDL2, that was an even greater reduction in the association. These results confirm that the AoO was the most appropriate measure of the HDL2 phenotype that was used in the assessment of potential modifiers.

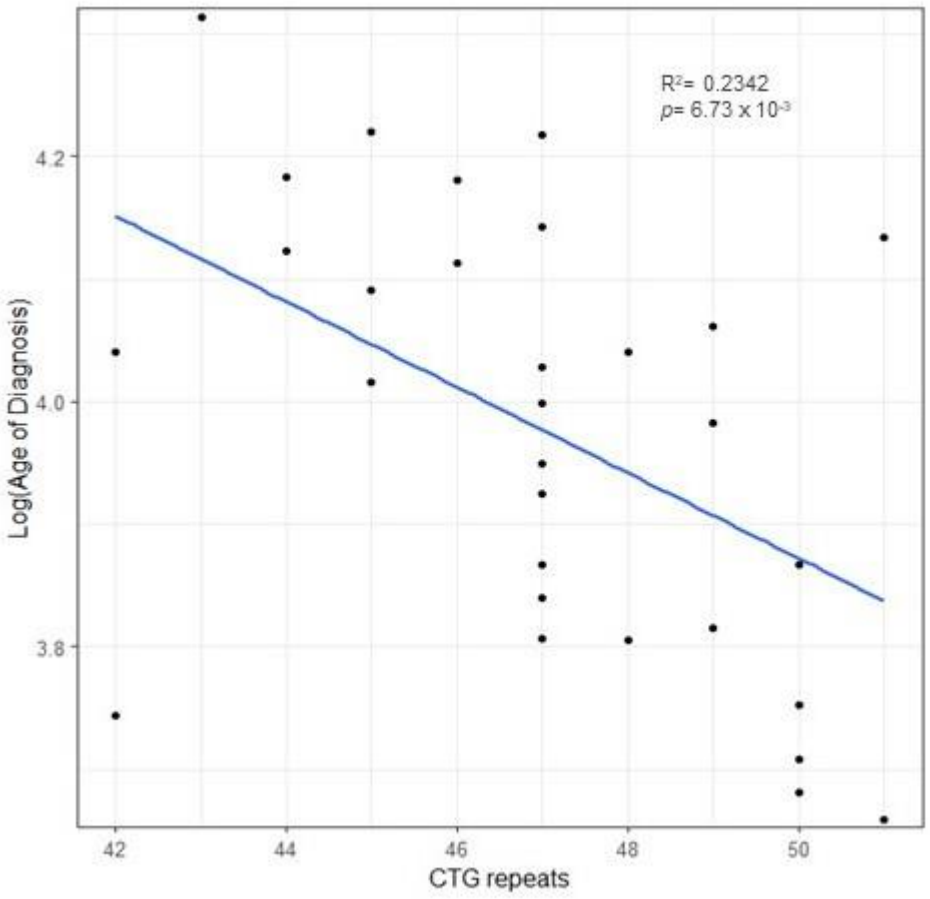


Figure 4.4. Linear regression analysis testing the association between the disease associated CTG repeat and the between log transformed AoD.

The R-square and *p*-values show a significant association (r^2 = 0.23 and p = 6.73 x 10^{-3}). The CTG repeat has a reduced association with the AoD than the AoO.

4.1.3.2. *JPH3* repeat tract modification

The CTG repeat was significantly associated with an earlier AoO and each repeat increase resulted in an earlier onset by 2.9 years.

The first potential modifier of the HDL2 phenotype (AoO) that was assessed was the effect of the *JPH3* repeat tract. Unlike *HTT* repeat tract assessment, this analysis only considered the effect of the CTG repeats as all disease alleles had the same typical allele structure with the only variation identified in the size of the CTG repeat. The CTG repeat was assessed through the construction of a linear regression analysis that tested the association with the AoO and linear models that revealed the effect in years. The linear regression analysis testing the association between the CTG repeat and the AoO was shown in Figure 4.3 (page 94). The results revealed that although a significant association was identified between the repeat and the age of onset, the CTG repeat accounted for less of the variation in the AoO (R-square value: 0.39), when compared with the CAG repeat in HD (R-square value: 0.59). Despite the difference identified between the two diseases, these HDL2 findings were similar to a previous study that also identified a reduced association (R-square value: 0.41) (Krause et al., 2015).

To determine the effect in years that the *JPH3* repeat tract had on the AoO, two linear models were constructed. In the model in Table 4.6, only the disease associated CTG repeat was assessed as an explanatory variable while the model in Table 4.7 (page 97) also assessed the non-disease associated CTG repeat and the sex of the HDL2 patients as explanatory variables. This differed from the *HTT* analysis as no other sequence component or atypical alleles structures were associated with the *JPH3* disease alleles thus other explanatory variables were explored. The results of the model shown in Table 4.6, revealed that every CTG repeat increase resulted in earlier AoO by 2.9 years (p = 3.45 x 10^{-3}). This CTG repeat effect was the same effect in years identified for the CAG repeat (2.9 years, p = 6.11 x 10^{-12}) in the *HTT* analysis.

Table 4.6. Linear model testing the effect of the disease associated CTG repeat on the AoO.

Model	r^2	*p*-value for model		Parameter values		
			Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
Ln (AoO)~ CTG (disease associated)	0.3861	3.45 x10^{-3}	30	**CTG (disease associated)**	**-2.948**	**3.45x10^{-3}**

The disease associated CTG repeat was assessed to determine the effect of each on the AoO. The R-square and *p*-values of the overall model show a significant association (r^2 = 0.39 and p = 3.45 x 10^{-3}). The disease associated CTG repeat was significant and indicated in **bold font**.

The results of the model shown in Table 4.7, revealed that despite the inclusion of additional explanatory variables, only the disease associated CTG repeat had a significant effect on the AoO (p = 5.59 x 10^{-3}). In line with the findings of the previous model, the CTG repeats were confirmed to result in a 2.9-year effect on the AoO. The assessment of the overall model revealed a reduction in the significance of the model despite a small increase in the R-square value (0.40) when compared with the R-square value (0.39) in the model in Table 4.6 (page 96). Thus, the non-disease associated CTG repeats and the sex of the HLD2 patients were not shown to have a significant effect on the AoO. Although the disease associated CTG repeat had a significant effect, comparable between the two diseases, it accounted for less of the variation and appeared not to be the only major driver of the disease phenotype.

Table 4.7. Linear model testing the effect of the disease and non-disease associated CTG repeats and the sex of the HDL2 patients on the AoO.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
Ln (AoO)~ CTG (disease associated) + CTG (non-disease associated) + Sex	0.4047	3.61 x10^{-2}	30	**CTG (disease associated)**	**-2.931**	**5.59x10^{-3}**
				CTG (non-disease associated)	0.408	4.93x10^{-1}
				Sex	-0.359	9.37 x10^{-1}

The disease and non-disease associated CTG repeat and the sex of the HDL2 patients were assessed to determine the effect on the AoO. The R-square and *p*-values of the overall model show a significant association (r^2 = 0.40 and p = 3.61 x 10^{-2}). The disease associated CTG repeat was significant and indicated in **bold font**.

4.1.3.3. *JPH3* haplogroup/haplotype modification

Patients with disease alleles that occurred on the larger haplotype 6 had a significantly earlier AoO.

The second potential modifier of the HDL2 phenotype (AoO) that was assessed was the effect of the background larger haplotype. The assessment of the background larger haplotype was conducted through the construction of a linear model that revealed the effect in years. Unlike the *HTT* analysis, only the larger haplotypes were assessed as the haplogroups and haplotypes (assessed with 5 tag-SNPs) were the same across all HDL2 patients. The only difference among the individuals affected with HDL2 was seen in the larger haplotypes that included the two-outer tag-SNPs.

In the linear model, the larger haplotypes and the CTG repeat were assessed as explanatory variables relative to a reference, larger haplotype 6: CC**CA**T**C**G. The results of the linear model shown in Table 4.8, revealed a significant correlation ($p = 1.479 \times 10^{-2}$) and an R-square value of 0.54. Although the inclusion of the larger haplotypes as an explanatory variable, increased the overall variation in the AoO, a reduction in the significance of the model was identified. The results revealed that the CTG repeats and the larger haplotype 8: CC**CA**T**C**A had a significant effect on the AoO when compared with the reference, larger haplotype 6: CC**CA**T**C**G. The effect in years revealed that every CTG repeat increase resulted in earlier AoO by 3.4 years ($p = 1.49 \times 10^{-3}$), an increase from the 2.9-year effect identified across the previous linear models. In the *HTT* analysis, the inclusion of the haplotypes also resulted in an increase in the effect from 2.9 to 3.1 years. The larger haplotype 6: CC**CA**T**C**G, resulted in an earlier AoO when compared with haplotype 8: CC**CA**T**C**A by 17.2 years ($p = 4.15 \times 10^{-2}$).

Table 4.8. Linear model testing the effect of the larger haplotypes on the AoO.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in years	*p*-value for explanatory variable
Ln (AoO)~ CTG (disease associated) + larger haplotypes	0.5406	1.479×10^{-2}	30	**CTG (disease associated)**	**-3.410**	**1.49×10^{-3}**
			9	**CCCATCA**	**17.194**	**4.15×10^{-2}**
			7	TCCATCA	5.641	3.48×10^{-1}
			9	TCCATCG	6.535	2.65×10^{-1}

The disease associated CTG repeat and the larger haplotype relative to a reference, larger haplotype 6: CC**CA**T**C**G were assessed to determine the effect on the AoO. The R-square and *p*-values of the overall model show a significant association ($r^2 = 0.54$ and $p = 1.479 \times 10^{-2}$). The disease associated CTG repeat and the larger haplotype 8: CC**CA**T**C**A were significant and indicated in **bold font**.

Although the linear model in Table 4.8 (page 98) showed the AoO effect in years for each of the larger haplotypes, the analysis did not provide information of the mean AoO of each of the larger haplotypes. The mean age of onset across the HDL2 patients was identified as 47.6 years (95% CI = 37.5 to 57.7). The estimated marginal mean AoO in years for each of the larger haplotypes was established as shown in Figure 4.5. The results confirmed that the larger haplotype 6: CC**CA**T**C**G had the earliest AoO at 41.2 years while the larger haplotype 8: CC**CA**T**C**A had the most delayed AoO at 56.1 years. These larger haplotypes differed only at the last tag-SNP. Thus, patients with the haplotype 6: CC**CA**T**C**G had the earliest mean AoO.

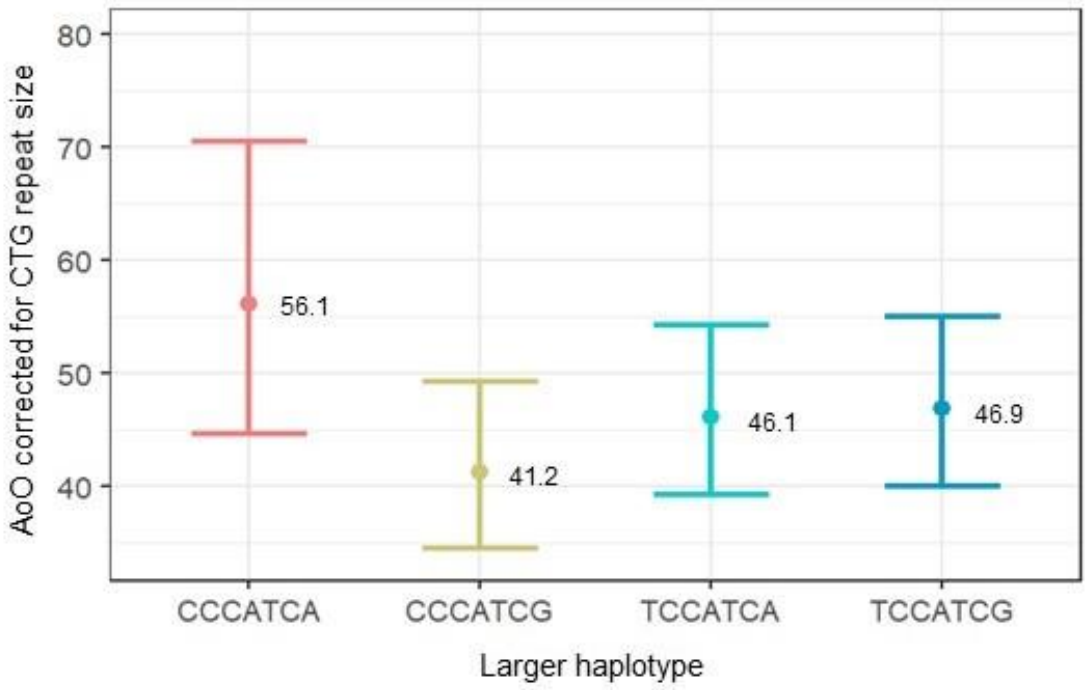


Figure 4.5. The estimated marginal mean AoO for each of the larger haplotypes, corrected for CTG repeat size.

The estimated marginal mean AoO for the larger haplotypes established using the emmeans function in R (*v*3.4.3). The earliest mean AoO was identified for the larger haplotype 6: CC**CA**T**C**G. The estimated marginal mean AoO for each of the larger haplotypes were as follows: larger haplotype 8: CC**CA**T**C**A (n=5, 56.1: 95% CI = 44.7 to 70.5), larger haplotype 6: CC**CA**T**C**G (n=10, 41.2: 95% CI = 34.5 to 49.3), larger haplotype 5: T**C**C**A**T**C**A (n=8, 46.1: 95% CI = 39.2 to 54.2) and larger haplotype 7: TC**CA**T**C**G (n=9, 46.9: 95% CI = 40.1 to 55.0).

4.1.3.4. CTG somatic expansion modification

Unlike the CAG repeats in HD, the CTG repeat does not appear to be the primary determinant of the ratio of somatic expansion.

The last modifier that was assessed for an effect on the AoO was the somatic expansion of the CTG repeat. In HD, somatic expansion significantly associated with an earlier AoO and poorer HD outcomes (Ciosi et al., 2019; Wright et al., 2019), however the scenario occurring in HDL2 remains unknown. Prior to the construction of the linear model that assessed the effect of somatic expansion on the AoO, the ratio of somatic expansion was assessed to establish an expansion score. The assessment of the ratio of somatic expansion was conducted through the construction of a linear regression analysis for the log transformed somatic expansion and CTG repeats. The results of the linear regression analysis shown in Figure 4.6 (page 101), demonstrated a significant association ($p < 2.2 \times 10^{-2}$) with an R-square value of 0.18. Thus, although there was a significant association identified, the CTG repeats only accounted for 18% of the variation in the ratio of somatic expansion. These findings revealed a considerable difference relative to the *HTT* CAG repeats, where a highly significant association was identified, and the repeat length accounted for 77% of the variation in somatic expansion. In this study, the CTG repeats do not appear to be the primary determinant of the ratio of somatic expansion.

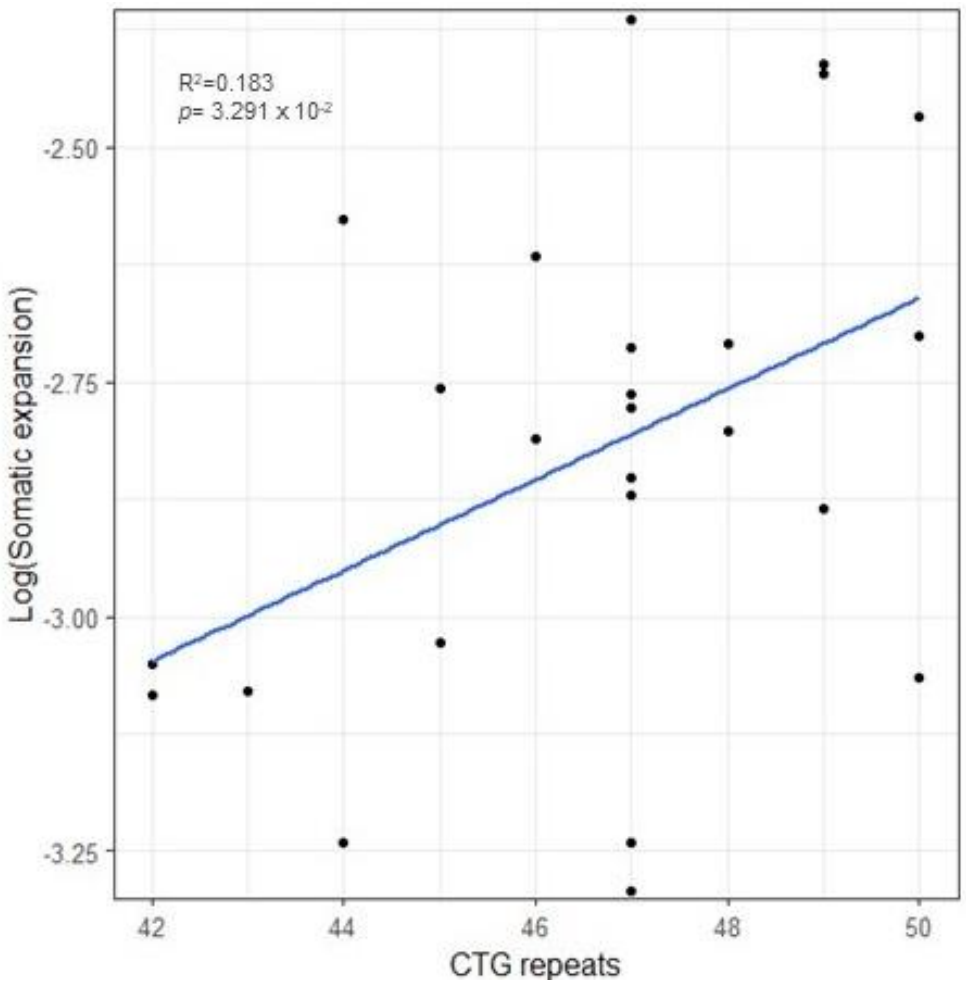


Figure 4.6. Linear regression analysis testing the association between CTG repeat and the ratio of somatic expansion.

The R-square and *p*-values show a significant association (r^2= 0.183 and $p = 3.3 \times 10^{23}$). The CTG repeat has a reduced association with the ratio of somatic expansion compared to the CAG repeat in HD patients.

To establish an expansion score, the effect on the ratio of somatic expansion was assessed through the inclusion of the age of sampling with the CTG repeat. In the linear model the explanatory variables were the CTG repeat, the age at sampling, and an age at sampling CTG repeat size interaction. The age of sampling was incorporated into the assessment due to the increase in instability of the repeat with increasing age, previously seen in HD and other repeat expansion disorders. The result of the linear model shown in Table 4.9 (page 102), did not however reveal the expected association ($p = 1.599 \times 10^{-1}$) with the ratio of somatic expansion. Thus, through the inclusion of the age at sampling in the assessment of the ratio of somatic expansion the effect was no longer significant, despite the CTG repeat explanatory variable still having a significant effect ($p = 2.47 \times 10^{-2}$).

Table 4.9. Linear model testing the effect of the CTG repeat and age at sampling on the ratio of somatic expansion.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Effect in ratio of SE	*p*-value for explanatory variable
Ln (SE) ~ CTG (disease associated) + age at sampling + CTG (disease associated) *age at sampling	0.216	$1.559x10^{-1}$	28	CTG (disease associated)	0.004	$2.74x10^{-2}$
				Age at sampling	0.0004	$3.72x10^{-1}$
				CTG (disease associated): Age at Sampling	0.0001	$6.33x10^{-1}$

The CTG repeat, age at sampling and the interaction between the two were assessed to determine the effect on the AoO. The *p*-values of the overall model did not show a significant association (r^2= 0.22 and $p = 1.559 \times 10^{-1}$).

Once the ratio of somatic expansion was assessed, the effect it had on the AoO was then considered. This was achieved through the construction of a linear model that included the effect of the expansion score. The expansion score was taken from the residuals of the ratio of somatic expansion linear model 1 in Table 4.9. In the linear model that assessed the effect on the AoO, the explanatory variables were the CTG repeat, expansion score and the larger haplotypes relative to a reference, the larger haplotype 6: CC**CA**T**C**G. The results of the linear model shown in Table 4.10, revealed a significant association ($p = 2.37 \times 10^{-2}$) that accounted for 68% of the variation in the AoO. However, the expansion score was not shown to have a significant effect ($p = 3.30 \times 10^{-1}$), indicating that there was no statically significant association between the ratio of somatic expansion and the age of onset in the HDL2 patients. The CTG repeats were the only variable that had a significant effect on the AoO, and every repeat increase resulted in an earlier onset by 3.4 years. This was a 0.5-year increase compared to the CTG repeat effect determined when only the CTG repeats were used in the assessment of the AoO (Table 4.6, page 96).

Table 4.10. Linear model testing the effect of the CTG repeat, larger haplotypes and the expansion score on the AoO.

Model	r^2	*p*-value for model	Parameter values			
			Sample size	Explanatory variable	Estimate in years	*p*-value for explanatory variable
Ln (AoO) ~ CTG (disease associated) + Larger Haplotype + Expansion score	0.6831	$2.37x10^{-2}$	28	**CTG (disease associated)**	**-3.364**	**$2.84x10^{-3}$**
				CC**CA**T**C**A	4.906	$4.29x10^{-1}$
				TC**CA**T**C**A	0.476	$9.19x10^{-1}$
				TC**CA**T**C**G	1.876	$6.95x10^{-1}$
				Expansion score	-6.861	$3.38x10^{-1}$

The CTG repeat, expansion score and the larger haplotypes relative to the larger haplotype 6: CC**CA**T**C**G were assessed to determine the effect on the AoO. The R-square *p*-values of the overall model how a significant association (r^2= 0.68 and $p = 2.37 \times 10^{-2}$). The CTG repeat had a significant effect and is indicated in **bold font.**

The CTG repeats showed considerably less somatic expansion than the CAG repeats of the same size.

The comparison of ratio of somatic expansion occurring in HD affected individuals versus the individuals affected with HDL2 revealed considerable differences between the two loci. The graphical representation of the comparison is shown in Figure 4.7. The ratio of somatic expansion of the *JPH3* repeats in the blood DNA of HDL2 patients was much lower than the ratio of somatic expansions of the *HTT* repeats in HD patients. Thus, at repeats of the same size the CAG repeats in the *HTT* repeat tract have an increase in somatic expansion when compared with the CTG repeats in the *JPH3* repeat tract.

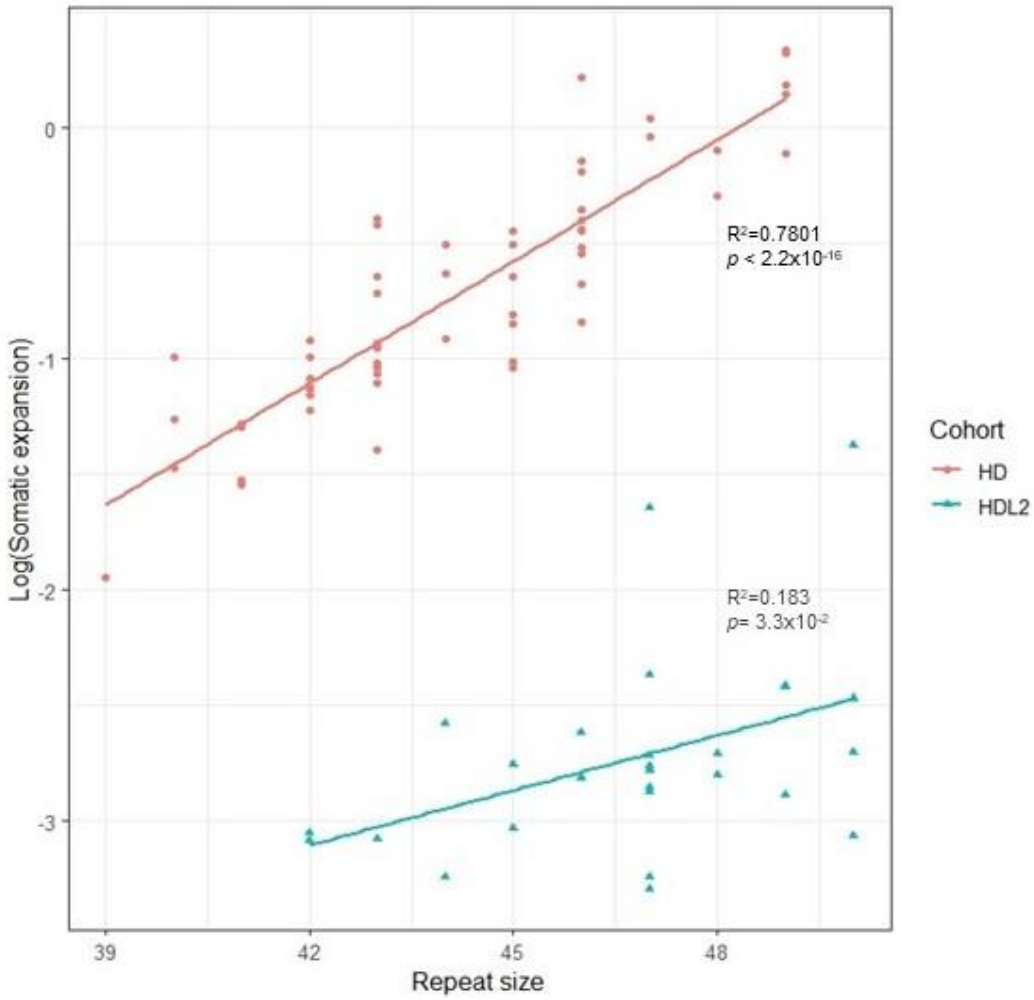


Figure 4.7. The comparison of the CAG/CTG repeat association with somatic expansion in the individuals affected with HD and HDL2.

The R-square and *p*-values for the CAG (r^2= 0.7802 and $p < 2.2 \times 10^{-16}$) and CTG (r^2= 0.183 and $p = 3.3 \times 10^{-2}$) repeat show a significant association. The CTG repeat have considerably less somatic expansion than the CAG repeats.

4.2. Discussion of *JPH3* analysis

The clinical phenotype of HDL2 is nearly indistinguishable from HD with the associated movement disorder, cognitive decline and psychiatric complications (Margolis et al., 2001; Walker et al., 2003). A recent clinical study comparing HD and HDL2 patients revealed that the two patient groups could not be distinguished on clinical examination and had similar neurodegeneration patterns noted on brain magnetic resonance imaging (Anderson et al., 2019a). The overall clinical similarity between HD and HDL2 suggests that a similar underlying molecular mechanism may be present. To investigate this the *JPH3* repeat tract sequence diversity, haplogroup/haplotype background diversity and potential modifiers of the HDL2 phenotype were assessed.

4.2.1. *JPH3* repeat tract sequence diversity

A limited amount of sequence diversity is identified in the *JPH3* repeat tract across the disease and non-disease alleles from various populations. Of the six different allele structures identified, one is the typical allele structure (variation only in the number of CTG repeat); while five are atypical allele structures (variation in the CTG repeat and/or flanking regions). Despite the five atypical allele structures only being identified in the non-disease alleles, the comparison of the allele types in the non-disease and disease alleles revealed no statistically significant frequency differences. In comparison to the *HTT* repeat tract, the *JPH3* repeat tract shows considerably reduced diversity. This is particularly true for the HDL2 disease alleles that all have the same typical allele structure and no sequence variation. These alleles are characteristic of the *JPH3* complied reference gene, which is comprised of a variable number of CTG repeats. The most common allele structure in the non-disease alleles is also the typical allele structure. Unlike the disease alleles, five atypical allele structures are also observed in the non-disease alleles.

These atypical allele structures are identified in the upstream and downstream flanking regions around the CTG repeat, as well as interrupting the CTG repeat itself. The variations in the flanking regions are included in this study as no previous study has assessed the sequence composition of the broader repeat region among disease and non-disease alleles. The atypical alleles only accounted for a small proportion (7.4%) of the non-disease alleles, with the most common, atypical 5 located in the downstream flanking region. Although the variants associated with all the atypical allele structures have been previously identified and submitted to dbSNP by the 1,000 Genomes Project, the only available information relates to the frequency, with no citations or clinical significance related to HDL2.

In addition to the South African, African ancestry non-disease alleles the inclusion of non-disease alleles from six other populations allowed for the *JPH3* repeat tract diversity to be established among non-disease alleles from various populations. These non-disease alleles have been assessed from presumed unaffected individuals from the populations studied in the 1,000 Genomes Project. This assessment contributes uniquely to expanding the knowledge of the *JPH3* repeat tract in individuals unaffected with HDL2 and provides population-specific sequence information for a locus that has not been previously assessed.

The comparison among populations shows differences in the occurrence of allele structures and CTG repeat size. In the African ancestry non-disease alleles, four of the five atypical allele structures (Atypical 1, 2, 4 and 5) are observed. While the only other atypical allele structure (Atypical 3) is observed in a single South Asian ancestry non-disease allele from Bangladesh. The other non-disease alleles from East Asian ancestry and European ancestry have no variation observed thus only have the typical allele structure. This highlights that repeat tract differences exist among various populations despite all being non-disease alleles.

Differences are identified in the mean CTG repeat size and repeat length distributions between the populations. The largest mean CTG repeat sizes are in two of the three African populations, from The Gambia (CTG: 16.8) and Kenya (CTG: 15.2). While the smallest mean repeat sizes are in the South Asian populations, from Japan (CTG: 14.5) and China (CTG: 14.1). Interestingly, the mean CTG repeat size from the South African (CTG: 14.5) population non-disease alleles are relatively lower than the other African populations. Looking at the repeat length distributions, the African populations have broader distributions compared to the European and Asian populations. Like the mean CTG repeat size, the Gambian non-disease alleles also have the widest repeat distribution, followed by the Kenyan and South African non-disease alleles. The largest and smallest CTG repeat sizes are also identified in these African populations, with CTG 33 from The Gambia and CTG 5 from South Africa. The repeat distributions observed among the other Asian and European non-disease alleles are relatively narrow.

In other repeat expansion diseases such as HD, the larger the repeat, the more unstable the repeat and thus the more prone to expansion into the disease range on vertical transmission (Goldberg et al., 1995; Semaka et al., 2006). In myotonic dystrophy type 1 (DM1), a wider CTG repeat size distribution is seen in individuals of European ancestry, while in Africa ancestry individuals there have been no reports of DM and there is a narrower repeat

distribution (Goldman, Ramsay and Jenkins, 1996). The larger mean repeat sizes and wider distributions observed in the African ancestry non-disease alleles may explain why HDL2 has only been identified in African or probable African ancestry patients, with the smaller CTG repeats common in the other populations being less likely to expand into the pathogenic range.

An unexpected finding of the sequence diversity analysis in the various populations is the high frequency of larger CTG repeats in the non-disease alleles of the Gambian individuals. In HD, the larger non-disease repeats fall into an intermediate range that are prone to expansion into the pathological range. Thus, if applied to HDL2, it is expected that there is potentially an increased predisposition to the occurrence of HDL2 in the Gambian population. To date, HDL2 has only been reported in two African ancestry populations, South Africa (Anderson et al., 2017; Baine, Krause and Greenberg, 2016; Krause et al., 2015) and Botswana (Ocampo, Daimari and Oyekunle, 2018). However, no studies have been published from The Gambia looking for patients with HDL2. A recent study conducted in over 50,000 participants revealed The Gambia was an important base for the West African slave trade (Micheletti et al., 2020), providing a possibility for the introduction of HDL2 that is seen in a large African-American pedigree (Pedigree W) from the South-eastern part of the United States (Margolis et al., 2001).

These findings suggest that there are more alleles with larger CTG repeats in African populations, that may form a pool of unstable alleles. The African alleles have also been around for longer and the population-specific findings show a subset of alleles have CTG repeats that are increasing on different haplotypes. None of these alleles appears to have migrated out of Africa into European populations which may provide an explanation for why HDL2 is not seen in non-African populations. Greater sequence diversity is also identified with four of the atypical alleles occurring exclusively in the African populations and these may thus be African specific.

This contrasts with the African disease alleles which appear to be free of any additional variation apart from the CTG repeat expansion. This is in stark contrast to the increased sequence diversity seen in the *HTT* gene, as most of the African disease alleles in the HD patients were of an atypical structure. Despite the clinical similarity seen among individuals affected with HD and HDL2, the comparison of the sequence diversity at the two loci shows considerable differences.

4.2.2. *JPH3* haplogroup/haplotype diversity

The construction of the *JPH3* haplogroups and haplotypes allowed for the largest assessment of the broader *JPH3* locus across disease and non-disease alleles. The haplogroup background constructed using three previously studied tag-SNPs (Krause et al., 2015), revealed seven different haplogroups. The disease alleles all occur on haplogroup 1: CAC, in keeping with the study by Krause et al. (2015), with only one disease allele identified from a patient with Moroccan ancestry not sharing the same disease haplogroup (22/23, 96%) (Krause et al., 2015). The haplogroup findings for the disease alleles are suggestive of a single origin of HDL2. In the non-disease alleles, all seven of the different haplogroups are identified. Despite the increase in diversity, the most common haplogroup is also haplogroup 1: CAC. The majority of the atypical allele structures (Atypical 2, 4 and 5) occur on haplogroup 4: CGC, suggesting that this haplogroup may be more prone to mutations than the other haplogroups.

The larger haplotypes constructed using all seven of the *JPH3* tag-SNPs show a large amount of haplotype diversity. A total of 83 different larger haplotypes are identified in the non-disease alleles while only four are observed in the disease alleles. In the disease alleles, the most common larger haplotype is haplotype 6: CC**CA**T**C**G, while in the non-disease alleles the most common larger haplotype is haplotype 12: CT**CA**T**C**G. These findings suggest that assessing the background *JPH3* haplotypes using all seven tag-SNPs may in fact be less informative especially in the non-disease alleles. To better study the haplotype background the two outmost SNPs are excluded from the assessment.

The haplotype background constructed using the five tag-SNPs show that like the haplogroup assessment, the disease alleles occur exclusively on haplotype 2: C**CA**T**C**. In the non-disease alleles, twenty-five haplotypes are observed. A considerable reduction in the diversity identified in larger haplotype assessment. Although the disease-associated haplotype 2: C**CA**T**C** is also present in the non-disease alleles, it is at a lower frequency. The comparison among the different populations show that this haplotype is only identified in the African ancestry individuals, suggesting that this haplotype is African specific. The most common haplotype in the non-disease alleles is haplotype 4: T**CA**T**C**, which is identified across all the populations. The inclusion of the two outermost tag-SNPs in the larger haplotype are less informative, in the non-disease alleles there is an increase in the recombination events. In the disease-associated alleles the haplotype composed of the five tag-SNPs are in linkage disequilibrium, suggesting a single origin for the mutation in the South African HDL2 patients which appears to be African specific.

Like the sequence diversity assessment, an increase in haplotype diversity is identified in the African ancestry non-disease alleles. The comparison between various populations suggests that there are African specific haplotypes, like haplotype 2: C**CA**T**C** that occurs only in African ancestry disease and non-disease alleles. Although larger non-disease CTG repeat sizes (in the range of uncertainty) are identified on this haplotype they are also present in other haplotypes. However, the other haplotypes are not associated with disease in South African HDL2 patients. At this stage, it cannot be excluded that these larger non-disease CTG repeat sizes occurring on the other haplotypes, that are most commonly present in the Gambian individuals, do expand into the pathogenic range. If expansions do occur on these haplotypes it would result in other disease-associated haplotypes in different African ancestry populations. Alternatively, if these haplotypes are not associated with disease in other African ancestry populations, these haplotypes may confer a protective effect thus preventing these larger CTG repeat sizes from expanding into the pathological range.

The comparison between the *HTT* haplotype diversity reveals considerable difference between the diseases. The various haplotypes identified across the African ancestry HD disease alleles contrast with the single disease haplotype identified in the HDL2 disease alleles. With diversity among the individuals affected with HDL2 only identified in the inclusion two of the outermost tag-SNPs to construct the larger haplotypes, which are less informative. The occurrence of haplotype 2: C**CA**T**C** exclusively in all the disease alleles is suggestive of a single origin of the HDL2 repeat expansion in South Africa.

4.2.3. Potential modifiers of HDL2

4.2.3.1. HDL2 phenotype assessment

Prior to assessing the potential modifiers of the HDL2 phenotype in the African ancestry individuals with HDL2, the most appropriate measure of the phenotype was established. The most characterised measure of the disease phenotype, used in HD studies is the onset of motor symptoms (Lee et al., 2012). Unlike the HD patients, age of onset information was available for the majority of the HDL2 patients. Despite this the age of diagnosis was also assessed to determine if the AoO was the most appropriate measure of the HDL2 phenotype. This assessment was carried out through testing the association of the CTG repeat, as the repeat size are the major driver of disease severity in repeat expansion disorders such as HD (Lee et al., 2012). The assessment revealed a higher association for the AoO when compared with AoD association. These results thus confirm that the AoO is the most appropriate measure of the HDL2 phenotype.

A considerably lower and less significant AoO association with CTG repeat length is however identified for the HDL2 patients when compared with the results of the AoD association with CAG repeat length in the HD patients. The reduced associations and results that are less significant, may however be attributed to the smaller number of HDL2 patients in the study (Morgan, 2017). Alternatively, it may be attributed to potential inaccuracy of age of onset data, especially if self-reported or from family member providing a clinical history rather than direct observation. Anecdotal evidence has shown that many patients present 5-7 years after reported onset.

Limitations of working on smaller sample sizes is the reduction of the statistical power of the study, outliers may have more of an effect to skew the results and associations could be over/underestimated (Button et al., 2013). Unfortunately, when studying diseases that are as rare as HDL2 the small sample size cannot be avoided and has thus led to less conclusive results when compared to the assessment of the potential modifiers of the HD phenotype. In the analysis of the potential modifiers of the HDL2 phenotype various components of the *JPH3* locus are assessed, to identify potential *cis*-acting modifiers. First focusing on the effect of the *JPH3* repeat tract, the haplogroup/haplotype background and lastly the somatic expansion of the CTG repeat.

4.2.3.2. *JPH3* repeat tract modification

In the assessment of the *JPH3* repeat tract, the CTG repeat is the only sequence component as no sequence variation is identified in the disease alleles. The assessment reveals a significant association for the CTG repeat, with the repeats accounting for 39% of the variation in the age of onset. Although this association is significant, the CTG repeat accounts for less of the variation in the AoO (R-square value: 0.39), in comparison to the CAG repeat association in HD (R-square value: 0.59). Despite differences identified between the two diseases, these AoO findings are in line with a previous study conducted on a similar cohort that showed the CTG repeats account for 41% of the variation in the AoD (Krause et al., 2015).

A recent clinical study by Anderson et al. (2019b) revealed comparable results between the association of the CAG repeat in HD (R-square value: 0.78) versus the CTG repeat HDL2 (R-square value: 0.74) and the age of onset (Anderson et al., 2019b). The increase in the association (R-squared value) may be attributed to the access to detailed clinical history with accurate age of diagnosis and onset information from clinical observation as well as the inclusion of larger disease-associated repeat sizes in the analysis. Research on HD has shown

that repeats greater than 53 violate linear model assumptions (Lee et al., 2012), thus larger repeats are excluded in this study.

The determination of the effect in years that the *JPH3* repeat tract had on the AoO, reveals every CTG repeat increase to result in an earlier AoO by 2.9 years. The non-disease associated CTG repeats and the sex of the HDL2 patients are also included to assess the effect on the AoO. However, these variables did not have a significant effect on the AoO, thus the only component within the *JPH3* repeat tract that is shown to modify the HDL2 phenotype is the disease associated CTG repeat. The effect that every repeat increase had on the disease phenotype in HDL2 and HD is the same (2.9 years). Although, the CTG repeat accounts for approximately 20% less of the variation in the HDL2 phenotype in comparison to variation attributed to the CAG repeat in the HD phenotype. This reduction may be due to the limited sample size in the HDL2 patients and the difference in disease phenotype assessed, as the age of onset is used in HDL2 and the age diagnosed is used for HD.

Alternatively, there may be another *cis/trans*-acting genetic modifier with a predominant effect on the HDL2 phenotype that is not seen in HD, despite the diseases being clinically indistinguishable. A different rate of disease progression has been previously identified between HD and HDL2 (Anderson et al., 2019b). The HDL2 patients presented three years earlier for diagnosis and clinical intervention from the time of disease onset when compared with the HD patients (Anderson et al., 2019b). This may support a different disease mechanism that causes more rapid HDL2 disease progression.

4.2.3.3. *JPH3* haplogroup/haplotype modification

The haplogroup/haplotype background is the second modifier assessed for an effect on the AoO. This broader component of the *JPH3* locus is assessed due to the association of certain haplotypes with disease and the hypothesis of haplotypes having *cis*-elements that predispose to repeat expansion (Warby et al., 2009; Warby et al., 2011). Unlike the *HTT* analysis, the disease associated haplogroups and haplotypes (assessed with 5 tag-SNPs) are the same across all HDL2 patients. Thus, only the larger haplotypes are assessed due to the difference among the individuals affected with HDL2.

The assessment of the effect of the haplotype background on the AoO reveals a significant association. Through the inclusion of the larger haplotypes as an explanatory variable in assessment of the AoO, there is an increase by 15% in the variation in the AoO and the effect size of the CTG repeat increases by 0.5 years, although the significance of the model is

reduced. This reduction in the significance may suggest that the R-square value is overestimated due to the effect of the small sample size (Morgan, 2017).

Despite this, the CTG repeats and the larger haplotype 6: CC**C**A**T**CG show a significant negative effect on the AoO in comparison to the other significant larger haplotype. Every CTG repeat increase is shown to result in an earlier AoO by 3.4 years and patients with the haplotype 6: CC**C**A**T**CG had a significantly earlier AoO by 17.2 years compared to the other larger haplotypes. Patients with alleles on haplotype 6: CC**C**A**T**CG have the earliest mean AoO of 41.2 years compared with haplotype 8: CC**C**A**T**CA that have the most delayed mean AoO of 56.1 years, although only identified in five HDL2 patients. Although the larger haplotype 6: CC**C**A**T**CG does appear modify the age of onset of the HDL2 patients it accounts for less of the variation than the CTG repeats. The mechanism of the effect of the larger haplotype 6: CC**C**A**T**CG remains unknown and appears to be quite distant as it is only seen when the larger haplotypes are assessed. This result is also of borderline significance ($p = 4.15x10^{-2}$) and will thus need to be verified using a larger sample size.

4.2.3.4. CTG somatic expansion modification

The CTG repeat size has a significant effect on the ratio of somatic expansion, although CTG repeats only account for 18% of the variation in the ratio of somatic expansion. This contrasts with the association of CAG repeats in the HD patients that account for 76% of the variation in the ratio of somatic expansion. Suggesting that although somatic expansion of the CTG repeats is occurring, the CTG repeat may not be the primary determinant that is driving somatic expansion. The comparison between the ratio of somatic expansion between the CTG and CAG repeats show that the HDL2 patients have significantly less somatic expansion than HD patients for the same expanded repeat size.

The effect of the ratio of somatic expansion is further assessed with the inclusion of the CTG repeat, the age at sampling and the interaction between the age at sampling and CTG repeat size. However, the results reveal no significant associations with the ratio of somatic expansion when the other two explanatory variables are considered. Despite this the residuals of the linear model were used to calculate the expansion score. The expansion score was then used to assess the effect of the ratio of somatic expansion on the AoO. The inclusion of the expansion score in assessment revealed that somatic expansion did not have a significant effect, indicating that at this stage there is no statically significant association between the ratio of somatic expansion and the age of onset in the HDL2 patients. Increasing the number of HDL2 patients may allow for an increase in the power of the statistical

analysis to reveal if these results are not biased by the small sample size and may show the CTG repeats are in fact less prone to somatic expansion and have less of an effect on HDL2 phenotype.

Unlike HD, the association of the CTG repeats and the ratio of somatic expansion does not appear to be significantly contributing to driving disease severity. Although, the effect of somatic expansion occurring at the location of neurodegeneration in the brain cannot be ruled out at this stage as only peripheral blood was assessed in this study. The ratio of somatic expansion in the blood may be an underrepresentation of the extent of expansion occurring in the brain of HDL2 patients (Kennedy et al., 2003; Telenius et al., 1994; Veitch et al., 2007). However, the assessment of the ratio of somatic expansion in African ancestry and European ancestry HD patients is also conducted in the peripheral blood of patients and reveals significant associations with the disease phenotype (Ciosi et al., 2019). Thus, a different mechanism of disease modification may be occurring in the HDL2 patients that is not driven by somatic expansion of the CTG repeats however follow up studies that are adequately powered will need to be conducted to assess if this is occurring.

The results of the *JPH3* analysis revealed less sequence diversity was observed across all individuals affected and unaffected with HDL2. The disease alleles all had the same typical allele structure, which occurred on a single disease haplotype 2: CCATC. The haplotype 2 was also only identified in African ancestry non-disease alleles lending support of an African specific single origin of the HDL2 mutation. The CTG repeat was confirmed to be the main driver of disease severity although when compared with the CAG repeat, it accounted for considerably less of the variation in the disease phenotype. When the broader haplotype background was considered, the larger haplotype 6: CCCATCG was associated with an earlier age of onset of disease by 17.2 years however the outer two tag-SNPs appear less informative, and its mechanism is uncertain. The CTG repeat showed considerably less somatic expansion than the CAG repeat and unlike HD, somatic expansion was not significantly associated with the age of onset in HDL2 although the sample size was a major limiting factor in the identification of significant correlations. These data suggest that there may be other *cis/trans*-acting modifiers yet to be identified that are specific to HDL2. The comparison of the *HTT* and *JPH3* loci indicated that despite the clinical similarities between HD and HDL2, stark contrasts were identified that suggest a different potential mechanism of disease modification may be occurring in the two diseases. With the only similarity seen in the association with the disease-associated repeats, making the clinical similarity more intriguing.

Section 4

Section 4 is made up of the concluding chapter of the study (Chapter 5), divided into four parts. The summary of findings (5.1), the limitations and unanswered questions (5.2), the future prospects (5.3) that have arisen from this research and the concluding remarks (5.4). Each of these were explored for both the *HTT* and *JPH3* analysis.

5. Chapter 5: Conclusion

This study addressed the unique aspects of the molecular basis of Huntington disease and Huntington disease-like 2 due to the access to African ancestry patients and using the expertise developed by the Glasgow laboratory. The individuals affected with HD and HDL2 were accessed through the Molecular diagnostic testing laboratory at the Division of Human Genetics, National Health Laboratory Service, and School of Pathology. The collaboration with the research unit at the Institute of Molecular, Cell and Systems Biology, University of Glasgow, that have performed extensive research focused on triplet repeat disorders, allowed for the high-throughput sequencing of the *HTT* and *JPH3* loci. This combination allowed for the novel assessment of the two Huntington disease loci in a previously understudied population and led to novel insights into disease mechanisms.

5.1. Summary of findings

5.1.1. *HTT* analysis

This is the first study to describe allele sequence diversity within the *HTT* repeat tract in African ancestry individuals affected and unaffected with HD. Substantial sequence diversity, evidenced by the presence of 17 different alleles structures and predominantly atypical alleles in the HD patients, is reported. Three novel allele structures were defined, which are either very rare in other populations or are African-specific. Typical allele structures are more frequent in the non-disease alleles, while the atypical allele structures are more frequently associated with the disease alleles. The most common non-disease allele structure (Q^1-2-2-7-2) is typical and has been reported as the most frequently occurring in European ancestry alleles (disease and non-disease). In the disease alleles, the most common allele structure (Q^1-2-0-9-2) is atypical, due to the loss of the CCGCCA proline intervening sequence.

The *HTT* repeat tract assessment of the effect on the disease phenotype, shows the Q^1-2-0-9-2 allele structure results in an earlier age of diagnosis of HD. This is a novel finding that contrasts with the European ancestry modifier (Ciosi et al., 2019; Wright et al., 2019). This disease allele structure occurs exclusively on haplotype B2, which is proposed as a ‘high-risk’ haplotype in African ancestry patients. A large amount of haplotype diversity is identified in the African ancestry individuals with differences in the most common disease-associated haplotype identified, when compared with European ancestry individuals, most likely the result of multiple HD mutation events.

Although haplotype B2 is present in European ancestry HD patients, it is very rare and differs from the African defined haplotype B2 by at least one tag-SNP, which is very rare in European individuals. The assessment of the African haplotype B2 in other worldwide populations show an increased frequency in African ancestry populations when compared with the considerably reduced frequencies across non-African populations. Although haplotype B2 is present in non-African populations, the occurrence may be a result of admixture between populations. The increased frequency in the African populations provides support for haplotype B2 being African-specific and of an African origin.

The presence of the European 'high-risk' haplogroup variants in two HD patients of African ancestry are identified. Although these 'high-risk' haplogroup variants were thought to be absent in African ancestry individuals (Baine et al., 2013; Warby et al., 2009), their presence may be attributed to low-level admixture between populations. Alternatively, these haplotype variants may have been enriched in European populations due to population bottlenecks as a result of migration events or genetic drift.

This is the first study to assess the ratio of somatic expansion of the CAG repeats in African ancestry individuals. Recent data in European ancestry HD patients have confirmed that somatic expansion of the *HTT* CAG repeat as a key driver of HD severity (Ciosi et al., 2019; Wright et al., 2019). The level of somatic expansion in blood DNA of HD patients is inversely associated with the age at onset, and directly correlates with disease progression.

Somatic expansion of the CAG repeat is significantly associated with the age of diagnosis of the HD patients of African ancestry, which parallels the age of onset association in HD patients of European ancestry. Interestingly, the typical allele structures appear to be somatically less stable compared with the atypical allele structure Q^1-2-0-9-2, characterised by loss of the CCGCCA proline intervening sequence. This disease allele structure has the earliest age of diagnosis and the lowest ratio of somatic expansion, suggesting that somatic expansion is not enhanced by the loss of the CCGCCA sequence in African ancestry patients. Instead, the results indicate that the CCGCCA sequence (intervening proline) does not mediate its modifier effect through somatic expansion.

The Q^1-2-0-9-2 allele structure characterised by the loss of the CCGCCA sequence that occurs on haplotype B2 is therefore proposed as a *cis*-acting modifier of HD in African ancestry patients with Huntington disease. It appears to modify disease by reducing the age of diagnosis, through a mechanism that is not driven by somatic instability. While the mechanism of disease

modification is currently unclear, it may act through the Q^1-2-0-9-2 allele structure or an unidentified *cis*/*trans*-acting modifier on haplotype B2 that is in linkage disequilibrium with the atypical structure. Further analyses are required to understand the modification on the age of diagnosis in these African ancestry individuals affected with HD.

This research contributes uniquely to the body of knowledge of the *HTT* repeat tract in individuals affected and unaffected with HD and gives population-specific sequence information for African ancestry individuals previously not assessed. This study provides novel data on an African population that can inform current research efforts into the precise molecular aetiology underlying the HD phenotype. This research also demonstrates the importance of studying African populations in general due to the increased diversity present (Choudhury et al., 2017), which would likely be missed when studying only non-African populations. The identification of a novel African modifier emphasizes the need to learn more about the underlying molecular pathogenesis of neurodegenerative diseases such as HD in Africa and the need to focus research on diverse populations, which may provide new insights and broaden current therapeutic approaches.

5.1.2. *JPH3* analysis

The *JPH3* locus has been poorly studied, with no previous research assessing the sequence diversity or potential modifiers of HDL2. This is the first study to describe the allele structures within the *JPH3* repeat tract of individuals affected and unaffected with HDL2. This assessment contributes novel sequence data on an African ancestry population that can be used to further research efforts into the precise molecular aetiology underlying the HDL2 phenotype. The HDL2 affected individuals in this study account for many of the known cases identified to date, thereby providing a comprehensive description of the sequence diversity in HDL2 patients. These patients are however all from South Africa and thus potentially do not capture all the diversity of this disease, although patients from South Africa account for ~60% of the cases described worldwide (Anderson et al., 2017; Baine, Krause and Greenberg, 2016).

This study has resulted in the development of a novel high-throughput sequencing workflow and data analysis pipeline to precisely assess the *JPH3* repeat tract. The detailed assessment of the *JPH3* locus has allowed for the identification of four additional tag-SNPs that have been used to delineate the haplotype background better.

This study also provides novel population information on the *JPH3* repeat tract sequence, including mean CTG repeat sizes and CTG repeat distributions. The inclusion of various presumed unaffected individuals reveals differences that appear to be population specific. The African ancestry populations, especially seen in the Gambian population have larger mean CTG repeats and wider repeat distributions that may be more likely to expand into the disease range. These findings may provide a possible explanation as to the reason HDL2 has only been reported in individuals of African or probable African ancestry as larger repeats are more unstable in the germline and thus prone to expansion into a pathological repeat size (Anderson et al., 2017). Smaller CTG repeats that are common in Asian and European ancestry individuals are likely to be more stable. This may be due to the presence of a *cis*-acting modifier such as a haplotype that has a protective effect on the repeat tract or may simply represent the loss of such larger alleles in the population bottleneck that was experienced by humans migrating out of Africa.

All *JPH3* disease alleles have the same typical allele structure and occur on a single disease haplotype. There is no sequence diversity in the disease alleles to account for variability in phenotype. The occurrence of haplotype 2: CCATC only in African ancestry individuals and exclusively on all the disease alleles is supportive of a single origin of the HDL2 repeat expansion in South Africa, and possibly Africa (Krause et al., 2015). Diversity within the *JPH3* locus on disease alleles is only identified when larger haplotypes (which include the seven tag-SNPs) are considered. These however proved less informative especially in the non-disease alleles where there is an increase in recombination and haplotype diversity.

Prior to the assessment of the ratio of somatic expansion of the CTG repeat, it was hypothesized that larger repeats would have more somatic expansion, due to the absence of intervening sequences within the *JPH3* repeat tract. In European ancestry individuals, intervening sequences within the *HTT* repeat tract act as stabilisers of the CAG repeat (Ciosi et al., 2019; Wright et al., 2019). In this study, although expanded CTG repeats are significantly associated with the ratio of somatic expansion, the comparison between the CTG and CAG repeats shows that the CTG repeats in HDL2 patients have less somatic expansion overall. The ratio of somatic expansion also did not have a significant effect on the age of onset of HDL2 patients which contrasts with the significant association seen in African and European HD patients (Ciosi et al., 2019; Wright et al., 2019).

The effect of somatic expansion occurring at the location of neurodegeneration in the brain cannot be ruled out as only the ratio of somatic expansion in peripheral blood has been assessed

in this study. The ratio of somatic expansion in the blood may be an underrepresentation of the extent of expansion occurring in the brain tissue of HDL2 patients (Kennedy et al., 2003; Telenius et al., 1994; Veitch et al., 2007). However, significant results are identified in the assessment of somatic expansion of HD patients' peripheral blood. Sampling error cannot be ruled out at this stage; although a different mechanism of disease modification may be occurring in the HDL2 patients that is not driven by the somatic expansion of the CTG repeats. A different disease mechanism may be supported by the different rate of disease progression that has previously been reported between HD and HDL2, with the disease progression of HDL2 patients occurring approximately three years faster (Anderson et al., 2019b).

The comparison with the *HTT* repeat tract and haplogroup/haplotype background, reveals considerable differences between the two loci and proposed disease mechanisms. The assessment of the potential modifiers of the HDL2 phenotype reveal a very different scenario occurring at the *JPH3* locus in comparison to the *HTT* locus. Of the components of the *JPH3* locus, the CTG repeat, drives most of the variation in age of onset of HDL2 patients. However, the association between the CTG repeat and the disease phenotype is reduced in comparison to the association of the CAG repeat in HD patient. Although the larger haplotype 6: CC**CA**T**C**G appears to modify the age of onset of the HDL2 patients it accounts for less of the variation. Its mechanism is however uncertain and not a result of the core tag-SNPs which are shared among all the larger haplotypes and may suggest a distant effect.

The comparison of the findings of the potential modifiers of the disease phenotypes indicates that despite the striking clinical similarities between HD and HDL2 (Anderson et al., 2019b; Margolis et al., 2001; Walker et al., 2003), there appear to be different mechanisms acting to modify phenotype, with the only similarity being the association with the repeat expansion. Although significant results were identified, the remaining genetic contributions to the variation in the age of onset for HDL2 patients remain unknown. This study has however highlighted differences seen in nearly all the components of the *HTT/JPH3* loci that contribute differently to disease – including repeat tract diversity, haplotype background and somatic instability. Thus, it is not unlikely that these differences extend to other *cis/trans*-acting modifiers yet to be identified that may be specific to HDL2.

5.2. Limitations and unanswered questions

5.2.1. *HTT* analysis

A limitation of this study is the relatively small size of the African ancestry HD patients sample. Despite this, the results were shown to be statistically significant. An attempt has been made to avoid the skewing of results due to outliers by removing patients with extremely large repeats and correcting for the repeat size in the assessment of the potential modifiers of HD. The small sample taken from a distinct geographical location in southern Africa, may also not be characteristic of all HD patients of African ancestry.

The lack of availability of age of onset information is another limitation of the study, to overcome this, the age of diagnosis was used as a proxy for the age of onset. Increasing the sample size of African ancestry HD patients with detailed phenotypic information would add valuable insights that could be used to directly compare African patients with studies conducted in European ancestry patients.

The number of tag-SNPs used to construct haplotypes in the study is also a limitation as only a subset was genotyped due to capacity constraints of the MassARRAY® System and funding constraints. Thirteen SNPs were selected from the larger list previously used to construct haplotypes of the *HTT* locus (Baine et al., 2013) and these mostly span the non-coding region of the *HTT* gene. Although the most common haplotype B2 was constructed from this subset tag-SNPs, the African-specific haplotype B2 could not be confirmed as the defining SNP, rs762855, was not genotyped. At this stage haplotype B2 cannot be ruled out as being the rare European-specific haplotype B2 that differs at this single SNP (Kay et al., 2018), despite being very unlikely.

The proposed *cis*-acting modifier, the Q^1-2-0-9-2 allele structure characterised by the loss of the CCGCCA sequence is in complete association with haplotype B2 in the disease alleles. However, at this stage the effect of the sequence differences versus a distant modifier on the AoD could not be separated out. Increasing the number of patients with the Q^1-2-0-9-2 disease-associated allele structure occurring on the haplotype B2 may allow for a statistically significant indication either way.

5.2.2. *JPH3* analysis

Like the *HTT* analysis, a major limitation of this study is the small sample size that is even smaller in the HDL2 patients. However, the rarity of HLD2 and its occurrence only in African ancestry individuals contributed to the small number of patients which is unavoidable in this

case. The clinical overlap with HD and other neurodegenerative diseases likely also contribute to the clinical misdiagnosis and fewer HDL2 patients being referred for HD phenocopy testing. Thus, fewer patients receive a molecular diagnosis unless the referring specialist is familiar with the Huntington disease phenocopy and specific genetic differences among ethnic groups. In African ancestry individuals or those with a distant African origin, a different diagnostic approach should be applied (Krause et al., 2015). Genetic testing is not widely available in Africa and thus many African cases of Huntington disease and Huntington disease-like 2 may not be diagnosed or molecularly confirmed.

The smaller sample sizes reduced the statistical power of the study. This can skew the result due to outliers that are not characteristic of the entire population. Associations may also be over- or under-estimated when conducting statistical analyses on small sample sizes (Button et al., 2013). Although significant results have been attained in the assessment of the effect of the repeat tract and haplotype background, the significance values for HDL2 are considerably lower compared to the results in HD. The ratio of somatic expansion does not appear to be associated with the age of onset however this may be due to the small sample size. Increasing the number of HDL2 patients will allow for the statistical power to be increased, to identify if some of the borderline significant results of this study have been biased by the small sample size and determine if somatic expansion does in fact modify HDL2.

Although the CTG repeat and the disease alleles occurring on the larger haplotype 6: CC**CA**T**C**G modify the HDL2 phenotype, there is a considerable amount of variation in the age of onset of HDL2 that cannot be fully explained. The differences identified in the comparison of the *HTT* and *JPH3* loci suggest that there may be unknown modifiers that are specific to HDL2. A limitation of all assessments of disease in both HD and HDL2 is the use of peripheral blood as a proxy for brain tissue, especially in the assessment of somatic expansion with the highest somatic instability noted in the brain (Kennedy et al., 2003; Shelbourne et al., 2007; Telenius et al., 1994).

5.3. Future prospects

5.3.1. *HTT* analysis

A prospect of this research is to increase the sample size of African ancestry HD patients along with detailed phenotype information, including age of onset. This will allow for confirmation of the findings in this study and potentially the identification of additional explanatory variables. Currently, the largest clinical research platform and observation study for HD is

Enroll-HD which has over 20,000 participants from 174 clinical sites (www.enroll-hd.net). The participants comprise HD patients with a molecular diagnosis, at-risk individuals with a family history of HD and unaffected individuals with a family history of HD. This provides a key resource for HD research and clinical trials as it collates patient biological samples and detailed clinical information. Enroll-HD has merged the Registry study in Europe and the COHORT study in North America to allow the standardisation of HD research. Participants from Latin America, Australia, New Zealand and some countries in Asia have also been incorporated. Although Enroll-HD has been set up as a worldwide study, no African sites or African HD patients are involved. This is a major omission, as individuals with African genetic background are not included.

When the findings of this research are considered, the increased genetic diversity and African specific modification of disease provides novel data in the context of HD. Further, from a clinical perspective, Huntington disease in Africa has not been systematically studied. It is considered to occur at a much lower frequency compared to European populations, however the prevalence is still in question due to the lack of sufficient population-wide data. The inclusion of African sites and African ancestry patients in a global HD study should thus be a major focus and could be implemented first in the major genetic centres in South Africa such as Johannesburg and Cape Town which are currently running HD genetics clinics. Alternatively, an African collaborative study between various genetics centres across the continent could be set up to increase the number of HD patients included in African led research and clinical trials. Research on African patients may yield novel insights into HD modifiers.

The DNA repair or maintenance genes identified as *trans*-acting modifiers of HD in European ancestry individuals in two GWAS studies should be explored in the African ancestry patients. Modifier loci that are directly or indirectly involved in DNA repair have been identified at *RRM2B/UBR5*, *FAN1*, *MSH3/DHFR*, *PMS1*, *PMS2*, *LIG1* and *MLH1* (Correia et al., 2015; GeM-HD Consortium, 2015; GeM-HD Consortium, 2019). The expression levels and polymorphisms found in these genes have significant effects on disease progression through the influence on somatic expansion of the CAG repeats in European ancestry individuals affected with HD. Genotyping polymorphisms in these DNA repair loci may identify associations with somatic expansion in the African HD patients. Although somatic expansion was found to have a significant association only with the disease-associated CAG repeats in our study, variation in the DNA repair genes may account for some of the remaining hereditary modifiers of disease, like the association seen in European ancestry patients. The reduced

somatic instability associated with the commonest *HTT* allele structure in African patients could be associated with a modifier mechanism. A GWAS approach could be applied to detect African specific variants which may then be correlated with expression levels to determine significant effects on disease progression. A major limiting factor will be an adequate number of African ancestry HD patients however, local and global attempts can be made to increase the pool of patients.

To allow for the differentiation of the haplotype B2 identified in this study as being either the African specific or the rare European haplotype B2, additional genotyping of the tag-SNP, rs762855 should be conducted on all the individuals affected and unaffected with HD that were identified as having the haplotype B2.

The sequence diversity analysis has identified novel genetic information that provides insight into the diversity present within the *HTT* disease alleles and identified the most characteristic disease allele structure. This genetic diversity can thus be used to inform current research efforts and clinical trials that are underway to test drugs to lower huntingtin protein and manage symptoms. Currently most *HTT* locus specific ASOs therapies are designed using European ancestry patients (Ionis Pharmaceuticals, 2017; Wave Life Sciences, 2019), and directed at SNPs or haplotypes common in European patients (Kay et al., 2019). However, these may prove ineffective when administered to the majority patients with African ancestry. The genome wide genetic diversity in African ancestry individuals also presents potential pharmacogenomic problems that would need to be examined/explored. Many variants that affect absorption, distribution, metabolism, and excretion of drugs have been identified (Grossman, 2009; Ramos et al., 2014). Drug development for HD should not exclude these populations but must rather be applied in a population specific manner so that all patients with disease benefit equally.

The analysis of methylation levels in HD patients is another new area of focus of the Glasgow laboratory that can be explored in African ancestry patients. In the *HTT* gene two cytosine phosphate guanine (CpG) clusters, CpG3 and CpG5 have been identified upstream and at downstream of the CAG repeats (Horvath et al., 2016), thus evaluating the comparison of the methylation levels between individuals affected and unaffected with HD may reveal disease associations. Previous analysis has identified large scale DNA methylation changes and accelerated epigenetic age in the brain tissue of HD patients (Horvath et al., 2016). This has recently been confirmed in an epigenome-wide association study (EWAS) that was conducted across three species and different tissue types (Lu et al., 2020). Findings from human blood

revealed that methylation levels at three loci were significantly associated with progression of motor symptoms in HD patients (Lu et al., 2020). Exploring the methylome in blood samples taken from African ancestry individuals may reveal a similar occurrence; alternatively, the comparison may further highlight the population specific differences this study has identified.

5.3.2. *JPH3* analysis

A prospect of this research would also include increasing the sample size of HDL2 patients and ensuring that detailed phenotype and age of onset information is obtained. This will allow for the confirmation of the findings in this study and potentially the identification of other factors modifying the age of onset. Like the scenario seen in the African ancestry HD patients, there is no Enroll-HD equivalent for HDL2 patients.

An African-led study, similar to what has been proposed in HD could be set up. This would however be unique given the relative absence of HDL2 in most other parts of the world. An African collaborative study from various genetic sites across the continent would increase the number of HDL2 patients. HDL2 patients identified in non-African countries could also be included. The research community studying HD phenocopies such as HDL2, is very small; this may provide a unique opportunity to establish a worldwide collaboration among groups studying HDL2. This will not only allow better information sharing to advancing HDL2 research, but also has the potential to increase the number of HDL2 patients available for study. This will increase the statistical power and the capacity to identify small and moderate effect modifiers of HDL2. Further, HDL2 patients with broader geographic origins could be studied.

The findings of the CTG repeat distribution suggest that HDL2 may be as frequent if not more frequent in The Gambia. Analysis of the haplotype diversity also suggests that there may be other disease haplotypes that are specific to, or more common in the Gambian population. However, no studies have been published from The Gambia looking for patients with HDL2 or the potential link of the West African slave trade (Micheletti et al., 2020) that could have introduced HDL2 in North American African ancestry individuals. A prospect would be to establish collaborations with genetic researchers in The Gambia to investigate whether HDL2 is present in this population. If HDL2 is identified in this African population, clinical efforts would be required to find these patients and address an unmet need from a genetic counselling and disease management perspective. Research prospects would also be increased. To date HDL2 has only been reported in two African countries, South Africa (Anderson et al., 2017;

Baine, Krause and Greenberg, 2016; Krause et al., 2015) and Botswana (Ocampo, Daimari and Oyekunle, 2018).

It would be important to expand the study of the broader group of trinucleotides repeat disorders as this may allow for common mechanisms of modification among the disease to be explored. Genes involved in the DNA damage response could underpin a common mechanism seen in the repeat expansion disorders (Jones, Houlden and Tabrizi, 2017). This also has the potential to provide therapeutic targets across multiple trinucleotide repeat disorders (Jones, Houlden and Tabrizi, 2017). Extensive research has been conducted in polyglutamine repeat expansion disorders focusing on the role of DNA repair or maintenance genes. These have been identified as *trans*-acting modifiers of HD in European ancestry individuals (Correia et al., 2015; GeM-HD Consortium, 2015; GeM-HD Consortium, 2019).

Variants in these modifier loci may provide a promising line of research that could be explored in the HDL2 patients. This may clarify the role of somatic expansion in the CTG repeats. In the HDL2 patients the main driver of somatic expansion may be a *trans*-acting modifier which would explain the very small association identified between the CTG repeats and somatic expansion. In another CTG repeat expansion disorder, myotonic dystrophy type 1 (DM1), variation in somatic expansion has been shown to contribute to the variation in age of onset (Morales et al., 2016). This is thought to be mediated through a polymorphism in the *MSH3* mismatch repair gene identified to have a key role in the variation of somatic expansion in DM1 patients (Morales et al., 2016). This demonstrates that this modifier effect is not just occurring in polyglutamine disorders like HD but extends to other trinucleotide repeat disorders that also result from a CTG repeat expansion.

5.4. Concluding remarks

This study has explored the unique aspects of the molecular basis of Huntington disease and Huntington disease-like 2 in South Africa. In general, African ancestry patients have been largely understudied as the focus of HD research is primarily European ancestry patients. This study has however emphasised the importance of studying genetically diverse populations and shown the locus heterogeneity within the *HTT* locus, the disease phenotype and highlighted differences among an HD phenocopy. In the *HTT* analysis, a large amount of sequence diversity was identified when compared with other European ancestry populations and novel likely African-specific alleles were defined. The CAG repeat was however shown to be the main determinant of disease severity which correlates with previous findings. Although the Q^1-

2-0-9-2 allele structure, characterised by loss of the CCGCCA sequence occurring on haplotype B2, was proposed as an African *cis*-acting modifier, that appears to modify disease diagnosis in individuals affected with HD. Although, the mechanism is not driven by somatic expansion which was an interesting finding as these patients had the earliest AoD, it was hypothesised they would have the highest amount of somatic expansion however the inverse was identified. These findings differ considerably to what has been previously identified in European ancestry HD patients and provide a different perspective of HD. It also highlights that the current therapeutic approaches may not be appropriate for the majority of African ancestry patients and the importance of developing therapies that include genetically diverse populations. While in the *JPH3* analysis, very little sequence diversity was identified at the *JPH3* locus and all the disease alleles had the same typical allele structure and occur on a single disease haplotype. The CTG repeat was confirmed to be the main driver of disease severity which correlates to previous findings and the effect of the CAG repeat in HD. The larger haplotype 6: CC**CA**T**C**G was also shown to modify disease although the mechanism is uncertain. A major consideration when studying rare diseases is the sample size which inevitably limits the power to identify significant correlations. The comparison of the findings of the *HTT* and *JPH3* loci has revealed that despite the clinical similarities between HD and HDL2, stark contrasts were identified that suggests a different potential mechanism of disease modification may be occurring in the two diseases. The only similarity being the association of the number of disease-associated repeats with disease severity.

Reference list

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., et al. (2018). The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Research*, 46(W1), W537-W544.

Agena Bioscience. 2015. *Single Nucleotide Polymorphism Detection with the iPLEX®Assay and the MassARRAY® System* [Online]. Available: https://agenabio.com/products/massarray-system/ [Accessed 07.09.2020].

Agena Bioscience. 2016. *MassARRAY System* [Online]. Available: https://agenabio.com/products/massarray-system/ [Accessed 10.09.2020].

Agilent Technologies. 2015. *Agilent 2100 Bioanalyzer 2100 Expert User's Guide* [Online]. Available: https://www.agilent.com/en/product/automated-electrophoresis/bioanalyzer-systems/bioanalyzer-instrument [Accessed 10.09.2020].

Ambrose, C. M., Duyao, M. P., Barnes, G., Bates, G. P., Lin, C. S., Srinidhi, J., Baxendale, S., Hummerich, H., et al. (1994). Structure and Expression of the Huntington's Disease Gene: Evidence Against Simple Inactivation Due to an Expanded CAG Repeat. *Somatic Cell and Molecular Genetics*, 20(1), 27-38.

Anderson, D. G., Haagensen, M., Ferreira-Correia, A., Pierson, R., Carr, J., Krause, A. & Margolis, R. L. (2019a). Emerging Differences Between Huntington's Disease-Like 2 and Huntington's Disease: A Comparison using MRI Brain Volumetry. *NeuroImage: Clinical*, 21(101666), 1-8.

Anderson, D. G., Krause, A. & Margolis, R. L. 2004 [Updated 2019 Jun 27]. *Huntington Disease-Like 2* [Online]. GeneReviews®. Seattle (WA): University of Washington, Seattle; 1993-2021. Available: https://www.ncbi.nlm.nih.gov/books/NBK1529/ [Accessed 07.09.2020].

Anderson, D. G., Walker, R. H., Connor, M., Carr, J., Margolis, R. L. & Krause, A. (2017). A Systematic Review of the Huntington Disease-Like 2 Phenotype. *Journal of Huntington's disease*, 6(1), 37-46.

Anderson, D. G. M., Ferreira-Correia, A., Rodrigues, F. B., Aziz, A. N., Carr, J., Wild, E. J., Margolis, R. L. & Krause, A. (2019b). Comparison of the Huntington's Disease like 2 and Huntington's Disease Clinical Phenotypes. *Movement Disorders Clinical Practice*, 6(4), 302-331.

Andrade, M. A. & Bork, P. (1995). HEAT Repeats in the Huntington's Disease Protein. *Nature Genetics*, 11(2), 115-116.

Andrew, S., Theilmann, J., Almqvist, E., Norremolle, A., Lucotte, G., Anvret, M., Sorensen, S. A., Turpin, J. C., et al. (1993). DNA Analysis of Distinct Populations Suggests Multiple Origins for the Mutation Causing Huntington Disease. *Clinical Genetics*, 43(6), 286-294.

Andrews, S. 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed 20.02.2018].

Avila-Giron, R. (1973). Medical and Social Aspects of Huntington's Chorea in the State of Zulia, Venezuela. *Advances in Neurology*, 1, 261-266.

Aziz, N. A., Jurgens, C. K., Landwehrmeyer, G. B., EHDN Registry Study Group., van Roon-Mom, W. M. C., van Ommen, G. J. B., Stijnen, T. & Roos, R. A. C. (2009). Normal and Mutant *HTT* Interact to Affect Clinical Severity and Progression in Huntington Disease. *Neurology*, 73(16), 1280-1285.

Baine, F. K., Kay, C., Ketelaar, M. E., Collins, J. A., Semaka, A., Doty, C. N., Krause, A., Greenberg, L. J., et al. (2013). Huntington Disease in the South African Population Occurs on Diverse and Ethnically Distinct Genetic Haplotypes. *European Journal of Human Genetics*, 21(10), 1120-1127.

Baine, F. K., Krause, A. & Greenberg, L. J. (2016). The Frequency of Huntington Disease and Huntington Disease-Like 2 in the South African Population. *Neuroepidemiology*, 46(3), 198-202.

Baliko, L., Csala, B. & Czopf, J. (2004). Suicide in Hungarian Huntington's Disease Patients. *Neuroepidemiology*, 23(5), 258-260.

Bañez-Coronel, M., Ayhan, F., Tarabochia, A. D., Zu, T., Perez, B. A., Khoramian Tusi, S., Pletnikova, O., Borchelt, D. R., et al. (2015). RAN Translation in Huntington Disease. *Neuron*, 88(4), 667–677.

Bates, G. P., Dorsey, R., Gusella, J. F., Hayden, M. R., Kay, C., Leavitt, B. R., Nance, M., Ross, C. A., et al. (2015). Huntington disease. *Nature Reviews Disease Primers*, 1(15005), 1-21.

Bauer, I., Gencik, M., Laccone, F., Peters, H., Weber, B. H. F., Feder, E. H., Weirich, H., Morris-Rosendahl, D. J., et al. (2002). Trinucleotide Repeat Expansions in the *Junctophilin-3* Gene are not Found in Caucasian Patients with a Huntington's Disease-Like Phenotype. *Annals of Neurology*, 51(5), 662.

Becanović, K., Nørremølle, A., Neal, S. J., Kay, C., Collins, J. A., Arenillas, D., Lilja, T., Gaudenzi, G., et al. (2015). A SNP in the *HTT* Promoter Alters NF-kB Binding and is a Bidirectional Genetic Modifier of Huntington Disease. *Nature neuroscience*, 18(6), 807-816.

Beckman Coulter. 2016. *Agencourt AMPure XP PCR Purification* [Online]. Available: https://www.beckman.co.za/reagents/genomic/cleanup-and-size-selection/pcr [Accessed 17.04.2020].

Benn-Torres, J., Bonilla, C., Robbins, C. M., Waterman, L., Moses, T. Y., Hernandez, W., Santos, E. R., Bennett, F., et al. (2008). Admixture and Population Stratification in African Caribbean Populations. *Annals of Human Genetics* 72(1), 90-98.

Bhattacharyya, A., Thakur, A. K., Chellgren, V. M., Thiagarajan, G., Williams, A. D., Chellgren, B. W., Creamer, T. P. & Wetzel, R. (2006). Oligoproline Effect on Polyglutamine Conformation and Aggregation. *Journal of Molecular Biology*, 355(3), 524-535.

Bhattacharyya, K. B. 2011. George Huntington. *Eminent Neuroscientists Their Lives and Works*. India: Association of Neuroscientists of Eastern India Academic Publishers.

Brock, G. J. R., Anderson, N. H. & Monckton, D. G. (1999). *Cis*-acting Modifiers of Expanded CAS/CTG Triplet Repeat Expandability: Associations with Flanking GC Content and Proximity to CpG Islands. *Human Molecular Genetics,* 8(6), 1061–1067.

Budworth, H. & McMurray, C. T. (2013). A Brief History of Triplet Repeat Diseases. *Methods in Molecular Biology,* 1010, 3-17.

Busan, S. & Weeks, K. M. (2013). The Role of Context in RNA Structure: Flanking Sequences Reconfigure CAG Motif Folding in *Huntingtin* Exon 1 Transcripts. *Biochemistry,* 52(46), 8219–8225.

Button, S. C., Ioannidis, P. A. J., Flint, J., Mokrysz, C., Nosek, A. B., Robinson, S. J. E. & Munafò, R. M. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Neuroscience,* 14, 365-376.

Chang, C. M., Yu, Y. L., Fong, K. Y., Wong, M. T., Chan, Y. W., Ng, T. H., Leung, C. M. & Chan, V. (1994). Huntington's Disease in Hong Kong Chinese: Epidemiology and Clinical Picture. *Clinical and Experimental Neurology,* 31, 43-51.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., et al. (2017). Whole-Genome Sequencing for an Enhanced Understanding of Genetic Variation Among South Africans. *Nature Communications,* 8(2062), 1-12.

Chung, M.-Y., Ranum, L. P. W., 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. *Nature Genetics,* 5, 254–258

Ciosi, M., Cumming, S. A., Alshammari, A. M., Symeonidi, E., Herzyk, P., McGuinness, D., Galbraith, J., Hamilton, G., et al. 2018. Library Preparation and MiSeq Sequencing for the Genotyping-By-Sequencing of the Huntington Disease *HTT* Exon One Trinucleotide Repeat and the Quantification of Somatic Mosaicism. Protocol Exchange.

Ciosi, M., Maxwell, A., Cumming, S. A., Hensman Moss, D. J., Alshammari, A. M., Flower, M. D., Durr, A., Leavitt, B. R., et al. (2019). A Genetic Association Study of Glutamine-Encoding DNA Sequence Structures, Somatic CAG Expansion, and DNA Repair Gene Variants, with Huntington Disease Clinical Outcomes. *EBioMedicine* 48, 568-580.

Correia, K., Harold, D., Kim, K.-H., Holmans, P., Jones, L., Orth, M., Myers, R. H., Kwak, S., et al. (2015). The Genetic Modifiers of Motor Onset Age (GeM MOA) Website: Genome-Wide Association Analysis for Genetic Modifiers of Huntington's Disease. *Journal of Huntington's disease,* 4(3), 279-284.

Di Maio, L., Squitieri, F., Napolitano, G., Campanella, G., Trofatter, J. A. & Conneally, M. P. (1993). Suicide Risk in Huntington's Disease. *Journal of Medical Genetics,* 30(4), 293-295.

DiFiglia, M., Sapp, E., Chase, K. O., Davies, S. W., Bates, G. P., Vonsattel, J. P. & Aronin, N. (1997). Aggregation of Huntingtin in Neuronal Intranuclear Inclusions and Dystrophic Neurites in Brain. *Science,* 277(5334), 1990-1993.

Djousse, L., Knowlton, B., Hayden, M. R., Almqvist, E. W., Brinkman, R., Ross, C. A., Margolis, R. L., Rosenblatt, A., et al. (2003). Interaction of Normal and Expanded CAG Repeat

Sizes Influence Age at Onset of Huntington Disease. *American Journal of Medical Genetics Part A,* 199A(3), 279-282.

Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., et al. (1993). Trinucleotide Repeat Length Instability and Age of Onset in Huntington's Disease. *Nature Genetics,* 4(4), 387-392.

Eichler, E. E., Holden, J. J. A., Popovich, B. W., Reiss, A. L., Snow, K., Thibodeau, S. N., Richards, S. C., Ward, P. A., et al. (1994). Length of uninterrupted CGG repeats determines instability in the *FMR1* gene. *Nature Genetics,* 8, 88–94.

Ferreira-Correia, A., Anderson, D. G., Cockcroft, K. & Krause, A. (2019). The Neuropsychological Deficits and Dissociations in Huntington Disease-Like 2: A Series of Case-Control Studies. *Neuropsychologia,* 136(107238).

Flower, M., Lomeikaite, V., Ciosi, M., Cumming, S., Morales, F., Lo, K., Hensman Moss, D., Jones, L., et al. (2019). *MSH3* Modifies Somatic Instability and Disease Severity in Huntington's and Myotonic Dystrophy Type 1. *Brain: a journal of neurology,* 142(7), 1876-1886.

Gellera, C., Meoni, C., Castellotti, B., Zappacosta, B., Girotti, F., Taroni, F. & DiDonato, S. (1996). Errors in Huntington Disease Diagnostic Test Caused by Trinucleotide Deletion in the *IT15* Gene. *American Journal of Human Genetics,* 59(2), 475-477.

GeM-HD Consortium (2015). Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell,* 162(3), 516-526.

GeM-HD Consortium (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell,* 178(4), 887-900.

Goldberg, Y. P., Kremer, B., Andrew, S. E., Theilmann, J., Graham, R. K., Squitieri, F., Telenius, H., Adam, S., et al. (1993). Molecular Analysis of New Mutations for Huntington's Disease: Intermediate Alleles and Sex of Origin Effects. *Nature Genetics,* 5(2), 174-179.

Goldberg, Y. P., McMurray, C. T., Zeisler, J., Almqvist, E., Sillence, D., Richards, F., Gacy, A. M., Buchanan, J., et al. (1995). Increased Instability of Intermediate Alleles in Families with Sporadic Huntington Disease Compared to Similar Sized Intermediate Alleles in the General Population. *Human Molecular Genetics,* 4(10), 1911-1918.

Goldman, A., Ramsay, M. & Jenkins, T. (1996). Ethnicity and Myotonic Dystrophy: A Possible Explanation for its Absence in Sub-Saharan Africa. *Annals of Human Genetics,* 60(1), 57-65.

Greenberg, J., Bardien, S. & Carr, J. (2008). Huntington's Disease-Like 2 in South Africa. *South African Medical Journal,* 98(3), 166.

Greenberg, L. J., Martell, R. W., Theilman, J., Hayden, M. R. & Joubert, J. (1991). Genetic Linkage Between Huntington Disease and the D4SI0 Locus in South African Families: Further Evidence Against Non-Allelic Heterogeneity. *Human Genetics,* 87(6), 701-708.

Greenstein, P. E., Vonsattel, J. G., Margolis, R. L. & Joseph, J. T. (2007). Huntington's Disease Like-2 Neuropathology. *Movement Disorders,* 22(10), 1416-1423.

Grossman, I. (2009). ADME Pharmacogentics: Current Practices and Future Outlook. *Expert Opinion on Drug Metabolism & Toxicology*, 5(5), 449-462.

Gusella, J. F. & MacDonald, M. E. (1995). Huntington's Disease: CAG Genetics Expands Neurobiology. *Current Opinion in Neurobiology* 5(5), 656-662.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., et al. (1983). A Polymorphic DNA Marker Genetically Linked to Huntington's Disease. *Nature,* 306(5940), 234-238.

Gutekunst, C.-A., Li, S.-H., Yi, H., Mulroy, J. S., Kuemmerle, S., Jones, R., Rye, D., Ferrante, R. J., et al. (1999). Nuclear and Neuropil Aggregates in Huntington's Disease: Relationship to Neuropathology. *The Journal of Neuroscience,* 19(7), 2522–2534.

Harper, P. (1992). The Epidemiology of Huntington's Disease. *Human Genetics,* 89(4), 365-376.

Harper, P. S. (1997). Trinucleotide Repeat Disorders. *Journal of Inherited Metabolic Disease* 20, 122-124.

Harper, P. S. (1999). Huntington's Disease: A Clinical, Genetic and Molecular Model for Polyglutamine Repeat Disorders. *Philosophical Transactions of the Royal Society B Biological Sciences,* 354(1386), 957-961.

Hayden, M. R., MacGregor, J. M. & Beighton, P. H. (1980). The Prevalence of Huntington's Chorea in South Africa. *South African Medical Journal,* 58(5), 193-196.

Hensman Moss, D. J., Pardiñas, A. F., Langbehn, D., Lo, K., Leavitt, B. R., Roos, R. A. C., Durr, A., Mead, S., et al. (2017). Identification of Genetic Variants Associated with Huntington's Disease Progression: A Genome-Wide Association Study. *The Lancet Neurology,* 16(9), 701-711.

Hensman Moss, D. J., Poulter, M., Beck, J., Hehir, J., Polke, J. M., Campbell, T., Adamson, G., Mudanohwo, E., et al. (2014). C9orf72 Expansions are the Most Common Genetic Cause of Huntington Disease Phenocopies. *Neurology,* 82(4), 292-299.

Hofferbert, S., Schanen, C. N., Chehab, F. & Francke, U. (1997). Trinucleotide Repeats in the Human Genome: Size Distributions for all Possible Triplets and Detection of Expanded Disease Alleles in a Group of Huntington Disease Individuals by the Repeat Expansion Detection Method. *Human Molecular Genetics,* 6(1), 77-83.

Holmes, S. E., O'Hearn, E., Rosenblatt, A., Callahan, C., Hwang, H. S., Ingersoll-Ashworth, R. G., Fleisher, A., Stevanin, G., et al. (2001). A Repeat Expansion in the Gene Encoding *Junctophilin-3* is Associated with Huntington Disease-Like 2. *Nature Genetics,* 29(4), 377-378.

Holmes, S. E., O'Hearn, E. E., McInnis, M. G., Gorelick-Feldman, D. A., Kleiderlein, J. J., Callahan, C., Kwak, N. G., Ingersoll-Ashworth, R. G., et al. (1999). Expansion of a Novel CAG Trinucleotide Repeat in the 5′ Region of *PPP2R2B* is Associated with SCA12. *Nature Genetics,* 23, 391-392.

Horvath, S., Langfelder, P., Kwak, S., Aaronson, J., Rosinski, J., Vogt, T. F., Eszes, M., Faull, R. L. M., et al. (2016). Huntington's Disease Accelerates Epigenetic Aging of Human Brain and Disrupts DNA Methylation Levels. *AGING,* 8(7), 1485-1512.

Huntington, G. (1872). On Chorea. *The Medical and Surgical Reporter*.

Illumina. 2016. *Optimizing Cluster Density on Illumina Sequencing Systems* [Online]. Available: https://emea.illumina.com/systems/sequencing-platforms/miseq.html [Accessed 05.04.2018].

Illumina. 2017. *An Introduction to Next-Generation Sequencing Technology* [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing.html [Accessed 05.04.2018].

Ionis Pharmaceuticals, I. 2017. *Ionis Pharmaceuticals Licenses IONIS-HTT Rx to Partner Following Successful Phase 1/2a Study in Patients with Huntington's Disease* [Online]. Available: https://ir.ionispharma.com/news-releases/news-release-details/ionis-pharmaceuticals-licenses-ionis-htt-rx-partner-following [Date 11.12.2017].

Jensen, M. A., Fukushima, M. & Davis, R. W. (2010). DMSO and Betaine Greatly Improve Amplification of GCRich Constructs in DeNovo Synthesis. *PLoS ONE,* 5(6), e11024.

Jiang, H., Mankodi, A., Swanson, M. S., Moxley, R. T. & Thornton, C. A. (2004). Myotonic Dystrophy Type 1 is Associated with Nuclear Foci of Mutant RNA, Sequestration of Muscleblind Proteins and Deregulated Alternative Splicing in Neurons. *Human Molecular Genetics,* 13(24), 3079-3088.

Jones, L., Houlden, H. & Tabrizi, S. J. (2017). DNA Repair in the Trinucleotide Repeat Disorders. *The Lancet Neurology,* 16(1), 88-96.

Kapa Biosystems. 2016. *KAPA Library Quantification Kit Illumina®Platforms Technical Data Sheet* [Online]. Available: https://rochesequencingstore.com/wp-content/uploads/2017/10/KAPA-Lib-Quant-ILMN_9.17-IfU_1.pdf [Accessed 09.03.2018].

Kay, C., Collins, J. A., Wright, G. E. B., Baine, F., Miedzybrodzka, Z., Aminkeng, F., Semaka, A., McDonald, C., et al. (2018). The Molecular Epidemiology of Huntington Disease is Related to Intermediate Allele Frequency and Haplotype in the General Population. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics,* 177(3), 346-357.

Kay, C., Collins, J. C., Caron, N. S., de Andrade Agostinho, L., Findlay-Black, H., Casal, L., Sumathipala, D., Dissanayake, V. H. W., et al. (2019). A Comprehensive Haplotype-Targeting Strategy for Allele-Specific *HTT* Suppression in Huntington Disease. *The American Journal of Human Genetics,* 105(6), 1112-1125.

Kay, C., Skotte, N. H., Southwell, A. L. & Hayden, M. R. (2014). Personalized Gene Silencing Therapeutics for Huntington Disease. *Clinical Genetics,* 86(1), 29-36.

Kennedy, L., Evans, E., Chen, C. M., Craven, L., Detloff, P. J., Ennis, M. & Shelbourne, P. F. (2003). Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum Mol Genet,* 12(24), 3359-3367.

Koide, R., Kobayashi, S., Shimohata, T., Ikeuchi, T., Maruyama, M., Saito, M., Yamada, M., Takahashi, H., et al. (1999). A Neurological Disease Caused by an Expanded CAG Trinucleotide Repeat in the TATA-Binding Protein Gene: A New Polyglutamine Disease? *Human Molecular Genetics,* 8(11), 2047-2053.

Kratter, H. & Finkbeiner, S. (2010). PolyQ Disease: Too Many Q's, Too Much Function? *Neuron,* 67(6), 897–899.

Kraus-Perrotta, C. & Lagalwar, S. (2016). Expansion, Mosaicism and Interruption: Mechanisms of the Cag Repeat Mutation in Spinocerebellar Ataxia Type 1. *Cerebellum Ataxias,* 3(20), 1-11.

Krause, A. (2015). Understanding the Genetic Diversity of South Africa's Peoples. *South African Medical Journal,* 105(7), 544-545.

Krause, A. & Greenberg, J. (2008). Genetic Testing for Huntington's Disease in South Africa. *South African Medical Journal,* 98(3), 193-194.

Krause, A., Mitchell, C., Essop, F., Tager, S., Temlett, J., Stevanin, G., Ross, C., Rudnicki, D., et al. (2015). Junctophilin 3 (JPH3) Expansion Mutations Causing Huntington Disease Like 2 (HDL2) are Common in South African Patients with African Ancestry and a Huntington Disease Phenotype. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics,* 168(7), 573-585.

Krauss, S., Griesche, N., Jastrzebska, E., Chen, C., Rutschow, D. s., Achmuller, C., Dorn, S., Boesch, S. M., et al. (2013). Translation of HTT mRNA with Expanded CAG Repeats is Regulated by the MID1–PP2A Protein Complex. *Nature Communications,* 4(1511), 1-9.

Kremer, B., Almqvist, E., Theilmann, J., Spence, N., Telenius, H., Goldberg, Y. P. & Hayden, M. R. (1995). Sex-Dependent Mechanisms for Expansions and Contractions of the CAG Repeat on Affected Huntington Disease Chromosomes. *American Journal of Human Genetics,* 57(2), 343-350.

Lee, J. M., Ramos, E. M., Lee, J. H., Gillis, T., Mysore, J. S., Hayden, M. R., Warby, S. C., Morrison, P., et al. (2012). CAG Repeat Expansion in Huntington Disease Determines Age at Onset in a Fully Dominant Fashion. *Neurology,* 78(10), 690-695.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., et al. (2016). Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature,* 536(7616), 285-291.

Lekoubou, A., Echouffo-Tcheugui, J. B. & Kengne, A. P. (2014). Epidemiology of Neurodegenerative Diseases in Sub-Saharan Africa: A Systematic Review. *BMC Public Health,* 14(653), 1-32.

Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv:1303.3997,* 2, 1-3.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics,* 25(16), 2078-2079.

Li, S.-H. & Li, X.-J. (2004). Huntingtin–Protein Interactions and the Pathogenesis of Huntington's Disease. *Trends in Genetics,* 20(3), 146-154.

Life Technologies. 2015. *Qubit® dsDNA HS Assay Kits* [Online]. Available: https://www.thermofisher.com/za/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit.html [Accessed 09.03.2018].

Liu, J., McFarland, K. N., Landrian, I., Wu, S. S., Bower, M., Hutter, D., Bushara, K., Teive, H. A. G., et al. (2014). Identifying Novel Interruption Motifs in Spinocerebellar Ataxia Type 10 Expansions. *Neurology and Clinical Neuroscience,* 2(2), 38-43.

Losekoot, M., van Belzen, M. J., Seneca, S., Bauer, P., Stenhouse, S. A., Barton, D. E. & European Molecular Genetic Quality, N. (2013). EMQN/CMGS Best Practice Guidelines for the Molecular Genetic Testing of Huntington disease. *European Journal of Human Genetics,* 21(5), 480-486.

Low, P. A., Allsop, J. L. & Halmagyi, G. M. (1974). Huntington's Chorea: The Rigid Form (Westphal Variant) Treated with Levodopa. *Medical Journal of Australia,* 1(11), 393-394.

Lu, A. T., Narayan, P., Grant, M. J., Langfelder, P., Wang, N., Kwak, S., Wilkinson, H., Chen, R. Z., et al. (2020). DNA Methylation Study of Huntington's Disease and Motor Progression in Patients and in Animal Models. *Nature Communications,* 11(4529), 1-15.

MacDonald, M. E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., et al. (1992). The Huntington's Disease Candidate Region Exhibits Many Different Haplotypes. *Nature Genetics,* 1(2), 99-103.

Machiela, M. J. & Chanock, S. J. (2015). LDlink: A Web-Based Application for Exploring Population-Specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants. *Bioinformatics,* 31(21), 3555-3557.

Magazi, D. S., Bonev, V., Moagi, M., Iqbal, Z., Dludla, M. & van der Meyden, C. H. (2008). Huntington's Disease: Genetic Heterogeneity in Black African Patients. *South African Medical Journal,* 98(3), 200-203.

Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., et al. (1996). Exon 1 of the HD Gene with an Expanded CAG Repeat is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice. *Cell,* 87(3), 493-506.

Margolis, R. L. & Holmes, S. E. (2003). Huntington's Disease-Like 2: A Clinical, Pathological, and Molecular Comparison to Huntington's Disease. *Clinical Neuroscience Research,* 3, 187-196.

Margolis, R. L., Holmes, S. E., Rosenblatt, A., Gourley, L., O'Hearn, E., Ross, C. A., Seltzer, W. K., Walker, R. H., et al. (2004). Huntington's Disease–Like 2 (HDL2) in North America and Japan. *Annals of Neurology,* 56(5), 670-674.

Margolis, R. L., Holmes, S. E., Rudnicki, D. D., O'Hearn, E., Ross, C. A., Pletnikova, O. & Troncoso, J. C. 2006. Huntington's Disease-Like 2. *In:* Wells, R. D. & Ashizawa, T. (eds.) *Genetic Instabilities and Neurological Diseases.* Academic Press.

Margolis, R. L., O'Hearn, E., Rosenblatt, A., Willour, V., Holmes, S. E., Franz, M. L., Callahan, C., Hwang, H. S., et al. (2001). A Disorder Similar to Huntington's Disease is Associated with a Novel CAG Repeat Expansion. *Annals of Neurology*, 50(3), 373-380.

Margolis, R. L. & Rudnicki, D. D. (2016). Pathogenic Insights from Huntington's Disease-Like 2 and other Huntington's Disease Genocopies. *Current Opinion in Neurology*, 29(6), 743-748.

Margolis, R. L., Rudnicki, D. D. & Holmes, S. E. (2005). Huntington's Disease Like-2: Review and Update. *Acta Neurologica Taiwanica*, 14(1), 1-8.

Mariani, L.-L., Tesson, C., Charles, P., Cazeneuve, C., Hahn, V., Youssov, K., Freeman, L., Grabli, D., et al. (2016). Expanding the Spectrum of Genes Involved in Huntington Disease Using a Combined Clinical and Genetic Approach. *JAMA Neurology*, 73 (9), 1105-1114.

Marshall, F. J. (2006). Tetrabenazine as Antichorea Therapy in Huntington Disease: A Randomized Controlled Trial. *Neurology* 66(3), 366-372.

Martin, M. (2011). Cutadapt Removes Adapter Sequence from High-Throughput Sequencing Reads. *EMBnet.journal*, 17(1), 10-12.

Marxreiter, F., Stemick, J. & Kohl, Z. (2020). Huntingtin Lowering Strategies. *International Journal of Molecular Sciences*, 21(6), 2146.

Maxwell, A. 2016. *RefGeneratr: Dynamic Multi-Loci/Multi-Repeat Tract Microsatellite Reference Sequence Generator* [Online]. Available: https://github.com/helloabunai/RefGeneratr [Accessed 04.05.2018].

Maxwell, A. 2018. *ScaleHD* [Online]. Available: https://github.com/helloabunai/ScaleHD [Accessed 20.02.2018].

Merritt, A. D., Conneally, P. M., Rahman N. F. & Drew, A. L. 1969. *Juvenile Huntington's Chorea*, Amsterdam, Excerpta Medica Foundation.

Micheletti, J. S., ., Bryc, K., Ancona Esselmann, G. S., Freyman, A. W., Moreno, E. M., Poznik, G. D., Shastri, J. A., 23andMe Research Team., et al. (2020). Genetic Consequences of the Transatlantic Slave Trade in the Americas. *The American Journal of Human Genetics*, 107(2), 265-277.

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shaw, P. D. & Marshall, D. (2013). Using Tablet for Visual Exploration of Second-Generation Sequencing Data. *Briefings in Bioinformatics*, 14(2), 193-202.

Montoya, A., Price, B. H., Menear, M. & Lepage, M. (2006). Brain Imaging and Cognitive Dysfunctions in Huntington's Disease. *Journal of Psychiatry & Neuroscience*, 31(1), 21-29.

Morgan, C. J. (2017). Use of Proper Statistical Techniques for Research Studies with Small Samples. *American Journal of Physiology-Lung Cellular Molecular Physiology*, 313, L873–L877.

Musova, Z., Mazanec, R., Krepelova, A., Ehler, E., Vales, J., Jaklova, R., Prochazka, T., Koukal, P., et al. (2009). Highly Unstable Sequence Interruptions of the CTG Repeat in the

Myotonic Dystrophy Gene. *American Journal of Medical Genetics Part A,* 149A(7), 1365–1374.

Nakashima, K., Watanabe, Y., Kusumi, M., Nanba, E., Maeoka, Y., Nakagawa, M., Igo, M., Irie, H., et al. (1996). Epidemiological and Genetic Studies of Huntington's Disease in the San-in Area of Japan. *Neuroepidemiology,* 15(3), 126-131.

Nance, M. A. & Myers, R. H. (2001). Juvenile Onset Huntington's Disease-Clinical and Research Perspectives. *Developmental Disabilities Research Reviews,* 7(3), 153-157.

Nasir, J., Floresco, S. B., O'Kusky, J. R., Diewert, V. M., Richman, J. M., Zeisler, J., Borowski, A., Marth, J. D., et al. (1995). Targeted Disruption of the Huntington's Disease Gene Results in Embryonic Lethality and Behavioral and Morphological Changes in Heterozygotes. *Cell,* 81(5), 811-823.

Neueder, A., Landles, C., Ghosh, R., Howland, D., Myers, R. H., Faull, R. L. M., Tabrizi, S. J. & Bates, G. P. (2017). The Pathogenic Exon 1 HTT Protein is Produced by Incomplete Splicing in Huntington's Disease Patients. *Scientific Reports,* 7(1), 1-10.

Nishi, M., Hashimoto, K., Kuriyama, K., Komazaki, S., Kano, M., Shibata, S. & Takeshima, H. (2002). Motor Discoordination in Mutant Mice Lacking Junctophilin Type 3. *Biochemical and Biophysical Research Communications,* 292, 318-324.

Nishi, M., Mizushima, A., Nakagawara, K.-i. & Takeshima, H. (2000). Characterization of Human Junctophilin Subtype Genes. *Biochemical and Biophysical Research Communications,* 273(3), 920-927.

Nørremølle, A., Budtz-Jørgensen, E., Fenger, K., Nielsen, J., Sørensen, S. & Hasholt, L. (2009). 4p16.3 Haplotype Modifying Age at Onset of Huntington Disease. *Clinical Genetics,* 75(3), 244-250.

Ocampo, C., Daimari, R. & Oyekunle, A., A. (2018). Huntington's Disease-Like 2 with an Expansion Mutation of the *Junctophilin-3* Gene; First Reported Case from Botswana. *Journal of Clinical Neuroscience,* 47, 126-127.

Orr, H. T. & Zoghbi, H. Y. (2007). Trinucleotide Repeat Disorders. *Annual Review of Neuroscience,* 30, 575-621.

Paradisi, I., Ikonomu, V. & Arias, S. (2012). Huntington Disease-Like 2 (HDL2) in Venezuela: Frequency and Ethnic Origin. *Journal of Human Genetics,* 58(1), 3-6.

Pouladi, M. A., Stanek, L. M., Xie, Y., Franciosi, S., Southwell, A. L., Deng, Y., Butland, S., Zhang, W., et al. (2012). Marked Differences in Neurochemistry and Aggregates Despite Similar Behavioural and Neuropathological Features of Huntington Disease in the Full-Length BACHD and YAC128 Mice. *Human Molecular Genetics,* 21(10), 2219-2232.

Pringsheim, T., Wiltshire, K., Day, L., Dykeman, J., Steeves, T. & Jette, N. (2012). The Incidence and Prevalence of Huntington's Disease: A Systematic Review and Meta-Analysis. *Movement Disorders,* 27(9), 1083-1091.

PTC Therapeutics Inc. 2018. *PTC Therapeutics and CHDI Foundation Announce a Collaboration on a Small-Molecule Therapeutic for Huntington's Disease* [Online]. Available:

https://ir.ptcbio.com/news-releases/news-release-details/ptc-therapeutics-and-chdi-foundation-announce-collaboration [Date 04.04.2018].

Ramos, E., Doumatey, A., Elkahloun, A. G., Shriner, D., Huang, H., Chen, G., Zhou, J., McLeod, H., et al. (2014). Pharmacogenomics, Ancestry and Clinical Decision Making for Global Populations. *The Pharmacogenomics Journal*, 14(3), 217-222.

Ravi, R., Walton, K. & Khosroheidari, M. (2018). MiSeq: A Next Generation Sequencing Platform for Genomic Analysis. *Methods in Molecular Biology*, 1706, 223-232.

Richards, R. I. & Sutherland, G. R. (1997). Dynamic Mutation: Possible Mechanisms and Significance in Human Disease. *Trends in Biochemical Science*, 22(11), 432-436.

Rodrigues Riccioppo, G. G., Walker, R. H., Brice, A., Cazeneuve, C., Russaouen, O., Teive, H. A. G., Munhoz, R. P., Becker, N., et al. (2008). Huntington's Disease-Like 2 in Brazil—Report of 4 Patients. *Movement Disorders*, 23(15), 2244-2255.

Roos, R. A. (2002). Polyglutamine Pathogenesis: Emergence of Unifying Mechanisms for Huntington's Disease and Related Disorders. *Neuron*, 35(5), 819-822.

Roos, R. A. (2010). Huntington's Disease: A Clinical Review. *Orphanet Journal of Rare Diseases*, 5(40), 1-8.

Ross, C. A. (1995). When More is Less: Pathogenesis of Glutamine Repeat Neurodegenerative Diseases. *Neuron*, 15(3), 493-496.

Ross, C. A. & Tabrizi, S. J. (2011). Huntington's Disease: From Molecular Pathogenesis to Clinical Treatment. *The Lancet Neurology*, 10(1), 83-98.

Rubinsztein, D. C., Amos, W., Leggo, J., Goodburn, S., Ramesar, J. R., Bontrop, R., McMahon, R., Barton, D. E., et al. (1994). Mutation Bias Provides a Model for Evolution of Huntington's Disease and Predicts a General Increase in Disease Prevalence. *Nature Genetics*, 7(4), 525-530.

Rubinsztein, D. C., Leggo, J., Coles, R., Almqvist, E., Biancalana, V., Cassiman, J. J., Chotai, K., Connarty, M., et al. (1996). Phenotypic Characterization of Individuals with 30-40 CAG Repeats in the Huntington Disease (HD) Gene Reveals HD Cases with 36 Repeats and Apparently Normal Elderly Individuals with 36-39 Repeats. *American Journal of Medical Genetics Part A*, 59(1), 16-22.

Rubinsztein, D. C., Leggo, J., Goodburn, S., Barton, D. E. & Ferguson-Smith, M. A. (1995). Haplotype Analysis of the A2642 and (CAG)n Polymorphisms in the Huntington's Disease (HD) Gene Provides an Explanation for an Apparent 'Founder' HD Haplotype. *Human Molecular Genetics*, 4(2), 203-206.

Rudnicki, D. D., Holmes, S. E., Lin, M. W., Thornton, C. A., Ross , C. A. & Margolis, R. L. (2007). Huntington's Disease–Like 2 is Associated with CUG Repeat-Containing RNA Foci. *Annals of Neurology*, 61(3), 272-282.

Rudnicki, D. D., Pletnikova, O., Vonsattel, J. P. G., Ross, C. & Margolis, R. (2008). A Comparison of Huntington Disease and Huntington Disease-Like 2 Neuropathology. *The Journal of Neuropathology & Experimental Neurology*, 67(4), 366-374.

Russell, L. 2020. Estimated Marginal Means, aka Least-Squares Means. 1.5.1 ed.

Santos, C., Wanderley, H., Vedolin, L., Pena, S. D., Jardim, L. & Sequeiros, J. (2008). Huntington Disease-Like 2: The First Patient with Apparent European Ancestry. *Clinical Genetics,* 73(5), 480-485.

Schulte, J. & Littleton, J. T. (2011). The Biological Function of the Huntingtin Protein and its Relevance to Huntington's Disease Pathology. *Current Trends in Neurology,* 5, 65-78.

Scrimgeour, E. M. & Simpson, S. A. (1992). Huntington Disease in Black African Populations. *Human Genetics,* 90(3), 1-2.

Seixas, A. I., Holmes, S. E., Takeshima, H., Pavlovich, A., Sachs, N., Pruitt, J. L., Silveira, I., Ross, C. A., et al. (2012). Loss of Junctophilin-3 Contributes to Huntington Disease-Like 2 Pathogenesis. *Annals of Neurology,* 71(2), 245-257.

Semaka, A., Creighton, S., Warby, S. & Hayden, M. R. (2006). Predictive Testing for Huntington Disease: Interpretation and Significance of Intermediate Alleles. *Clinical Genetics,* 70(4), 283-294.

Shelbourne, P. F., Keller-McGandy, C., Bi, W. L., Yoon, S., Dubeau, L., Veitch, N. J., Vonsattel, J. P., Wexler, N. S., et al. (2007). Triplet Repeat Mutation Length Gains Correlate with Cell-Type Specific Vulnerability in Huntington Disease Brain. *Human Molecular Genetics,* 16(10), 1133-1142.

Shinde, D. (2003). Taq DNA Polymerase Slippage Mutation Rates Measured by PCR and Quasi-Likelihood Analysis: (CA/GT)n and (A/T)n Microsatellites. *Nucleic Acids Research,* 31(3), 974-980.

Squitieri, F., Andrew, S. E., Goldberg, Y. P., Kremer, B., Spence, N., Zeisler, J., Nichol, K., Theilman, J., et al. (1994). DNA Haplotype Analysis of Huntington Disease Reveals Clues to the Origins and Mechanisms of CAG Expansion and Reasons for Geographic Variations of Prevalence. *Human Molecular Genetics,* 3(12), 2103-2114.

Stephens, M. & Donnelly, P. (2003). A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *American Journal of Human Genetics,* 73(5), 1162-1169.

Stephens, M., Smith, N. J. & Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *American Journal of Human Genetics,* 68(4), 978-989.

Sturrock, A. & Leavitt, B. R. (2010). The Clinical and Genetic Features of Huntington Disease. *Journal of Geriatric Psychiatry and Neurology,* 23(4), 243-259.

Sun, Y.-M., Zhang, Y.-B. & Wu, Z.-Y. (2016). Huntington's Disease: Relationship Between Phenotype and Genotype. *Molecular Neurobiology,* 54(1), 342-348.

Takano, H. & Gusella, J. F. (2002). The Predominantly HEAT-Like Motif Structure of Huntingtin and its Association and Coincident Nuclear Entry with Dorsal, an NF-kB/Rel/Dorsal Family Transcription Factor. *BMC Neuroscience,* 3(1), 15.

Takeshima, H., Komazaki, S., Nishi, M., Iino, M. & Kangawa, K. (2000). Junctophilins: A Novel Family of Junctional Membrane Complex Proteins. *Molecular Cell*, 6, 11-22.

Telenius, H., Kremer, B., Goldberg, Y. P., Theilmann, J., Andrew, S. E., Zeisler, J., Adam, S., Greenberg, C., et al. (1994). Somatic and Gonadal Mosaicism of the Huntington Disease Gene CAG Repeat in Brain and Sperm. *Nature Genetics*, 6(4), 409-414.

The Huntington's Disease Collaborative Research Group (1993). A Novel Gene Containing a Trinucleotide Repeat that is Expanded and Unstable on Huntington's Disease Chromosomes. *Cell*, 72(6), 971-983.

The U.S.–Venezuela Collaborative Research Project & Wexler, N. S. (2004). Venezuelan Kindreds Reveal that Genetic and Environmental Factors Modulate Huntington's Disease Age of Onset. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10), 3498-3503.

UniQure Biopharma B.V. 2019. *Safety and Proof-of-Concept (POC) Study with AMT-130 in Adults with Early Manifest Huntington Disease* [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT04120493 [Accessed 08.09.2020].

Veitch, N. J., Ennis, M., McAbney, J. P., Shelbourne, P. F., Monckton, D. G. & Project, T. U.-V. C. R. (2007). Inherited CAG·CTG allele length is a major modifier of somatic mutation length variability in Huntington disease. *DNA repair*, 6, 789-796.

Via, M., Gignoux, C. R., Roth, L. A., Fejerman, L., Galanter, J., Choudhry, S., Toro-Labrador, G., Viera-Vera, J., et al. (2011). History Shaped the Geographic Distribution of Genomic Admixture on the Island of Puerto Rico. *PLoS ONE*, 6(1), 1-8.

Vonsattel, J. P. & DiFiglia, M. (1998). Huntington Disease. *Journal of Neuropathology and Experimental Neurology*, 57(5), 369-384.

Voyager Therapeutics Inc. 2020. *Voyager Therapeutics Announces Fourth Quarter and Full Year 2019 Financial Results and Corporate Updates* [Online]. Available: https://www.globenewswire.com/news-release/2020/03/03/1994663/0/en/Voyager-Therapeutics-Announces-Fourth-Quarter-and-Full-Year-2019-Financial-Results-and-Corporate-Updates.html [Date 03.03.2020].

Walker, F. O. (2007). Huntington's Disease. *The Lancet*, 369(9557), 218-228.

Walker, R. H., Jankovic, J., O'Hearn, E. & Margolis, R. L. (2003). Phenotypic Features of Huntington's Disease-Like 2. *Movement Disorders*, 18(12), 1527-1530.

Warby, S. C., Montpetit, A., Hayden, A. R., Carroll, J. B., Butland, S. L., Visscher, H., Collins, J. A., Semaka, A., et al. (2009). CAG Expansion in the Huntington Disease Gene is Associated with a Specific and Targetable Predisposing Haplogroup. *American Journal of Human Genetics*, 84(3), 351-366.

Warby, S. C., Visscher, H., Collins, J. A., Doty, C. N., Carter, C., Butland, S. L., Hayden, A. R., Kanazawa, I., et al. (2011). *HTT* Haplotypes Contribute to Differences in Huntington Disease Prevalence Between Europe and East Asia. *European Journal of Human Genetics*, 19(5), 561-566.

Wave Life Sciences. 2019. *Wave Life Sciences Announces Topline Data and Addition of Higher Dose Cohort in Ongoing Phase 1b/2a PRECISION-HD2 Trial in Huntington's Disease* [Online]. Available: https://www.biospace.com/article/releases/wave-life-sciences-announces-topline-data-and-addition-of-higher-dose-cohort-in-ongoing-phase-1b-2a-precision-hd2-trial-in-huntington-s-disease/ [Date 30.12.2019].

Wave Life Sciences 2021. *Wave Life Sciences Provides Update on Phase 1b/2a PRECISION-HD Trials* [Online]. Available: https://ir.wavelifesciences.com/news-releases/news-release-details/wave-life-sciences-provides-update-phase-1b2a-precision-hd [Date 29.03.2021].

Wilburn, B., Rudnicki, D. D., Zhao, J., Weitz, T. M., Cheng, Y., Gu, X., Greiner, E., Park, C. S., et al. (2011). An Antisense CAG Repeat Transcript at *JPH3* Locus Mediates Expanded Polyglutamine Protein Toxicity In Huntington's Disease-Like 2 Mice. *Neuron,* 70(3), 427-440.

Wild, E. J., Mudanohwo, E. E., Sweeney, M. G., Schneider, S. A., Beck, J., Bhatia, K. P., Rossor, M. N., Davis, M. B., et al. (2008). Huntington's Disease Phenocopies are Clinically and Genetically Heterogeneous. *Movement Disorders,* 23(5), 716-720.

Wild, E. J. & Tabrizi, S. J. (2007a). The Differential Diagnosis of Chorea. *Practical Neurology,* 7(7), 360-373.

Wild, E. J. & Tabrizi, S. J. (2007b). Huntington's Disease Phenocopy Syndromes. *Current Opinion in Neurology* 20(6), 681-687.

Wright, G. E. B., Collins, J. A., Kay, C., McDonald, C., Dolzhenko, E., Xia, Q., Bečanović, K., Semaka, A., et al. (2019). Length of Uninterrupted CAG Repeats, Independent of Polyglutamine Size, Results in Increased Somatic Instability and Hastened Age of Onset in Huntington Disease. *The American Journal of Human Genetics,* 104(6), 1116-1126.

Yang, S., Yang, H., Huang, L., Chen, L., Qin, Z., Li, S. & Li, X. J. (2020). Lack of RAN-mediated toxicity in Huntington's disease knock-in mice. *Proceedings of the National Academy of Sciences,* 117(8), 4411-4417.

Yu, S., Fimmel, A., Fung, D. & Trent, R. J. (2000). Polymorphisms in the CAG Repeat-A Source of Error in Huntington Disease DNA Testing. *Clinical Genetics,* 58(6), 469-472.

Zoghbi, H. Y. & Orr, H. T. (2000). Glutamine Repeats and Neurodegeneration. *Annual Reviews of Neuroscience,* 23, 217-247.

Appendices

Appendix 1. Ethics clearance certificate.

Appendix 2.1. Demographic information for patients affected with HD.

Appendix 2.2. Demographic information for patients affected with HDL2.

Appendix 3. *HTT*/*JPH3* PCR reagents and amplification conditions.

Appendix 4.1. *HTT* MiSeq PCR primers.

Appendix 4.2. *JPH3* MiSeq PCR primers.

Appendix 5. *HTT*/*JPH3* tag-SNPs used for haplotype assignment.

Appendix 6. Haplotype B2 frequencies in populations from the 1,000 Genomes Project.

Appendix 7. *JPH3* repeat tract sequence and variants that cause atypical structures.

Appendix 8. *JPH3* larger haplotypes in disease and non-disease alleles.

Appendix 9. Plagiarism form.

Appendix 10. Submissible article.

Appendix 1. Ethics clearance certificate.

R14/49 «Tit init name»

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M1704130

NAME: (Principal Investigator)	Ms Jessica Levesley and Prof Darren Monckton
DEPARTMENT:	Human Genetics University of the Witwatersrand Division of Human Genetics
PROJECT TITLE:	Investigating Allele Sequence Diversity at the Huntington Disease-associated Genetic Loci (Huntington and junctophilin-3) in the South African Population
DATE CONSIDERED:	Adhoc
DECISION:	Approved unconditionally
CONDITIONS:	Sub-study (M110443)
SUPERVISOR:	Dr Fiona Baine and Prof Amanda Krause
APPROVED BY:	Professor P. Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	17/05/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review in April and will therefore be due in the month of April each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 25/05/2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 2.1. Demographic information for patients affected with HD.

Patient	Expanded CAG size	Sex	AoD (years)	AoO (years)
1	46	F	56.3	-
2	41	F	61.3	-
3	55	F	22.9	20.0
4	46	F	33.4	30.0
5	46	F	43.8	38.0
6	46	M	39.9	39.0
7	52	M	37.1	34.0
8	44	F	48.7	-
9	44	F	63.4	-
10	49	M	39.7	39.0
11	43	M	57.0	-
12	43	M	58.8	-
13	42	M	69.7	-
14	40	M	73.7	-
15	43	F	41.6	39.0
16	42	M	62.7	35.0
17	42	F	63.1	-
18	44	F	55.5	-
19	45	F	38.3	35.0
20	46	F	NA	-
21	46	M	41.3	40.0
22	42	M	50.7	-
23	43	F	48.1	47.0
24	43	M	28.7	24.0
25	51	M	33.1	-
26	46	M	54.3	-
27	45	F	45.5	-
28	47	F	52.6	-
29	49	F	43.6	28.0
30	42	F	53.3	43.0
31	46	F	37.6	-
32	49	F	44.5	-
33	58	F	34.9	-
34	48	F	46.0	-
35	45	F	39.0	-
36	46	M	47.5	43.0
37	40	F	62.0	-
38	41	F	41.5	-
39	39	M	74.8	-
40	54	F	28.4	-
41	45	F	48.7	45.0
42	43	F	51.9	-
43	43	F	50.9	39.0
44	47	F	49.9	-
45	48	F	35.7	-
46	40	F	66.6	62.0
47	41	M	59.5	-
48	41	M	63.4	-
49	45	F	49.0	-
50	51	M	42.8	-
51	45	F	44.0	-
52	42	F	59.9	-
53	43	F	58.8	-
54	42	M	50.5	50.0
55	51	M	36.1	29.0
56	43	M	46.3	42.0
57	43	F	67.6	45.0

58	43	M	48.8	-
59	42	M	58.5	-
60	43	F	72.0	-
61	49	F	37.0	-
62	41	F	77.0	-
63	46	F	49.0	-
64	42	M	57.0	-
65	45	M	46.1	43.0
66	49	F	34.1	-
67	53	M	27.0	-
68	46	F	42.8	36.0
69	45	F	34.7	29.0
70	46	F	47.4	42.0
71	43	F	58.6	-
72	43	F	51.9	-
73	43	M	44.9	45.0
74	73	M	12.9	10.0
75	43	M	38.5	-

Demographic information of individuals affected with HD showing the CAG repeat sizes of expanded alleles, sex, age of diagnosis and age of onset. The age of onset was available for 28 of the 75 patients, whereas the age of diagnosis was available for all but one individual.

Appendix 2.2. Demographic information for patients affected with HDL2.

Patient	Expanded CTG size	Sex	AoD (years)	AoO (years)
1	42	F	56.8	-
2	42	F	42.3	-
3	43	F	74.7	64.0
4	44	M	65.6	62.0
5	44	F	49.1	49.0
6	44	M	61.7	-
7	45	F	59.8	46.0
8	45	F	68.0	58.0
9	45	M	55.4	52.0
10	46	F	61.1	50.0
11	46	F	65.4	-
12	46	F	70.1	-
13	47	M	67.8	50.0
14	47	F	54.5	50.0
15	47	M	56.2	51.0
16	47	F	50.7	49.0
17	47	F	45.0	-
18	47	M	46.5	40.0
19	47	M	63.0	-
20	47	M	47.8	-
21	47	F	51.9	49.0
22	48	F	56.8	49.0
23	48	M	52.0	-
24	49	F	58.0	42.0
25	49	F	53.6	51.0
26	49	M	50.7	-
27	50	M	47.8	-
28	50	M	39.7	38.0
29	50	F	42.7	24.0
30	50	F	40.8	30.0
31	51	F	62.4	59.0
32	51	M	48.7	38.0
33	53	M	34.8	42.0

Demographic information of individuals affected with HDL2 showing the CTG repeat sizes of expanded alleles, sex, age of diagnosis and age of onset. The age of onset was available for 22 of the 33 patients, whereas the age of diagnosis was present for all the individuals affected with HDL2.

Appendix 3. *HTT/JPH3* PCR reagents and amplification conditions**.**

The *HTT* repeat tract PCR reaction mixture for one sample

Reagents	**1x sample (µl)**
H_2O	1.8
10X buffer	1.5
10% DMSO	1.5
HTT-F Miseq primer @5µM	4
HTT-R Miseq primer @5µM	4
Taq polymerase	0.2
DNA template at 10 ng/µl	2
Total	**15**

The *HTT* repeat tract PCR conditions

Cycling parameters	**Cycle number**
96ºC for 5 minutes	1 Cycle
96ºC for 45 seconds	28 Cycles
58.5ºC for 45 seconds	
72ºC for 3 minutes	
72ºC for 10 minutes	1 Cycle

The *JPH3* repeat tract PCR reaction mixture for one sample

Reagents	**1x sample (µl)**
H_20	7.3
10X buffer	1.5
JPH3-F MiSeq primers @ 5µM	2
JPH3-R Miseq primers @ 5µM	2
Taq polymerase	0.2
DNA template at 10ng/ul	2
Total	**15**

The *JPH3* repeat tract PCR conditions

Cycling parameters	**Cycle number**
96ºC for 5 minutes	1 Cycle
96ºC for 30 seconds	28 Cycles
58.5ºC for 30 seconds	
72ºC for 1 minutes	
72ºC for 10 minutes	1 Cycle

The 10X buffer (cat. no. SM0005, Thermo Scientific, ABgene®UK) is made up of; 45 mM Tris-HCl (pH 8.8), 11 mM $(NH_4)_2SO_4$, 4.5 mM MgCl2, 0.113 mg/ml BSA, 4.4 µM EDTA, 1 mM each of dATP, dCTP, dGTP and dTTP)], 0.048 % (v/v) 2-Mercaptoethanol (βME, BioReagent, for molecular biology, 99%, Sigma-Aldrich). The Taq polymerase used in both PCR reactions was from Sigma (Product # D4545, Sigma, Poole, UK).

Appendix 4.1. *HTT* MiSeq PCR primers.

Nextera XT v2 indexes	Primer name	Sequence P5 adaptor-Common Part-Barcode-Sequencing primer binding site-Spacer-Locus specific primer
S502	HTT F S502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTAATATATGCGACCCTGGAAAAGCTGATGA
S503	HTT F S503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATACTTGCGACCCTGGAAAAGCTGATGA
S505	HTT F S505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGCGACCCTGGAAAAGCTGATGA
S506	HTT F S506	AATGATACGGCGACCACCGAGATCTACACACTGCATAACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCGCGACCCTGGAAAAGCTGATGA
S507	HTT F S507	AATGATACGGCGACCACCGAGATCTACACAAGGAGTAACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATGCGACCCTGGAAAAGCTGATGA
S508	HTT F S508	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCGACCCTGGAAAAGCTGATGA
S510	HTT F S510	AATGATACGGCGACCACCGAGATCTACACCGTCTAATACACTCTTTCCCTACACGACGCTCTTCCGATCTTATTAGCGACCCTGGAAAAGCTGATGA
S511	HTT F S511	AATGATACGGCGACCACCGAGATCTACACTCTCTCCGACACTCTTTCCCTACACGACGCTCTTCCGATCTGCGACCCTGGAAAAGCTGATGA
S513	HTT F S513	AATGATACGGCGACCACCGAGATCTACACTCGACTAGACACTCTTTCCCTACACGACGCTCTTCCGATCTGCGACCCTGGAAAAGCTGATGA
S515	HTT F S515	AATGATACGGCGACCACCGAGATCTACACTTCTAGCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTTGCGACCCTGGAAAAGCTGATGA
S516	HTT F S516	AATGATACGGCGACCACCGAGATCTACACCCTAGAGTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCGACCCTGGAAAAGCTGATGA
S517	HTT F S517	AATGATACGGCGACCACCGAGATCTACACGCGTAAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGCGACCCTGGAAAAGCTGATGA
S518	HTT F S518	AATGATACGGCGACCACCGAGATCTACACCTATTAAGACACTCTTTCCCTACACGACGCTCTTCCGATCTTAATATTGCGACCCTGGAAAAGCTGATGA
S520	HTT F S520	AATGATACGGCGACCACCGAGATCTACACAAGGCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTCATATAGCGACCCTGGAAAAGCTGATGA
S521	HTT F S521	AATGATACGGCGACCACCGAGATCTACACGAGCCTTAACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGCGACCCTGGAAAAGCTGATGA
S522	HTT F S522	AATGATACGGCGACCACCGAGATCTACACTTATGCGAACACTCTTTCCCTACACGACGCTCTTCCGATCTATACTGCGACCCTGGAAAAGCTGATGA

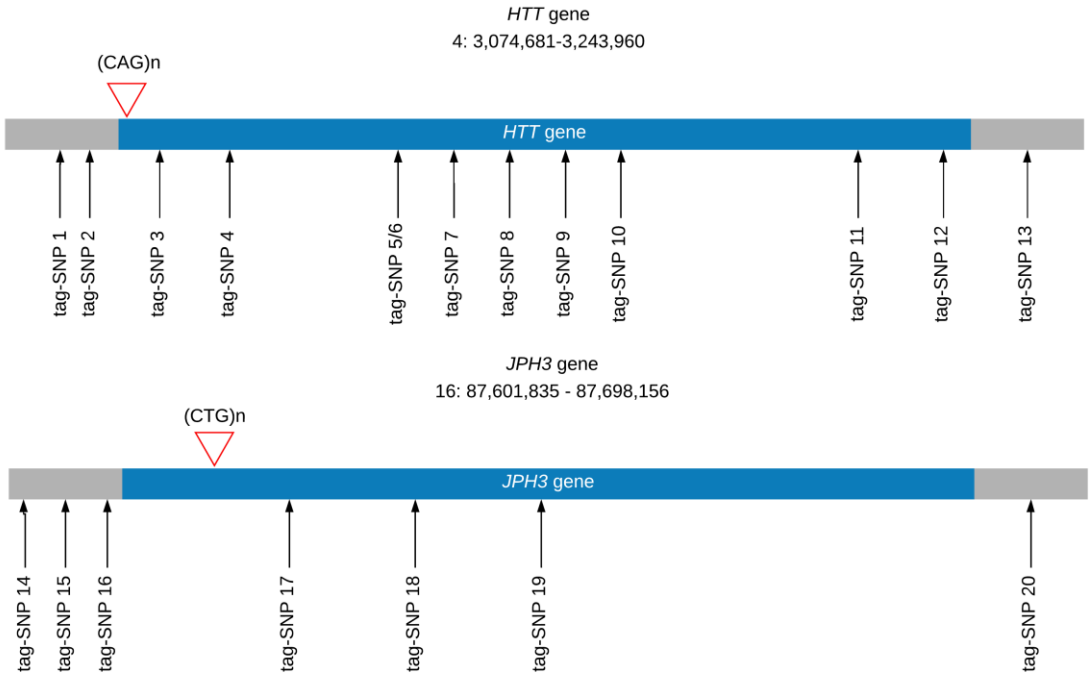
Nextera XT v2 indexes	Primer name	Sequence P7 adaptor-Barcode-Sequencing primer binding site-Spacer-Locus specific primer
N701	HTT R N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCAGCGGCTGTGCCTGC
N702	HTT R N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGCAGCGGCTGTGCCTGC
N703	HTT R N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCAGCGGCTGTGCCTGC
N704	HTT R N704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC
N705	HTT R N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC
N706	HTT R N706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGCAGCGGCTGTGCCTGC
N707	HTT R N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCAGCGGCTGTGCCTGC
N710	HTT R N710	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
N711	HTT R N711	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGCAGCGGCTGTGCCTGC
N712	HTT R N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCAGCGGCTGTGCCTGC
N714	HTT R N714	CAAGCAGAAGACGGCATACGAGATTCATGAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGCAGCGGCTGTGCCTGC
N715	HTT R N715	CAAGCAGAAGACGGCATACGAGATCCTGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCAGCGGCTGTGCCTGC
N716	HTT R N716	CAAGCAGAAGACGGCATACGAGATTAGCGAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGCAGCGGCTGTGCCTGC
N718	HTT R N718	CAAGCAGAAGACGGCATACGAGATGTAGCTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC
N719	HTT R N719	CAAGCAGAAGACGGCATACGAGATTACTACGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCAGCGGCTGTGCCTGC
N720	HTT R N720	CAAGCAGAAGACGGCATACGAGATAGGCTCCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
N721	HTT R N721	CAAGCAGAAGACGGCATACGAGATGCAGCGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC
N722	HTT R N722	CAAGCAGAAGACGGCATACGAGATCTGCGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC
N723	HTT R N723	CAAGCAGAAGACGGCATACGAGATGAGCGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC
N724	HTT R N724	CAAGCAGAAGACGGCATACGAGATCGCTCAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC
N726	HTT R N726	CAAGCAGAAGACGGCATACGAGATGTCTTAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGCAGCGGCTGTGCCTGC
N727	HTT R N727	CAAGCAGAAGACGGCATACGAGATACTGATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGCAGCGGCTGTGCCTGC
N728	HTT R N728	CAAGCAGAAGACGGCATACGAGATTAGCTGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGCAGCGGCTGTGCCTGC
N729	HTT R N729	CAAGCAGAAGACGGCATACGAGATGACGTCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGCAGCGGCTGTGCCTGC

Appendix 4.2. *JPH3* MiSeq PCR primers.

Nextera XT v2 indexes	Primer name	Sequence P5 adaptor-Common Part-Barcode-Sequencing primer binding site-Spacer-Locus specific primer
S502	JPH3 F S502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGGGTCTAAGATGCCACCGCATTCGG
S503	JPH3 F S503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATATTCGCTAAGATGCCACCGCATTCGG
S505	JPH3 F S505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCTAAGATGCCACCGCATTCGG
S506	JPH3 F S506	AATGATACGGCGACCACCGAGATCTACACACTGCATAACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCCTAAGATGCCACCGCATTCGG
S507	JPH3 F S507	AATGATACGGCGACCACCGAGATCTACACAAGGAGTAACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGCCTAAGATGCCACCGCATTCGG
S508	JPH3 F S508	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAAGATGCCACCGCATTCGG
S510	JPH3 F S510	AATGATACGGCGACCACCGAGATCTACACCGTCTAATACACTCTTTCCCTACACGACGCTCTTCCGATCTTATGCCTAAGATGCCACCGCATTCGG
S511	JPH3 F S511	AATGATACGGCGACCACCGAGATCTACACTCTCTCCGACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAAGATGCCACCGCATTCGG
S513	JPH3 F S513	AATGATACGGCGACCACCGAGATCTACACTCGACTAGACACTCTTTCCCTACACGACGCTCTTCCGATCTCTAAGATGCCACCGCATTCGG
S515	JPH3 F S515	AATGATACGGCGACCACCGAGATCTACACTTCTAGCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTGCTAAGATGCCACCGCATTCGG
S516	JPH3 F S516	AATGATACGGCGACCACCGAGATCTACACCCTAGAGTACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTAAGATGCCACCGCATTCGG
S517	JPH3 F S517	AATGATACGGCGACCACCGAGATCTACACGCGTAAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTAAGATGCCACCGCATTCGG
S518	JPH3 F S518	AATGATACGGCGACCACCGAGATCTACACCTATTAAGACACTCTTTCCCTACACGACGCTCTTCCGATCTATAGCTCCTAAGATGCCACCGCATTCGG
S520	JPH3 F S520	AATGATACGGCGACCACCGAGATCTACACAAGGCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCTTCCTAAGATGCCACCGCATTCGG
S521	JPH3 F S521	AATGATACGGCGACCACCGAGATCTACACGAGCCTTAACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTAAGATGCCACCGCATTCGG
S522	JPH3 F S522	AATGATACGGCGACCACCGAGATCTACACTTATGCGAACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCGCTAAGATGCCACCGCATTCGG

Nextera XT v2 indexes	Primer name	Sequence P7 adaptor-Barcode-Sequencing primer binding site-Spacer-Locus specific primer
N701	JPH3 R N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGGCGGGCAGCTGCAGAG
N702	JPH3 R N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGGCGGGCAGCTGCAGAG
N703	JPH3 R N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGGCGGGCAGCTGCAGAG
N704	JPH3 R N704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG
N705	JPH3 R N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG
N706	JPH3 R N706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGGCGGGCAGCTGCAGAG
N707	JPH3 R N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGGCGGGCAGCTGCAGAG
N710	JPH3 R N710	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGGCGGGCAGCTGCAGAG
N711	JPH3 R N711	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGGCGGGCAGCTGCAGAG
N712	JPH3 R N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGGCGGGCAGCTGCAGAG
N714	JPH3 R N714	CAAGCAGAAGACGGCATACGAGATTCATGAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGGCGGGCAGCTGCAGAG
N715	JPH3 R N715	CAAGCAGAAGACGGCATACGAGATCCTGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGGCGGGCAGCTGCAGAG
N716	JPH3 R N716	CAAGCAGAAGACGGCATACGAGATTAGCGAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGGCGGGCAGCTGCAGAG
N718	JPH3 R N718	CAAGCAGAAGACGGCATACGAGATGTAGCTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG
N719	JPH3 R N719	CAAGCAGAAGACGGCATACGAGATTACTACGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGGCGGGCAGCTGCAGAG
N720	JPH3 R N720	CAAGCAGAAGACGGCATACGAGATAGGCTCCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGGCGGGCAGCTGCAGAG
N721	JPH3 R N721	CAAGCAGAAGACGGCATACGAGATGCAGCGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG
N722	JPH3 R N722	CAAGCAGAAGACGGCATACGAGATCTGCGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG
N723	JPH3 R N723	CAAGCAGAAGACGGCATACGAGATGAGCGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG
N724	JPH3 R N724	CAAGCAGAAGACGGCATACGAGATCGCTCAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG
N726	JPH3 R N726	CAAGCAGAAGACGGCATACGAGATGTCTTAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGGCGGGCAGCTGCAGAG
N727	JPH3 R N727	CAAGCAGAAGACGGCATACGAGATACTGATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGGCGGGCAGCTGCAGAG
N728	JPH3 R N728	CAAGCAGAAGACGGCATACGAGATTAGCTGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGGCGGGCAGCTGCAGAG
N729	JPH3 R N729	CAAGCAGAAGACGGCATACGAGATGACGTCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCCAGGCGGGCAGCTGCAGAG

Appendix 5. *HTT*/*JPH3* tag-SNPs used for haplotype construction.

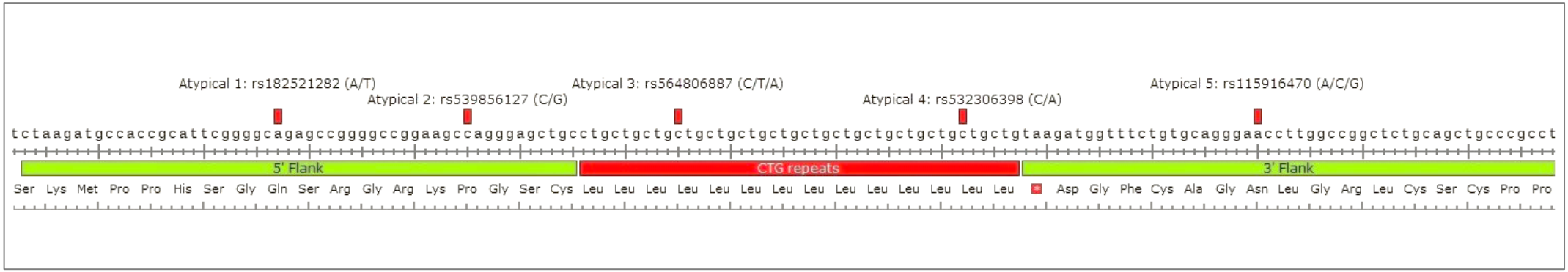


tag-SNPs	rs number	Location	Gene
1	rs2857936	3,060,583	Upstream of *HTT*
2	rs762855	3,073,068	Upstream of *HTT*
3	rs3856973	3,078,446	*HTT*
4	rs10015979	3,107,715	*HTT*
5	rs363075	3,135,947	*HTT*
6	rs363064	3,139,683	*HTT*
7	rs363102	3,147,289	*HTT*
8	rs4690073	3,158,423	*HTT*
9	rs363099	3,160,329	*HTT*
10	rs363096	3,178,294	*HTT*
11	rs2276881	3,229,934	*HTT*
12	rs362307	3,240,118	*HTT*
13	rs1006798	3,256,646	*MSANTD1* (Downstream of *HTT)*
14	rs3849253	87,090,788	Upstream of *JPH3*
15	rs8051218	87,402,857	*MAP1LC3B* (Upstream of *JPH3)*
16	rs2544635	87,589,874	Upstream of *JPH3*
17	rs3849254	87,605,987	*JPH3*
18	rs3860288	87,624,492	*JPH3*
19	rs918368	87,644,538	*JPH3*
20	rs2303771	87,755,258	*KLHDC4* (Downstream *JPH3)*

Appendix 6. Haplotype B2 frequencies in populations from the 1,000 Genomes Project.

Overall population	Abbreviation	Population	Frequency (%)
American	MXL	Mexican Ancestry from Los Angeles USA	0.0
	PUR	Puerto Ricans from Puerto Rico	3.4
	CLM	Colombians from Medellin, Colombia	0.5
	PEL	Peruvians from Lima, Peru	0.0
African	YRI	Yoruba in Ibadan, Nigeria	8.3
	LWK	Luhya in Webuye, Kenya	9.1
	GWD	Gambian in Western Divisions in the Gambia	6.6
	MSL	Mende in Sierra Leone	9.4
	ESN	Esan in Nigeria	7.6
	ASW	Americans of African Ancestry in SW USA	6.6
	ACB	African Caribbeans in Barbados	9.9
East Asian	CHB	Han Chinese in Beijing, China	0.5
	JPT	Japanese in Tokyo, Japan	1.0
	CHS	Southern Han Chinese	0.5
	CDX	Chinese Dai in Xishuangbanna, China	0.5
	KHV	Kinh in Ho Chi Minh City, Vietnam	1.0
South Asian	GIH	Gujarati Indian from Houston, Texas	0.0
	PJL	Punjabi from Lahore, Pakistan	0.0
	BEB	Bengali from Bangladesh	0.0
	STU	Sri Lankan Tamil from the UK	0.5
	ITU	Indian Telugu from the UK	0.0
European	TSI	Toscani in Italia	0.0
	FIN	Finnish in Finland	0.0
	GBR	British in England and Scotland	0.0
	IBS	Iberian Population in Spain	0.0
	CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	0.5

Appendix 7. *JPH3* repeat tract sequence and variants that cause atypical structures.



Appendix 8. *JPH3* larger haplotypes identified in disease and non-disease alleles.

Haplogroup	Haplotype	Larger haplotype	Non-disease			Disease		
			CTG range	n	%	CTG range	n	%
1: CAC	1: C**CA**C**C**	1: TCCACCA	13-22	5	0.5	-	-	-
		2: TCCACCG	5-19	23	2.4	-	-	-
		3: CCCACCG	5-20	20	2.1	-	-	-
		4: CCCACCA	14-17	8	0.8	-	-	-
	2: C**CA**T**C**	5: TCCATCA	8-33	11	1.1	44-53	8	25.0
		6: CCCATCG	8-23	34	3.5	42-50	10	31.3
		7: TCCATCG	8-25	22	2.3	45-50	9	28.1
		8: CCCATCA	8-29	18	1.9	42-51	5	15.6
	3: T**CA**T**C**	9: TTCATCG	13-18	33	3.4	-	-	-
		10: TTCATCA	13-27	35	3.7	-	-	-
		11: CTCATCA	13-19	28	2.9	-	-	-
		12: CTCATCG	8-18	52	5.4	-	-	-
	4: T**CA**C**C**	13: TTCACCA	14-19	18	1.9	-	-	-
		14: CTCACCG	11-22	29	3.0	-	-	-
		15: TTCACCG	14-17	28	2.9	-	-	-
		16: CTCACCA	11-19	15	1.6	-	-	-
2: CAT	5: C**CA**T**T**	17: TCCATTA	8-19	17	1.8	-	-	-
		18: TCCATTG	6-19	5	0.5	-	-	-
		19: CCCATTA	8-19	11	1.1	-	-	-
		20: CCCATTG	14-26	10	1.0	-	-	-
	6: T**CA**T**T**	21: CTCATTA	13-28	27	2.8	-	-	-
		22: CTCATTG	8-19	29	3.0	-	-	-
		23: TTCATTA	8-19	43	4.5	-	-	-
		24: TTCATTG	11-17	36	3.8	-	-	-
	7: C**CA**C**T**	25: CCCACTG	15	1	0.1	-	-	-
		26: TCCACTA	14	1	0.1	-	-	-
		27: TCCACTG	13-26	3	0.3	-	-	-
	8: T**CA**C**T**	28: CTCACTA	14-16	5	0.5	-	-	-
		29: CTCACTG	14-18	13	1.4	-	-	-
		30: TTCACTA	14-16	3	0.3	-	-	-
		31: TTCACTG	14-15	7	0.7	-	-	-
3: CGC	9: C**CG**C**C**	32: CCCGCCA	13-19	6	0.6	-	-	-
		33: CCCGCCG	12-26	3	0.3	-	-	-
		34: TCCGCCA	15-26	10	1.0	-	-	-
		35: TCCGCCG	6-18	22	2.3	-	-	-
	10: C**CG**T**C**	36: CCCGTCG	12-17	13	1.4	-	-	-
		37: TCCGTCG	13-18	32	3.3	-	-	-
		38: CCCGTCA	12-18	24	2.5	-	-	-
		39: TCCGTCA	13-17	37	3.9	-	-	-
	11: T**CG**C**C**	40: CTCGCCA	13-19	4	0.4	-	-	-
		41: TTCGCCG	12-15	7	0.7	-	-	-
		42: TTCGCCA	13-16	8	0.8	-	-	-
		43: CTCGCCG	13-15	2	0.2	-	-	-
	12: T**CG**T**C**	44: TTCGTCG	8-19	20	2.1	-	-	-
		45: TTCGTCA	13-17	13	1.4	-	-	-

		46: CTCGTCA	8-16	22	2.3	-	-	-
		47: CTCGTCG	12-17	12	1.3	-	-	-
4: CGT	13: C**CG**C**T**	48: TCCGCTA	12-15	3	0.3	-	-	-
		49: CCCGCTA	13-3	4	0.4	-	-	-
		50: TCCGCTG	13	1	0.1	-	-	-
		51: CCCGCTG	14	1	0.1	-	-	-
	14: C**CGTT**	52: TCCGTTG	14-15	4	0.4	-	-	-
		53: TCCGTTA	14-15	2	0.2	-	-	-
		54: CCCGTTG	13-16	6	0.6	-	-	-
		55: CCCGTTA	13-15	3	0.3	-	-	-
	15: T**CG**C**T**	56: CTCGCTG	15	1	0.1	-	-	-
	16: T**CGTT**	57: CTCGTTA	12-18	8	0.8	-	-	-
		58: TTCGTTA	14-18	7	0.7	-	-	-
		59: TTCGTTG	12-14	3	0.3	-	-	-
		60: CTCGTTG	14-16	2	0.2	-	-	-
5: AAC	17: C**AA**C**C**	61: CCAACCG	8-16	3	0.3	-	-	-
	18: C**AAT**C	62: TCAATCA	16	1	0.1	-	-	-
		63: TCAATCG	14	1	0.1	-	-	-
		64: CCAATCG	13-31	3	0.3	-	-	-
	19: T**AA**C**C**	65: CTAACCG	13-16	3	0.3	-	-	-
		66: CTAACCA	16	1	0.1	-	-	-
		67: TTAACCA	15-17	4	0.4	-	-	-
		68: TTAACCG	14-28	10	1.0	-	-	-
	20: T**AAT**C	69: TTAATCG	14-17	13	1.4	-	-	-
		70: CTAATCA	14-16	6	0.6	-	-	-
		71: CTAATCG	14-17	5	0.5	-	-	-
		72: TTAATCA	14-16	6	0.6	-	-	-
6: AAT	21: C**AA**C**T**	73: CCAACTA	16	1	0.1	-	-	-
	22: C**AATT**	74: CCAATTA	14	1	0.1	-	-	-
		75: CCAATTG	20	1	0.1	-	-	-
		76: TCAATTA	18	1	0.1	-	-	-
	23: T**AATT**	77: CTAATTA	14-17	9	0.9	-	-	-
		78: TTAATTG	14-16	7	0.7	-	-	-
		79: CTAATTG	14-16	6	0.6	-	-	-
		80: TTAATTA	15-16	3	0.3	-	-	-
	24: T**AA**C**T**	81: TTAACTA	16	1	0.1	-	-	-
		82: TTAACTG	8	1	0.1	-	-	-
7 AGC	25: T**AGCC**	83: TTAGCCG	17	1	0.1	-	-	-
Total	**25**	**83**	**958**			**32**		

Appendix 9. Plagiarism form.

600997:Levesley_Full_Thesis.pdf

ORIGINALITY REPORT

11%	8%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	orca.cf.ac.uk Internet Source	1%
2	www.ncbi.nlm.nih.gov Internet Source	<1%
3	discovery.ucl.ac.uk Internet Source	<1%
4	Submitted to University of Witwatersrand Student Paper	<1%
5	theses.gla.ac.uk Internet Source	<1%
6	open.uct.ac.za Internet Source	<1%
7	open.library.ubc.ca Internet Source	<1%
8	link.springer.com Internet Source	<1%
9	www.nature.com Internet Source	<1%

Appendix 10. Submissible article.

A *cis*-acting modifier of Huntington disease in African ancestry patients

Jessica Levesley, [1] Fiona K. Baine-Savanhu, [1] Marc Ciosi, [2] Alastair Maxwell, [2] Darren G. Monckton [2] and Amanda Krause [1, *]

Abstract

Huntington disease (HD) is a dominantly inherited neurodegenerative disorder caused by the expansion of a polyglutamine encoding CAG repeat in the *HTT* gene. Recently, it has been established that disease severity in HD is best predicted by the number of pure CAGs (Q^1) rather than total glutamines encoded (Q^1+Q^2). Along with data revealing DNA repair gene variants as *trans*-acting modifiers of HD severity, these data reveal somatic expansion of the CAG repeat as a key driver of HD onset. Using high-throughput ultra-deep DNA sequencing, we have determined the precise sequence and somatic expansion profiles of the *HTT* repeat tract of 68 HD affected and 158 unaffected African ancestry individuals. A high level of sequence diversity was observed in African ancestry individuals, with three novel likely African-specific atypical allele structures identified. The most common disease-associated allele structure in 30/68 HD individuals was defined as atypical, due to the loss of the proline encoding CCGCCA sequence (P^1=0). This disease-associated allele (Q^1-2-0-9-2) was associated with an earlier age of diagnosis of approximately 7.1 years. It occurred exclusively on the background B2 haplotype in disease-associated alleles. Notably, relative to the typical *HTT* expansions, Q^1-2-0-9-2 disease-associated alleles displayed reduced somatic expansion in blood DNA. We propose that the loss of the CCGCCA sequence occurring on haplotype B2 is an African *cis*-acting modifier, that appears to modify disease diagnosis in individuals affected with HD, through a mechanism that is not driven by somatic expansion.

[1] Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
[2] Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary & Life Sciences, University of Glasgow, Glasgow, United Kingdom.
*Correspondence: Amanda.Krause@wits.ac.za

Introduction

Huntington disease (HD [MIM: 143100]) is a dominantly inherited neurodegenerative disorder, caused by the expansion of the CAG repeat tract in exon 1 of the huntingtin (*HTT*) gene (The Huntington's Disease Collaborative Research Group, 1993). The expanded CAG size in HD patients is 36 or more repeats, and inversely correlates with the age of onset (AoO) (Roos, 2010; Walker, 2007). Thus, the longer the CAG repeat the earlier the onset of HD symptoms (The Huntington's Disease Collaborative Research Group, 1993). Expanded CAG repeats are not only unstable in the germline, with a bias toward repeat length increases in successive generations (Duyao et al., 1993), but are also unstable in somatic cells. Factors such as the length of the repeat, the age of the individual and cell type affect the repeat tract stability (Kennedy et al., 2003; Telenius et al., 1994; Veitch et al., 2007).

The reference sequence for the *HTT* repeat tract comprises a region that codes for polyglutamine with a variable number of pure CAG repeats (Q^1) and a number of additional glutamine codons (Q^2) *i.e.* CAACAG. This is followed by a region coding for polyproline which includes a number of proline codons upstream of the pure CCG repeat tract (P^1) *i.e.* CCGCCA, a variable number of CCG repeats (P^2) and downstream CCT codons (P^3). The *HTT* trinucleotide repeat has been reported to be a hotspot for mutations (Ciosi et al., 2019b; GeM-HD Consortium, 2019; Wright et al., 2019; Yu et al., 2000). The mutations that have been identified alter the structure of the sequence relative to the typical reference allele structure, resulting in apparent duplication or loss of part of the intervening sequence and the downstream CCT repeats (Figure 1) (Gellera et al., 1996; Goldberg et al., 1995; Yu et al., 2000). In other repeat expansion disorders such as spinocerebellar ataxia type 1, myotonic dystrophy type 1 and fragile X syndrome, repeat stability and disease severity are affected by mutations within the repeat tract (*i.e.* atypical allele structure) (Kraus-Perrotta and Lagalwar, 2016; Liu et al., 2014). Mutation within the *HTT* repeat tract may have a similar effect in HD.

A recent study of the *HTT* repeat tract in over 800 European HD patients revealed that atypical allele structures are more frequent than previously thought (~8% of non-disease associated alleles and ~3% of disease-associated alleles) (Ciosi et al., 2019). Various studies have shown that HD severity is best predicted by the number of pure CAGs (Q^1) rather than the total length of glutamines encoded ($Q^1 + Q^2$) (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). Given that somatic expansion of the CAG repeat is also best predicted by the number of pure CAGs and coupled with the observation that CAG length independent variation in age at onset is associated with DNA repair gene variants, these data confirm that somatic expansion is a key driver of HD onset (Ciosi et al., 2019; Gem-HD Consortium, 2019; Wright et al., 2019).

Although HD has been reported worldwide, there are distinct geographic differences in prevalence, with the lowest rates in African populations and those with African ancestry (Pringsheim et al., 2012; Warby et al., 2011). These differences are reflected in South Africa, where HD has been reported in three different groups; white, coloured and black African but at varying prevalence estimates (8.27, 2.42 and 0.75 per 100,000 individuals respectively) (Baine, Krause and Greenberg, 2016).

In general, there is a greater amount of genetic diversity present in individuals of African ancestry (Choudhury et al., 2017) Specifically at the HTT locus, a large number of haplotypes have been previously defined, and among South African individuals a specific C haplotype (C-SA) has been identified. We therefore hypothesised that in line with the broad genetic heterogeneity observed in Africa across many loci (Choudhury et al., 2017), we except more *HTT* repeat tract diversity in disease-associated alleles. This diversity may similarly impact on the HD phenotype as seen in European patients and could provide insights into the role of sequence diversity in modifying disease. The hypothesis was tested using a high-throughput ultra-deep DNA sequencing assay for the *HTT* repeat tract (Ciosi et al., 2018). This allowed for the investigation of sequence variants in this complex region, the background haplotypes

they occur on, and the characterisation of somatic expansion, as potential genetic modifiers of the HD phenotype in individuals of African ancestry.

Subjects and Methods

Subjects

Blood DNA samples were sourced from the archives of the Division of Human Genetics, University of the Witwatersrand (Wits) and the National Health Laboratory Services (NHLS) in Johannesburg, South Africa. The study population comprised 68 individuals affected with HD and 158 individuals unaffected with HD of African ancestry (disease-associated alleles: 68; non-disease associated alleles: 384). The individuals affected with HD were originally referred for molecular confirmation of disease status following a clinical diagnosis of HD. They comprised 42 females and 26 males, with an age of diagnosis ranging between 22 and 77 years; addition patient information is shown in the supplementary information (Supplementary Table 1). Ethical approval for this study was obtained from the Wits Human Research Ethics Committee (Medical) (certificate numbers: M1704130, sub-study M110443).

HTT repeat tract sequence diversity

The *HTT* repeat tract sequencing followed an established high-throughput sequencing protocol developed to characterise the repeat tract precisely (Ciosi et al., 2018; Ciosi et al., 2019). Following sequencing on a MiSeq platform, genotyping was carried out using a Galaxy analysis pipeline and ScaleHD (v0.251). Typical alleles were defined as those that conform to the reference allele structure (locus reference genomic: LRG_763) comprising a polyglutamine coding region with a variable number of CAG repeats (Q^1) and the glutamine codons after CAG repeat (Q^2) *i.e.* CAACAG; followed by a polyproline encoding region made up of the proline codons before the CCG repeats (P^1) *i.e.* CCGCCA, a variable number of CCG repeats (P^2) and two downstream CCT codons (P^3). The *HTT* repeat tract reference sequence is shown

in Figure 1. The atypical alleles were defined based on deviations from this reference sequence resulting from variants within the repeat tract at the Q^2, P^1 and P^3 regions.

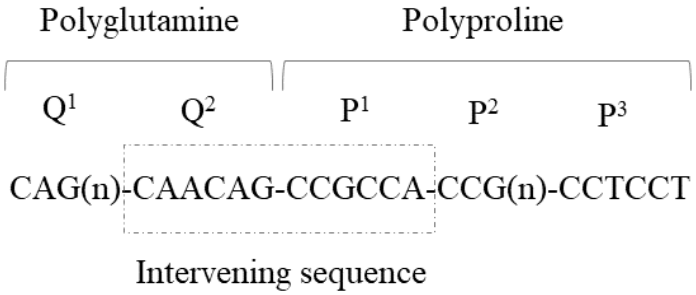


Figure 1. The *HTT* repeat tract reference sequence is composed of a region that codes for polyglutamine with a variable number of CAG repeats (Q^1) and downstream glutamine codons (Q^2). This is followed by a region coding for polyproline which includes proline codons before the CCG repeats (P^1), a variable number of CCG (P^2) repeats and downstream CCT codons (P^3). The CAACAG-CCGCCA is the intervening sequence between the CAG/CCG repeats.

HTT background haplotypes

The allele structures were investigated in the context of background haplotypes for the *HTT* locus spanning ~196 063 kb on chromosome 4. Thirteen tag single nucleotide polymorphisms (tSNPs) were selected from previously studied haplotype SNPs, to define haplogroups A, B and C, and the South African specific haplogroup variant C-South Africa (C-SA) (Supplementary Table 2) (Baine et al., 2013). The SNPs were genotyped using a MassARRAY® System from Agena Bioscience™ and haplotypes constructed using manual and statistical phasing. Manual phasing was achieved using Mendelian principles for related individuals, homozygous genotypes and repeat tract associations. Statistical phasing was performed using PHASE (*v*2.1.1) which employs a Bayesian inference model. Two samples from the sequence diversity analysis were excluded due to unsuccessful SNP genotyping. The LDhap tool from the LDlink suite was used to derive haplotype frequencies for the tSNPs in

the 1000 Genome Project populations (Machiela and Chanock, 2015) used to define the most common disease-associated haplotype.

Quantification of HTT somatic expansions

The ratio of somatic expansions $\left(\frac{number\ of\ somatic\ expansion\ products}{number\ of\ progenitor\ allele\ products}\right)$ was quantified from the MiSeq read count distributions as described previously (Ciosi et al., 2019). Aligning the sequenced reads to synthetic references with different CAG repeat sizes allowed for quantification of somatic expansion (N+1 to 10). The somatic expansion score was then calculated as the residuals of the linear model of the ratio of somatic expansion when the CAG repeat size and the age at sampling and their interaction were considered (Warby et al., 2011).

Statistical analysis

Potential genetic modifiers of HD were investigated with multiple linear models using the programme R (*v3.4.3*). The lm function was used to determine associations between the HD phenotype and various explanatory variables: the *HTT* repeat tract, background haplotypes and somatic expansion score. When studying the HD phenotype, the age at onset of motor symptoms is the most well defined and frequently used measure of disease severity (Lee et al., 2012). However, in our patients, less than 50% had AoO information. The age of diagnosis (AoD) was used as an acceptable proxy for AoO (strong correlation between AoO and AoD, $R^2 = 0.68$, Supplementary Figure 2). AoD analyses were only conducted on the subset of patients (64) carrying CAG repeat sizes between 39 and 52 repeats, as repeats >53 CAG repeats violate linear model assumptions (Lee et al., 2012). In the multiple linear models the reference for the allele structures was the Q^1-2-2-P^2-2 allele structure and for the haplotypes was the B2 haplotype. To determine if the variation in the AoD was better explained by allele structure or the background haplotype, the goodness of fit of the R-squared of the Q^1-2-0-9-2 allele structure model and the B2 haplotype model in 5000 bootstrapped samples was assessed.

Results

The most common sequence structure on HD disease-associated alleles of African ancestry is characterised by the loss of the CCGCCA sequence.

A total of 226 samples from individuals of African ancestry, 68 affected with HD and 158 unaffected (68 disease-associated alleles and 384 non-disease-associated alleles) were sequenced and genotyped. Seventeen different allele structures were identified and defined as either typical or atypical (Table 1). The eight defined as typical had a variable number of CAG repeats, and a variable number of CCGs ranging from 6 to 13 repeats. The other nine structures were defined as atypical due to variants resulting in an apparent loss or duplication of the intervening sequences CAACAG (Q^2 = 0 or 4), CCGCCA (P^1 = 0 or 4) and/or an addition to the two downstream CCTs (P^3 = 3). Of the 17 allele structures defined, three (one typical and two atypical) were unique to this study. Schematics for the disease and non-disease allele structures are shown in the supplementary information (Supplementary Figure 1).

The most common disease-associated allele Q^1-2-0-9-2 (30/68= 44.1%) had an atypical structure as defined by the absence of the CCGCCA sequence (P^2 = 0). The high frequency of atypical alleles contrasts with European populations where the most common allele structure in both the disease and non-disease associated alleles is typical. Although, also present in the unaffected individuals, it occurred only on a small percentage of non-disease-associated alleles (~ 7%). The typical allele structures were more frequent among non-disease-associated alleles (with eight defined) than among disease alleles (with three defined) (Fisher exact test, $p = 2.25 \times 10^{-2}$, 235/384 = 61.2%, 95% CI = 63.0 to 65.2% vs 31/68 = 45.6%, 95% CI = 45.9 to 46.5%). The most common non-disease-associated allele Q^1-2-2-7-2 (99/384, 25.8%) had a typical structure, with variability occurring only in the length of the CAG repeat.

When comparing disease and non-disease-associated alleles, two typical structures (Q^1-2-2-7-2 and Q^1-2-2-10-2) occurred more frequently in the non-disease-associated alleles (10/68 < 99/384, Fisher exact test $p = 1.29 \times 10^{-2}$ and 20/68 < 71/384, Fisher exact test $p = 4.831 \times 10^{-2}$). Among the atypical structures, five differed significantly between the disease and non-disease-associated alleles. Three structures (Q^1-2-2-6-3, Q^1-2-0-9-2 and Q^1-0-0-9-2) were more frequent in disease-associated alleles (4/68 > 2/384 Fisher exact test $p = 5.587 \times 10^{-3}$, 30/68 > 27/384 Fisher exact test $p = 3.119 \times 10^{-13}$ and 2/68 > 0/365 Fisher exact test $p = 2.235 \times 10^{-2}$) and two (Q^1-4-2-7-3 and Q^1-2-2-9-3) were more frequent in non-disease-associated alleles (0/68 < 22/362 Fisher exact test $p = 5.878 \times 10^{-2}$, 0/68 < 92/384 Fisher exact test $p < 9.142 \times 10^{-8}$) (Table 1).

Table 1. Summary of the African *HTT* disease and non-disease associated alleles.

Allele types	**Allele structures Q^1-Q^2-P^1-P^2-P^3**	***HTT* repeat tract**						**Allele frequency**			
		CAG		CAACAG	CCGCCA	CCG	CCT	Non-disease 384		Disease 68	
								n	%	n	%
Typical alleles	Q^1-2-2-6-2	14-17	-	2	2	6	2	11	2.86	-	-
	Q^1-2-2-7-2	15-28	41-55	2	2	7	2	99	25.78	10	14.71
	Q^1-2-2-8-2	17	-	2	2	8	2	5	1.30	-	-
	Q^1-2-2-9-2	15-28	40	2	2	9	2	29	7.55	1	1.47
	Q^1-2-2-10-2	11-20	40-54	2	2	10	2	71	18.49	20	29.41
	Q^1-2-2-11-2	12-21	-	2	2	11	2	18	4.69	-	-
	Q^1-2-2-12-2	17	-	2	2	12	2	1	0.26	-	-
	*Q^1-2-2-13-2	17	-	2	2	13	2	1	0.26	-	-
Typical alleles subtotal								235	61.2%	31	45.6%
Atypical alleles	Q^1-2-2-4-3	23	-	2	2	4	3	1	0.26	-	-
	Q^1-2-2-6-3	15-23	42-44	2	2	6	3	2	0.52	4	5.88
	Q^1-2-2-9-3	12-21	-	2	2	9	3	92	23.96	-	-
	Q^1-2-2-10-3	16	-	2	2	10	3	1	0.26	-	-
	*$Q^1$4-2-4-3	-	42	4	2	4	3	-	-	1	1.47
	Q^1-4-2-7-3	14-19	-	4	2	7	3	22	5.73	-	-
	*Q^1-4-2-10-2	16-19	-	4	2	10	2	4	1.04	-	-
	Q^1-2-0-9-2	16-32	40-58	2	0	9	2	27	7.03	30	44.12
	Q^1-0-0-9-2	-	39-46	0	0	9	2	-	-	2	2.94
Atypical alleles subtotal								149	38.8%	37	54.4%

a The novel allele structures which are unique to this study are indicated with an asterisk (*).

The comparison of African alleles with the European alleles previously described (among 807 individuals carrying an HD disease-associated allele) (Ciosi et al., 2019) revealed differences in the allele structures defined and allele frequencies. Among European *HTT* alleles, 92.7% of the non-disease and 97.3% of disease-associated alleles had a typical allele structure; (Ciosi et al., 2019) compared to the African ancestry *HTT* alleles where only 61.2% of non-disease alleles were typical and 45.6% of disease alleles were atypical (Non-disease: 235/384 >748/807 Fisher exact test $p < 2.2 \times 10^{-16}$, Disease alleles: 31/68 >785/807 Fisher exact test $p < 2.2 \times 10^{-16}$). The most common allele structure Q^1-2-2-7-2 was the same for both non-disease and disease alleles in European individuals. This typical allele structure was also the most common in African non-disease-associated alleles. However, in the African disease associated alleles the most common allele was the atypical structure Q^1-2-0-9-2, which is rare (0.2%) among European individuals carrying HD alleles (Ciosi et al., 2019).

The loss of the CCGCCA sequence in disease alleles occurred exclusively on haplotype B2.

Background haplotypes were constructed for 224 individuals with African ancestry (67 affected and 157 unaffected with HD), shown in Table 2. The largest proportion of non-disease-associated alleles occurred on haplogroup C (159/381 = 41.7%) with haplotype C5 being the most common (94/381 = 24.8%). The allele structures occurring on haplotype C5 were predominantly typical in structure however, atypical structures (Q^1-2-2-10-3, Q^1-2-0-9-2, and Q^1-4-2-10-2) were also present. On disease alleles, the largest proportion of allele structures occurred on haplogroup B (33/67= 49.3%), with haplotype B2 being the most common (32/67 = 47.8%). The allele structures occurring on haplotype B2 were all atypical (Q^1-0-0-9-2, Q^1-4-2-4-3, and Q^1-2-0-9-2). The most common disease-associated allele structure Q^1-2-0-9-2 was found exclusively on haplotype B2.

Table 2. Summary of the background haplogroups and haplotypes associated with disease and non-disease allele structures.

Haplogroups	Haplotypes	Allele structures	Non-disease		Disease	
			n	%	n	%
A	A2a	Q^1-2-2-7-2	-	-	1	1.5
	A2b	Q^1-2-2-7-2	13	3.4	1	1.5
	A4a	Q^1-2-2-7-2	5	1.3	3	4.5
		Q^1-2-2-12-2	1	0.3	-	-
		Q^1-2-2-13-2	1	0.3	-	-
	A4b	Q^1-2-2-7-2	2	0.5	5	7.5
	A6	Q^1-2-2-7-2	34	8.9	-	-
B	B1	Q^1-2-2-9-2	1	0.3	1	1.5
	B2	Q^1-2-0-9-2	25	6.6	29	43.3
		Q^1-0-0-9-2	-	-	2	3.0
		Q^1-4-2-4-3	-	-	1	1.5
C	C2	Q^1-4-2-7-3	21	5.5	-	-
		Q^1-2-2-8-2	5	1.3	-	-
	C4	Q^1-2-2-9-2	21	5.5	1	1.5
	C4c	Q^1-2-2-6-2	11	2.9	-	-
	C5	Q^1-2-2-10-2	69	18.1	19	28.4
		Q^1-2-2-11-2	18	4.7	-	-
		Q^1-2-2-10-3	1	0.3	-	-
		Q^1-2-0-9-2	2	0.5	-	-
		Q^1-4-2-10-2	4	1.0	-	-
	C8	Q^1-2-2-9-2	7	1.8	-	-
***C-SA**	C3	Q^1-2-2-10-2	1	0.3	-	-
		Q^1-2-2-9-3	90	23.6	-	-
	C9	Q^1-2-2-6-3	2	0.5	4	6.0
		Q^1-2-2-7-2	7	1.8	-	-
	C10	Q^1-2-2-4-3	1	0.3	-	-
Other	O	Q^1-2-2-7-2	36	9.4	-	-
		Q^1-2-2-9-3	2	0.5	-	-
		Q^1-2-2-10-2	1	0.3	-	-
Total			**381**		**67**	

a *C-SA is a C haplogroup variant that is specific to South African individuals (Baine et al., 2013).

An increased haplotype frequency of the tSNP alleles which define the African B2 haplotype was also observed in the African general population controls in the 1000 Genome Project Populations (Figure 2). The highest frequencies were observed in the following African and African-derived populations; African Caribbean's in Barbados - ACB (9.9%), Mende in Sierra Leone - MSL (9.4%), Luhya in Webuye, Kenya -LWK (9.1%), Yoruba in Ibadan, Nigeria - YRI (8.3%), Esan in Nigeria - ESN (7.6%), Gambian in Western Divisions in the Gambia - GWD (6.6%) and Americans of African Ancestry in SW USA - ASW (6.6%).

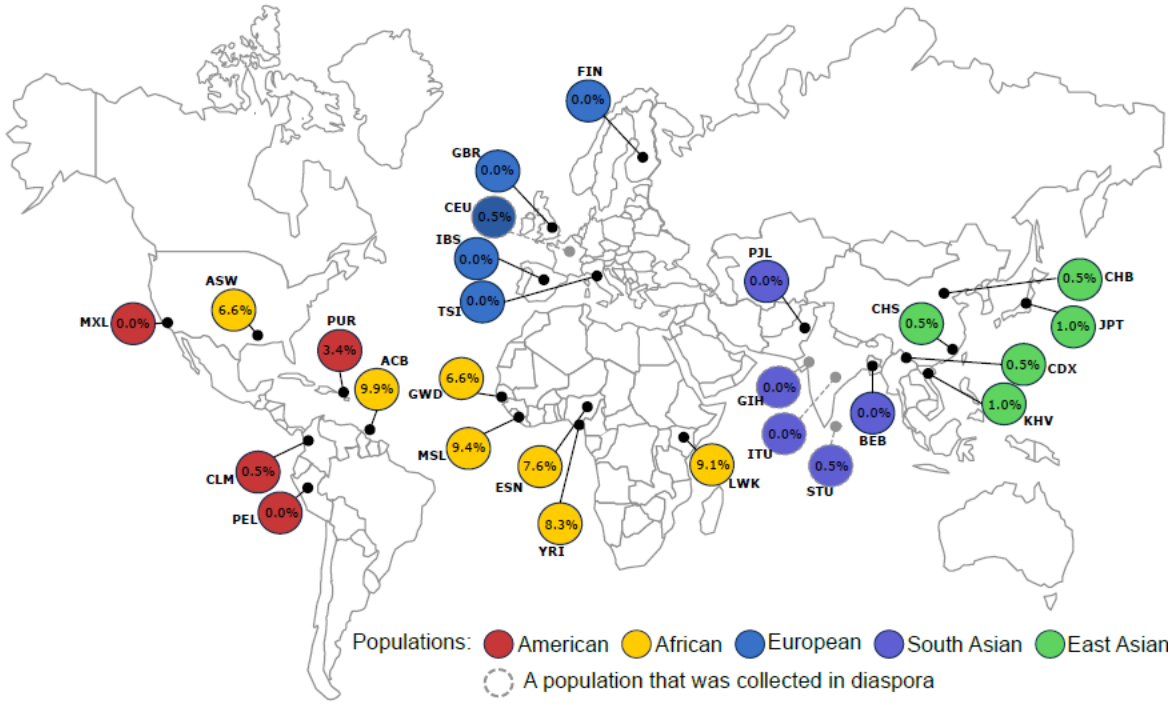


Figure 2. The frequency of the B2 haplotype in the 1000 Genome Project populations. Constructed using the LDhap tool from the LDlink suite.(Machiela and Chanock, 2015)

Disease-associated alleles with the loss of CCGCCA sequence had the lowest somatic expansion score.

As expected, the CAG repeat size and the age at sampling was shown to have a significant association ($p = 7.81 \times 10^{-16}$) with the ratio of somatic CAG expansions observed in blood DNA (Supplementary Table 3 Model 1). A larger effect was observed for the CAG repeat size, as every additional CAG repeat resulted in an increase to the ratio of somatic expansions by 0.131.

The effect of the disease allele structures revealed that the ratio of somatic expansions was associated with the following disease-associated allele structures, Q^1-0-0-9-2 ($p = 7.66 \times 10^{-4}$), Q^1-2-0-9-2 ($p = 1.02 \times 10^{-5}$) and Q^1-2-2-6-3 ($p = 1.14 \times 10^{-2}$) were significantly lower than the typical alleles Q^1-2-2-P^2-2 (Supplementary Figure 4 and Supplementary Table 3, Model 2).

Patients with disease alleles Q^1-0-0-9-2, Q^1-2-0-9-2 and Q^1-2-2-6-3 were shown to have a decrease in the ratio of somatic expansions of 0.257, 0.168 and 0.151, respectively.

The typical disease alleles Q^1-2-2-P^2-2 were associated with the highest ratio of somatic expansion with an estimated marginal mean of 0.598 (95% CI = 0.535 to 0.669). While the lowest was associated with the atypical allele structures Q^1-2-0-9-2: 0.418 (95% CI = 0.380 to 0.460) and Q^1-0-0-9-2: 0.323 (95% CI = 0.227 to 0.458), both alleles are characterised by loss of the CCGCCA sequence. The remaining two disease allele structures, Q^1-2-2-6-3 and Q^1-4-2-4-3 had estimated marginal means of 0.436 (95% CI = 0.348 to 0.545) and 0.443 (95% CI = 0.288 to 0.680) respectively.

Potential modifiers of the HD phenotype

The loss of the CCGCCA sequence was associated with a significantly earlier AoD.

A significant negative association ($p = 2.545 \times 10^{-14}$) was detected between the CAG repeat size and the AoD, accounting for approximately 60% of the variation in the HD phenotype (Supplementary Figure 3). Two components within the *HTT* repeat tract were shown to have a significant effect on the HD phenotype (Supplementary Table 4, Model 1), the CAG repeat size (AoD effect in years= 2.9-year earlier, $p = 6.11 \times 10^{-12}$) and the CCGCCA sequence (AoD effect in years = 4.0-year earlier for the absence of CCGCCA, $p = 7.34 \times 10^{-4}$). Each additional CAG repeat and the loss of the CCGCCA sequence resulted in an earlier AoD in individuals affected with HD.

The most common disease allele structure Q^1-2-0-9-2 was shown to have a significant negative effect on the HD phenotype in comparison to the other disease allele structures (R^2=0.6097, $p = 1.355 \times 10^{-10}$) (Figure 2). The Q^1-2-0-9-2 disease allele structure was also the only allele structure shown to have a significant negative effect on the HD phenotype in comparison to

the Q^1-2-2-P^2-2 grouped typical allele structures (AoD effect in years = 7.1-year earlier, $p = 8.15 \times 10^{-4}$) (Supplementary Table 4, Model 2). Thus, patients carrying the Q^1-2-0-9-2 disease allele structure had a significantly earlier AoD than patients carrying the Q^1-2-2-P^2-2 allele structures.

The estimated marginal mean for the AoD (in years) for each allele structure were as follows; Q^1-2-0-9-2: 45.5 (95% CI = 43.0 to 48.2), Q^1-0-0-9-2: 47.1 (95% CI = 38.1 to 58.1), Q^1-4-2-4-3: 50.4 (95% CI = 37.4 to 67.9), Q^1-2-2-P^2-2: 53.0 (95% CI = 49.9 to 56.3) and Q^1-2-2-6-3: 56.9 (95% CI = 48.9 to 66.0).

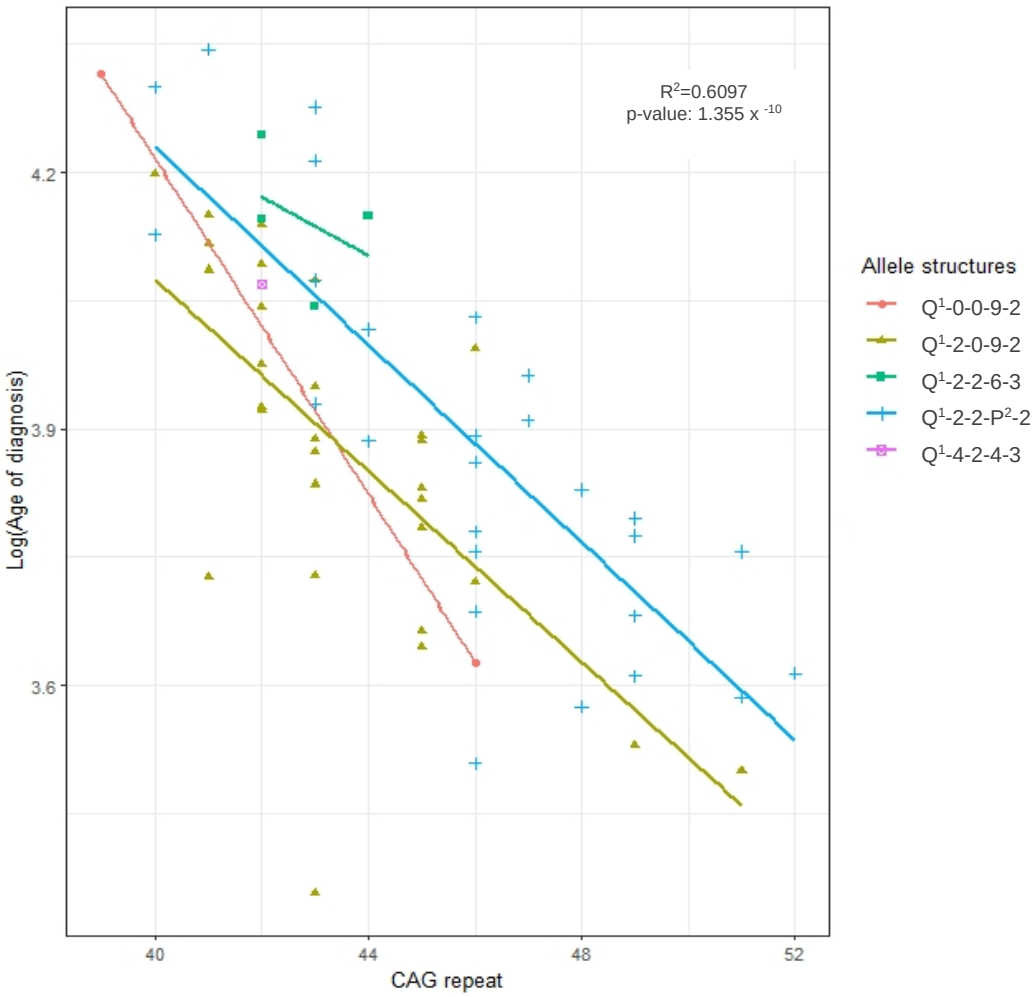


Figure 3. Linear regression model testing the association between the HD-associated CAG repeats and log transformed (AoD) for each of the disease allele structures.

Patients with allele structures on haplotype B2 had the earliest AoD.

Haplogroup A, C and haplogroup variant C-SA were shown to have a significant positive effect on the HD phenotype (delayed the age of diagnosis) when compared with the most common haplogroup B (Supplementary Table 4, Model 3). Patients with allele structures occurring on haplogroup B had a significantly earlier AoD compared with haplogroup A (8.6 years, $p = 0.0137$), haplogroup C (6.2 years, $p = 0.0224$) and haplogroup variant C-SA (11.8 years, $p = 0.0124$).

The estimated marginal mean of the AoD (in years) for the haplotypes were as follows; A2a: 47.2 (95% CI = 35.0 to 63.4), A2b: 53.8 (95% CI = 39.9 to 72.4), A4a: 60.6 (95% CI = 48.5 to 75.7), A4b: 53.6 (95% CI = 46.9 to 61.3), B1: 65.5 (95% CI = 48.6 to 88.2), B2: 45.5 95% CI = 43.0 to 48.1), C4: 46.8 (95% CI = 34.5 to 63.4), C5: 52.5 (95% CI = 48.5-56.8) and C9: 56.5 (95% CI = 48.6 to 65.7). Patients with alleles structures on haplotype B2 had a significantly earlier AoD when compared with the other significant haplotypes; A4a (16.8 years, $p = 0.0177$), A4b (9.1 years, $p = 0.0290$), B1 (22.3 years, $p = 0.0188$), C5 (7.8 years, $p = 0.00676$) and C9 (12.3 years, $p = 0.00785$) (Supplementary Table 3, Model 4).

The assessment of the allele structure Q^1-0-2-9-2 compared to haplotype B2 as a better explanation of the earlier AoD, revealed neither to have more of a significant effect (Supplementary Figure 5). The effect to the AoD cannot be separated out as the 95% confidence interval of the difference between the R-square of the two models contains zero. The haplotype is thus not better associated with the AoD than the Q^1-2-0-9-2 allele structure.

Somatic expansion had a significant effect on AoD.

After correcting for the effects of CAG length (Q^1) and age at sampling, residual variation in the somatic expansion ratio, the somatic expansion score, was shown to have a significant

negative association (p = 0.0120) with the HD phenotype (AoD effect in years = 10.6-year earlier per 1-unit increase in somatic expansion score) (Supplementary Table 4, Model 5) *i.e.* individuals with higher individual-specific rates of somatic expansion in blood DNA have an earlier AoD than expected.

Discussion

This study set out to characterise the *HTT* repeat tract sequence diversity in African ancestry disease and non-disease-associated alleles and ultimately assess potential genetic modifiers of the HD phenotype.

A large amount of sequence diversity was observed in the *HTT* repeat tract in both non-disease and disease-associated alleles. Of the 17 different allele structures identified, eight were defined as typical (variation only in the number of CAG/CCG repeats); while nine were atypical (variation present throughout the *HTT* repeat tract). Less variation was identified in the non-disease-associated alleles, with typical allele structures being more frequent (Table 1); while atypical allele structures were more frequently observed in disease-associated alleles.

Within non-disease-associated alleles, the typical allele structure Q^1-2-2-7-2 was the most common (N=99/384, Table 1). This allele structure has been previously shown to be the most common in both European ancestry non-disease and disease-associated associated alleles (Ciosi et al., 2019). In contrast, the atypical allele structure Q^1-2-0-9-2, characterised by loss of the CCGCCA sequence (intervening proline) was the most common African disease associated allele N=30/68, Table 1). Thus, revealing that in our cohort of African ancestry HD patients have a different main disease-associated allele compared to European populations characterised previously. Although this allele structure has been previously identified in European ancestry individuals, it is very rare especially among individuals affected with HD (non-disease alleles: 32/807= 4%; disease alleles: 1/807= 0.1%) (Ciosi et al., 2018).

Novel structures were identified in the disease and non- disease (Ciosi et al., 2018) alleles studied here, with three of the 17 allele structures being unique to this study. This is possibly due to these allele structures being very rare in previously studied populations or, more likely, specific to African ancestry individuals. Through studying individuals of African ancestry, novel and atypical allele structures have been characterised despite the limited cohort size. The differences between atypical allele frequencies in African alleles and recently reported European alleles (European atypical non-disease: ~3%; disease: ~8% vs African ancestry non-disease: ~39%; disease: ~54%) (Ciosi et al., 2019), highlights the importance of research across diverse populations to better understand the full range of diversity, especially for African populations where a greater amount of genetic diversity is present.

Analysis of the broader *HTT* locus in individuals of African ancestry revealed that the largest proportion of non-disease-associated alleles occurred on haplogroup C (haplotype C5), while the largest proportion of disease-associated alleles occurred on haplogroup B (haplotype B2). The comparison of the European ancestry haplotypes revealed the largest proportion of non-disease-associated alleles occur on haplogroup C, while the largest proportion of disease-associated alleles occurred on haplogroup A (Warby et al., 2011).

The most common disease-associated allele structure (Q^1-2-0-9-2) in our African ancestry cohort, characterised by the loss of the CCGCCA sequence, occurred exclusively on haplotype B2. Although, haplotype B2 has been identified in individuals of European ancestry, it is rare and differs to the African haplotype at a single tag-SNP rs762855 where an A allele is found on European chromosomes relative to the G allele on African chromosomes (Kay et al., 2018) Previous research by Baine *et al.,* found the African B2 haplotype exclusively in African ancestry or admixed individuals (Baine et al., 2013). Although rs762855 was not included in the subset of SNPs used to define the haplotypes in this study, the B2 haplotype we have defined is likely the African-specific haplotype, future work will however be conducted for

confirmation. We have also identified the presence of haplotype variants A2a and A2b, in two of our African patients suggesting that although rare, European 'high-risk' haplotypes are present in African ancestry individuals, potentially as a result of low-level admixture.

Recent data have confirmed somatic expansion of the *HTT* CAG repeat as a key driver of HD severity (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). Indeed, in European ancestry HD patients, individual-specific rates of somatic expansion in blood DNA are inversely associated with age at onset, and directly correlated with disease progression (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). Here, we have replicated this effect in an African ancestry HD cohort demonstrating an inverse association between individual-specific rates of somatic expansion in blood DNA and AoD as a proxy for age at onset. Whether individual-specific rates of somatic expansion in African HD patients are driven by the same set of DNA repair gene variants as observed in European populations (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019), is yet to be determined. However, given the increased genetic diversity observed in African populations, it seems likely that additional African-specific genetic variants will also be operating, providing a potential route to identify additional modifier genes, not accessible to studies of European patients, and which may have potential as novel therapeutic targets.

It has also recently been determined that HD severity is best explained by the length of the pure CAG repeat tract (Q^1) and not by the length of the polyglutamine tract encoded ($Q^1 + Q^2$) (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019). Since the degree of somatic expansion is also best predicted by pure CAG length (Q^1) (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019), these data suggest that the potential for somatic expansion is more important than the number of glutamines encoded in the inherited allele. However, although after correcting for pure CAG (Q^1) the number of additional glutamine encoding codons (Q^2) had no detectable impact on somatic expansion profiles in the blood

DNA of European HD individuals, counter to that predicted under a polyglutamine toxicity model loss ($Q^2 = 0$) of the CAACAG sequence was still associated with an earlier age at onset than expected based on their pure CAG length (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019).

Only two disease-associated alleles lacking the CAACAG sequence ($Q^2 = 0$) were detected in our African ancestry cohort, precluding an assessment of the impact of this structure on HD severity. Rather, we determined that the most common disease-associated allele structure in this study was characterised by loss of the CCGCCA sequence ($P^1 = 0$) and patients carrying these alleles had an earlier AoD by 4.0 years compared to patients with the CCGCCA sequence ($P^1 = 2$). Significant effects were also identified when comparing the most common disease allele structures Q^1-2-0-9-2 with the reference allele, Q^1-2-2-P^2-2, with patients having an earlier AoD by 7.1 years.

The loss of the CCGCCA sequence (intervening proline) is thus proposed as a *cis*-acting modifier of the HD phenotype in African patients. As the number of prolines encoded by the variable CCG repeat (P^2) has not been revealed as a modifier of HD onset (Panegyres et al., 2006), it seems unlikely that the phenotypic consequence of loss of the CCGCCA sequence is mediated at the protein level. Since inter-locus CAG repeat instability is modified by the flanking sequence (Brock, Anderson and Monckton, 1999), it seems plausible that polymorphisms in the flanking DNA could mediate changes in somatic instability.

Our data support this model in that loss of the CCGCAA sequence was associated with altered somatic expansion scores. However, contrary to the prediction that higher GC content would increase somatic expansion rates, and the expectation that a disease accelerating variant would increase somatic expansion rates, we detected a decreased somatic expansion rate in the blood DNA of individuals with loss of the CCGCCA sequence on their disease-associated allele.

Thus, unless this effect on somatic expansion is reversed in the critical brain regions, we speculate that the disease accelerating effect of the CCGCCA cassette is mediated by some other pathway. As we have previously speculated for the CAG-independent disease modifying effect of the CAACAG sequence (Q^2) (Ciosi et al., 2019a; GeM-HD Consortium, 2019; Wright et al., 2019), these could include effects on the *HTT* transcript so as to alter folding, splicing, canonical or RAN translation (Bañez-Coronel et al., 2015; Busan and Weeks, 2013; Krauss et al., 2013; Neueder et al., 2017). Alternatively, it is possible there is an effect mediated by a linked variant (Becanović et al., 2015).

Indeed, the analysis of background haplotypes revealed that all the loss of the CCGCCA sequence disease-associated alleles were present on the same haplotype B2, and that this haplotype was negatively associated with the HD phenotype, when compared with haplotypes A4a, A4b, B1, C5 and C9. Haplotype B2 can thus be designated a 'high-risk' haplotype in African ancestry individuals due to its virtually complete association with the loss of the CCGCCA sequence disease-associated alleles. The assessment of the allele structure Q^1-0-2-9-2 compared to haplotype B2 as a better explanation of the earlier AoD, revealed neither to have more of a significant effect. Thus, at this stage, we cannot distinguish between a direct effect or the effect of a *cis*-acting modifier in linkage disequilibrium. An effect of loss of the CCGCCA sequence on disease-associated alleles has not been seen in European ancestry patients; however, minimal variation has been identified in the P^1 proline sequence in such populations (Ciosi et al., 2019; GeM-HD Consortium, 2019; Wright et al., 2019).

In conclusion, this is the first study to describe allele sequence diversity within the *HTT* repeat tract in African ancestry individuals. Substantial sequence diversity, evidenced by the presence of predominantly atypical allele structures, is reported. Three novel allele structures were defined, which are either very rare in other populations or are African-specific. The most common HD disease-associated allele structure (Q^1-2-0-9-2) in our patients is atypical,

characterized by loss of the CCGCCA sequence, and results in an earlier age of diagnosis by 7.1 years. This disease allele structure occurs exclusively on haplotype B2, which we propose as a 'high-risk' haplotype in African ancestry patients. Interestingly, this disease-associated allele structure had the lowest ratio of somatic expansion in blood DNA, suggesting that the disease accelerating effect of this haplotype is not mediated by an increase in the frequency of somatic expansions. We propose the loss of the intervening proline sequence occurring on haplotype B2 is a *cis*-acting modifier of HD in our African ancestry patients that accelerates disease onset through a mechanism that is not driven by somatic instability.

Supplementary figures:

Supplementary Figure 1. The disease and non-disease allele structures defined in the African ancestry individuals.

Supplementary Figure 2. Linear regression model testing log transformed AoD as HD phenotype.

Supplementary Figure 3. Association between the CAG repeat size and log transformed AoD.

Supplementary Figure 4. The association between the somatic expansion estimates and CAG repeat size, grouped by disease allele structures.

Supplementary Figure 5. The goodness of fit comparison of the Q^1-2-0-9-2 allele structure and B2 haplotypes model R-square via nonparametric bootstrapping.

Supplementary tables:

Supplementary Table 1. Individuals affected with HD demographic information.

Supplementary Table 2. The tSNPs used to construct the *HTT* haplotypes.

Supplementary Table 3. Linear models testing the association between the ratio of somatic expansion and various explanatory variables.

- Model 1. Association between the ratio of somatic expansion and the CAG repeat size and age at sampling
- Model 2. Association between the ratio of somatic expansion and the *HTT* disease allele structures

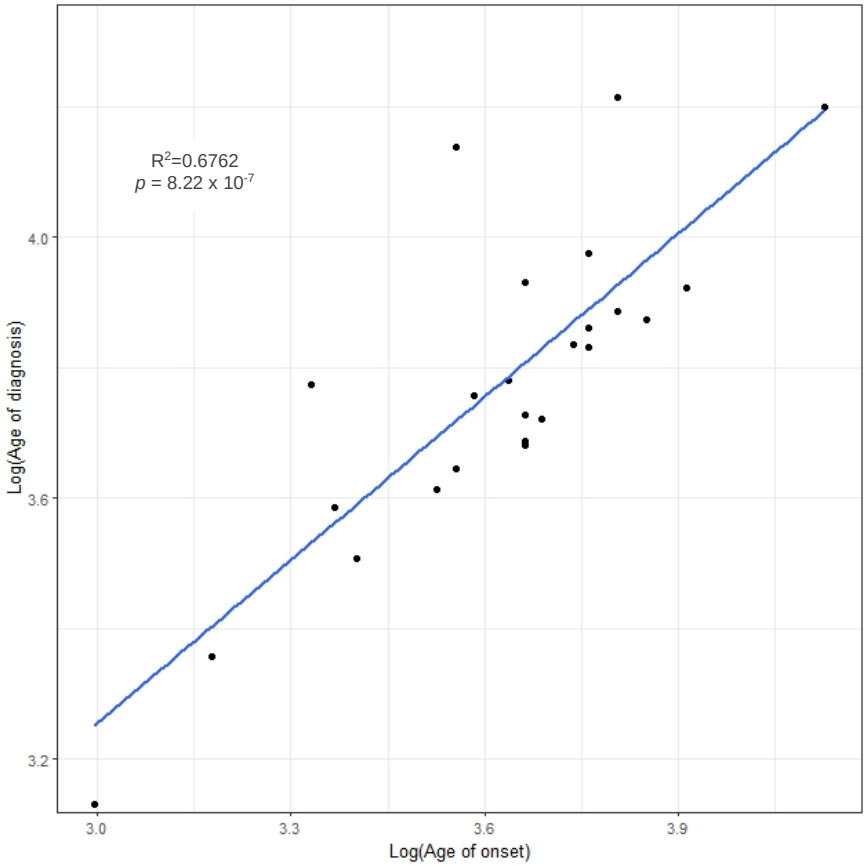
Supplementary Table 4. Multiple linear models testing the association between the HD phenotype and various explanatory variables.

- Model 1 Association between the AoD and the individual components of the *HTT* repeat tract.
- Model 2. Association between the AoD and the *HTT* disease allele structures.
- Model 3. Association between the AoD and the background haplogroups.
- Model 4. Association between the AoD and the background haplotypes.
- Model 5. Association between the AoD and the expansion score

		Glutamine		[Proline]				
		Q^1	Q^2	P^1	P^2	P^3	n	(%)
Reference:	Q^1-2-2-P^2-2:	CAG(n)-CAACAG - [CCGCCA-CCG(n)-CCTCCT]						
Disease allele structures							**68**	
Typical alleles:	Q^1-2-2-P^2(7,9,10)-2:	CAG(n)-CAACAG - [CCGCCA-CCG(7,9,10)-CCTCCT]					31	(46.5)
Atypical alleles:	Q^1-2-2-6-3:	CAG(n)-CAACAG - [CCGCCA-CCG(6)-**CCTCCTCCT**]					4	(5.9)
	Q^1-4-2-4-3:	CAG(n)-CAACAG-**CAACAG** - [CCGCCA-CCG(4)-**CCTCCTCCT**]					1	(1.5)
	Q^1-2-0-9-2:	CAG(n)-CAACAG - [CCG(9)-CCTCCT]					30	(44.1)
	Q^1-0-0-9-2:	CAG(n) - [CCG(9)-CCTCCT]					2	(2.9)
Non-disease allele structures							**384**	
Typical alleles:	Q^1-2-2-P^2(6-13)-2:	CAG(n)-CAACAG - [CCGCCA-CCG(6-13)-CCTCCT]					235	(61.2)
Atypical alleles:	Q^1-2-2-4-3:	CAG(n)-CAACAG - [CCGCCA-CCG(4)-**CCTCCTCCT**]					1	(0.3)
	Q^1-2-2-6-3:	CAG(n)-CAACAG - [CCGCCA-CCG(6)-**CCTCCTCCT**]					2	(0.5)
	Q^1-2-2-9-3:	CAG(n)-CAACAG - [CCGCCA-CCG(9)-**CCTCCTCCT**]					92	(24.0)
	Q^1-2-2-10-3:	CAG(n)-CAACAG - [CCGCCA-CCG(10)-**CCTCCTCCT**]					1	(0.3)
	Q^1-4-2-7-3:	CAG(n)-CAACAG-**CAACAG** - [CCGCCA-CCG(7)-**CCTCCTCCT**]					22	(5.7)
	Q^1-4-2-10-2:	CAG(n)- CAACAG-**CAACAG** - [CCGCCA-CCG(10)-CCTCCT]					4	(1.0)
	Q^1-2-0-9-2:	CAG(n)- CAACAG - [CCG(9)-CCTCCT]					27	(7.0)

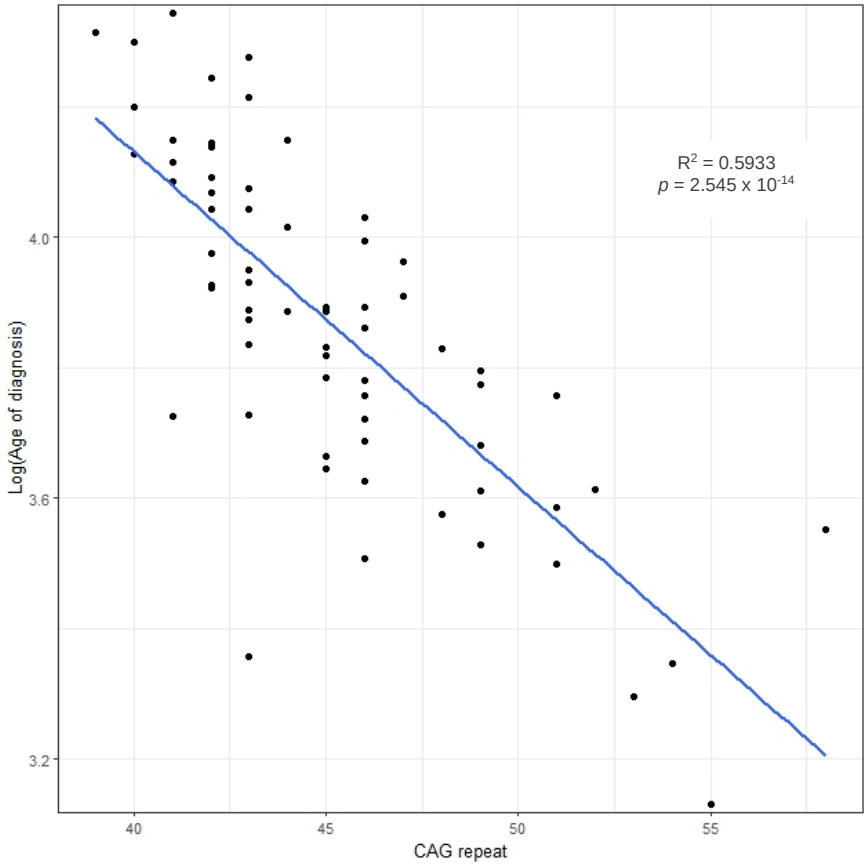
Supplementary Figure 1. The disease and non-disease allele structures defined in the individuals of African ancestry.

The *HTT* repeat tract is composed of a region that codes for polyglutamine with a variable number of CAG repeats (Q^1) and a CAACAG (Q^2) sequence. This is followed by a region coding for polyproline which includes a CCGCCA (P^1) sequence, a variable number of CCG (P^2) repeats and downstream CCT repeats (P^3). Typical alleles were defined as such if the allele structure was characteristic of the *HTT* reference gene (locus reference genomic: LRG_763). The atypical alleles were defined based on deviations from the reference gene, as a result of variants within the repeat tract indicated in bold. The polyglutamine region is upstream of the polyproline region in square brackets.



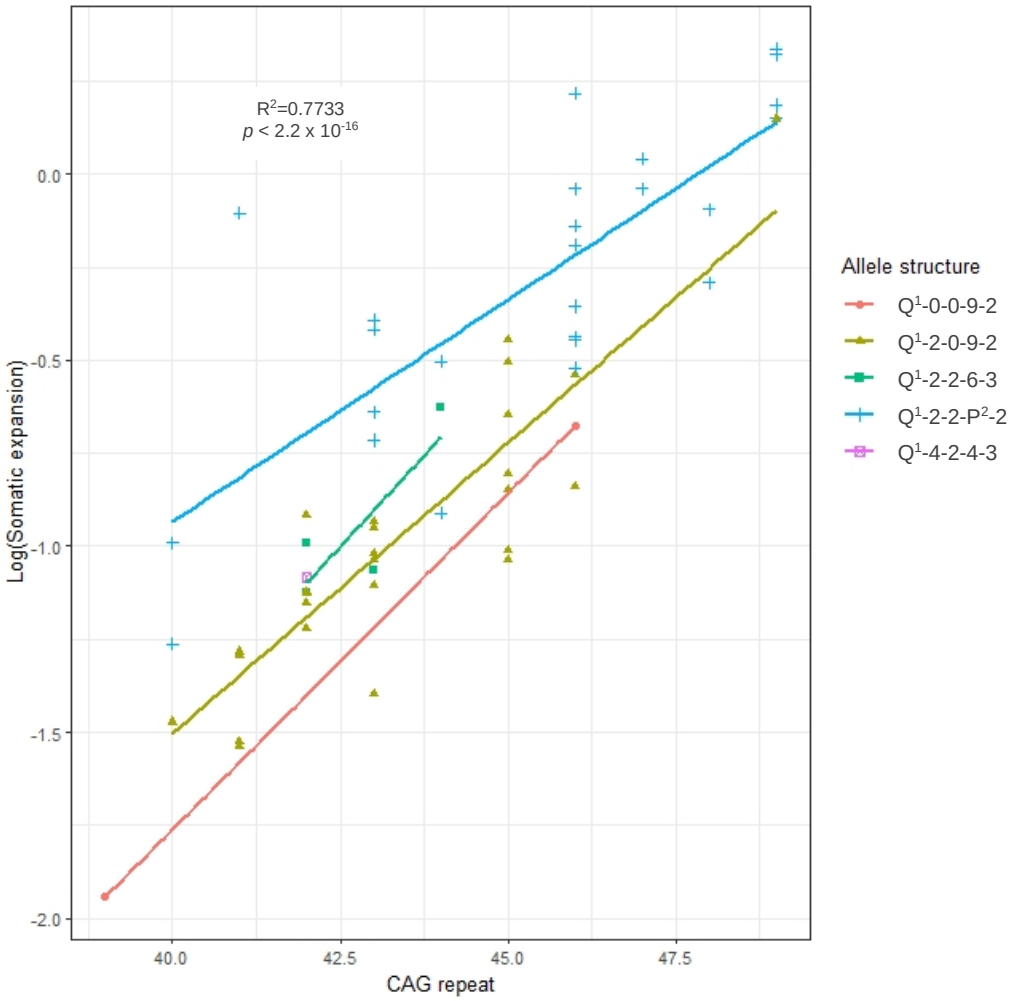
Supplementary Figure 2. Linear regression model of the log-transformed (AoD) with the log-transformed (AoO).

When studying the HD phenotype, the age at onset of motor symptoms is used as it is the most well characterised measure of disease severity. However, in our patients, less than 50% had AoO information and therefore the AoD was assessed. The relationship between the natural log transformed AoD and AoO, for 23 of the 68 patients for whom AoO data was available was assessed. A highly significant correlation was shown between the AoD and AoO, with approximately 68% of the variation in the AoD due to the AoO. It was therefore concluded that the AoD could be used as a reasonable proxy in this study.



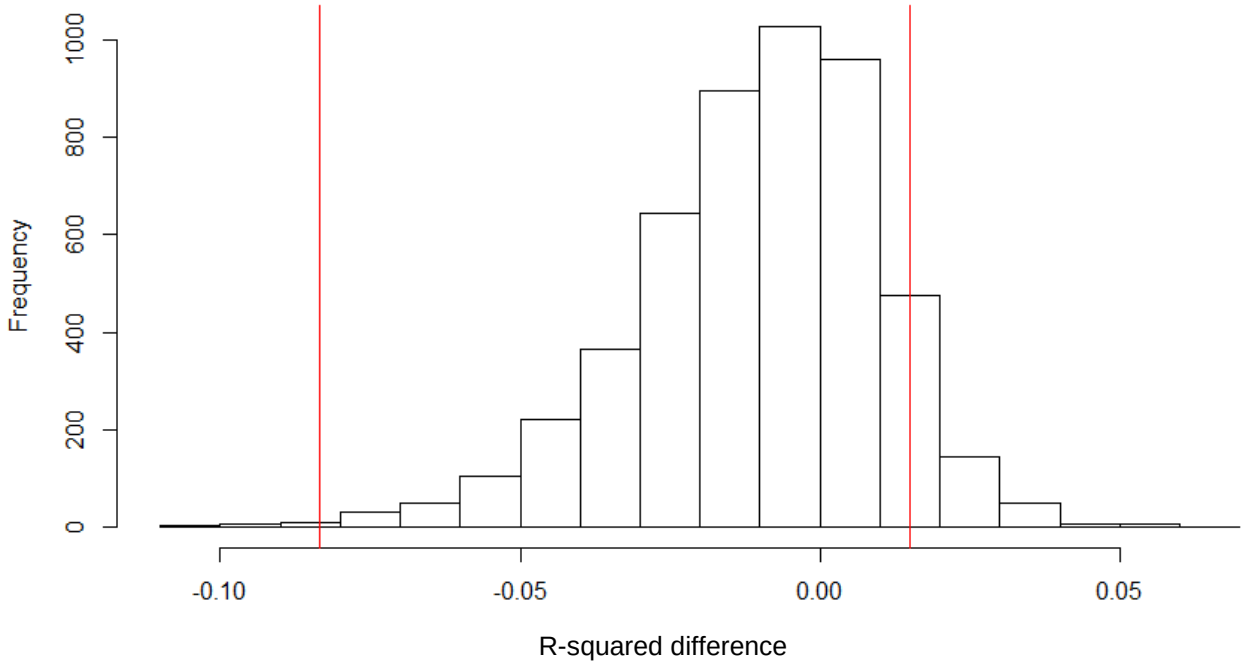
Supplementary Figure 3. Linear regression model testing the association between the HD-associated CAG repeats and log transformed (AoD).

A significant negative association was detected between the CAG repeat size and the AoD, accounting for approximately 60% of the variation in the HD phenotype.



Supplementary Figure 4. Linear model testing the association between the level of somatic expansion and CAG repeats size for the disease allele structures.

The amount of somatic expansion of the CAG repeats was measured by counting the ratio of reads larger (n+1 to 10 repeats) than the progenitor CAG repeat. The combined typical allele structures Q^1-2-2-P^2-2 have the highest relative ratio of somatic expansions and the lowest was present in the atypical allele structures, Q^1-2-0-9-2 and Q^1-0-0-9-2, both of which are characterised by a loss of the CCGCCA sequence (intervening proline). The Q^1-0-0-9-2 disease allele structure was however excluded as it was only present in two patients.



Supplementary Figure 5. The frequency distribution of the difference between the Q^1-2-0-9-2 allele structure model and the B2 haplotype model R-squared among 5000 bootstrapped samples.

The red lines indicate the 95% confidence interval of the difference between the R-square of the Q^1-2-0-9-2 and B2 models, contains 0. There is thus no statistical indication that Q^1-2-0-9-2 or B2 better explains the variation in AoD observed (*i.e.* B2 is not better associated with AoD than Q^1-2-0-9-2).

Supplementary Table 1. Demographic information of individuals affected with HD showing the disease associated CAG repeats (Q^1), allele structures (Q^1-Q^2-P^1-P^2-P^3), age of diagnosis (AoD) and age of onset (AoO). The age of onset information was only available for 24 patients whereas the age of diagnosis was available in all except for one individual.

No	Disease associated-allele CAG (Q^1)	Disease associated allele structure. Q^1-Q^2-P^1-P^2-P^3	AoD	AoO
1	39	Q^1-0-0-9-2	74.8	NA
2	40	Q^1-2-2-10-2	73.7	NA
3	40	Q^1-2-2-9-2	62.0	NA
4	40	Q^1-2-0-9-2	66.6	62.0
5	41	Q^1-2-0-9-2	61.3	NA
6	41	Q^1-2-0-9-2	41.5	NA
7	41	Q^1-2-0-9-2	59.5	NA
8	41	Q^1-2-0-9-2	63.4	NA
9	41	Q^1-2-2-7-2	77.0	NA
10	42	Q^1-2-2-6-3	69.7	NA
11	42	Q^1-2-0-9-2	62.7	35.0
12	42	Q^1-2-2-6-3	63.1	NA
13	42	Q^1-2-0-9-2	50.7	NA
14	42	Q^1-2-0-9-2	53.3	43.0
15	42	Q^1-2-0-9-2	59.9	NA
16	42	Q^1-2-0-9-2	50.5	50.0
17	42	Q^1-4-2-4-3	58.5	NA
18	42	Q^1-2-0-9-2	57.0	NA
19	43	Q^1-2-2-6-3	57.0	NA
20	43	Q^1-2-0-9-2	58.8	NA
21	43	Q^1-2-0-9-2	41.6	39.0
22	43	Q^1-2-0-9-2	48.1	47.0
23	43	Q^1-2-0-9-2	28.7	24.0
24	43	Q^1-2-0-9-2	51.9	NA
25	43	Q^1-2-2-10-2	50.9	39.0
26	43	Q^1-2-2-10-2	58.8	NA
27	43	Q^1-2-0-9-2	46.3	42.0
28	43	Q^1-2-2-10-2	67.6	45.0
29	43	Q^1-2-0-9-2	48.8	NA
30	43	Q^1-2-2-10-2	72.0	NA
31	44	Q^1-2-2-7-2	48.7	NA
32	44	Q^1-2-2-6-3	63.4	NA
33	44	Q^1-2-2-7-2	55.5	NA
34	45	Q^1-2-0-9-2	38.3	35.0
35	45	Q^1-2-0-9-2	45.5	NA
36	45	Q^1-2-0-9-2	39.0	NA
37	45	Q^1-2-0-9-2	48.7	45.0
38	45	Q^1-2-0-9-2	49.0	NA
39	45	Q^1-2-0-9-2	44.0	NA
40	45	Q^1-2-0-9-2	46.1	43.0
41	46	Q^1-2-2-10-2	56.3	NA
42	46	Q^1-2-2-10-2	33.4	30.0
43	46	Q^1-2-2-7-2	43.8	38.0
44	46	Q^1-2-2-10-2	39.9	39.0
45	46	Q^1-2-2-10-2	NA	NA
46	46	Q^1-2-0-9-2	41.3	40.0
47	46	Q^1-2-0-9-2	54.3	NA
48	46	Q^1-0-0-9-2	37.6	NA
49	46	Q^1-2-2-10-2	47.5	43.0
50	46	Q^1-2-2-7-2	49.0	NA
51	46	Q^1-2-2-7-2	42.8	36.0
52	47	Q^1-2-2-10-2	52.6	NA
53	47	Q^1-2-2-10-2	49.9	NA

54	48	Q^1-2-2-7-2	46.0	NA
55	48	Q^1-2-2-10-2	35.7	NA
56	49	Q^1-2-2-10-2	39.7	39.0
57	49	Q^1-2-2-10-2	43.6	28.0
58	49	Q^1-2-2-10-2	44.5	NA
59	49	Q^1-2-2-10-2	37.0	NA
60	49	Q^1-2-0-9-2	34.1	NA
61	51	Q^1-2-0-9-2	33.1	NA
62	51	Q^1-2-2-7-2	42.8	NA
63	51	Q^1-2-2-10-2	36.1	29.0
64	52	Q^1-2-2-7-2	37.1	34.0
65	53	Q^1-2-2-10-2	27.0	NA
66	54	Q^1-2-2-10-2	28.4	NA
67	55	Q^1-2-2-7-2	22.9	20.0
68	58	Q^1-2-0-9-2	34.9	NA

Supplementary Table 2. The tSNPs used to construct the *HTT* haplotypes.

No	rs number	Location on chromosome 4
1	rs2857936	3060583
2	rs762855	3073068
3	rs3856973	3078446
4	rs10015979	3107715
5	rs363075	3135947
6	rs363064	3139683
7	rs363102	3147289
8	rs4690073	3158423
9	rs363099	3160329
10	rs363096	3178294
11	rs2276881	3229934
12	rs362307	3240118
13	rs1006798	3256646

Location on chromosome 4: Homo sapiens (human) genome assembly GRCh38.p12 from Genome Reference Consortium.

Supplementary Table 3. Multiple linear models testing the association between the ratio of somatic expansion and various explanatory variables.

	Model	r^2	Adjusted r^2	p-value	Parameter values						
					Sample size	Explanatory variable	Estimate	Exponential estimate	% change	Estimate in ratio of SE	Pr(>|t|)
1.	Ln (SE) ~ CAG + Age at sampling + CAG*Age at sampling	0.7581	0.7449	$<2.2e^{-16}$	60	**CAG**	**0.211**	**1.235**	**23.450**	**0.131**	**$7.81e^{-16}$**
						Age at sampling	**0.015**	**1.015**	**1.486**	**0.008**	**$1.79e^{-3}$**
						CAG*Age at sampling	0.000	1.000	-0.039	0.000	$7.64e^{-1}$
2.	Ln (SE) ~ CAG + Age at sampling + CAG*Age at sampling + Allele structures	0.8513	0.8309	$2.2e^{-16}$	60	**CAG**	**0.185**	**1.204**	**20.359**	**0.114**	**$2.59e^{-3}$**
						Age at sampling	0.032	1.033	3.290	0.018	$5.21e^{-1}$
					2	**Q^1-0-0-9-2**	**-0.617**	**0.539**	**-46.061**	**-0.257**	**$7.66e^{-4}$**
					30	**Q^1-2-0-9-2**	**-0.358**	**0.699**	**-30.114**	**-0.168**	**$1.02e^{-5}$**
					4	**Q^1-2-2-6-3**	**-0.317**	**0.728**	**-27.154**	**-0.151**	**$1.14e^{-2}$**
					1	Q^1-4-2-4-3	0.301	1.351	35.084	0.196	$1.80e^{-1}$
						CAG*Age at sampling	-0.001	0.999	-0.059	0.000	$6.13e^{-1}$

The 'reference' used in the linear model 2.2 was the most common disease and non-disease allele structure (Q^1-2-2-P^2-2). The explanatory variables with a significant association to the log transformed AoD are indicated in bold.

Supplementary Table 4. Multiple linear models testing the association between the HD phenotype and various explanatory variables.

	Model	r^2	Adjusted r^2	p-value	Parameter values						
					Sample size	Explanatory variable	Estimate	Exponential estimate	% change	Estimate in years	Pr(>|t|)
1.	Ln (AoD)~ CAG + CAACAG + CCGCCA + CCG + CCT	0.6088	0.5745	$1.444e^{-10}$	64	**CAG**	**-0.059**	**0.943**	**-5.727**	**-2.902**	**$6.11e^{-12}$**
					0=2, 1=61, 2=1	CAACAG	-0.031	0.969	-3.066	-1.554	0.498
					0=31, 1=33	**CCGCCA**	**0.077**	**1.080**	**7.953**	**4.029**	**$7.34e^{-4}$**
					7/9/10=59, 4/6=5	CCG	-0.005	0.995	-0.509	-0.258	0.799
					2=59, 3=5	CCT	0.042	1.043	4.293	2.175	0.6801
2.	Ln (AoD)~ CAG + Allele structures	0.6097	0.5755	$1.355e^{-10}$	64	**CAG**	-0.059	0.943	-5.743	-2.910	$5.71e^{-12}$
					2	Q^1-0-0-9-2	-0.119	0.888	-11.212	-5.681	$2.88e^{-1}$
					29	**Q^1-2-0-9-2**	**-0.152**	**0.859**	**-14.079**	**-7.133**	**$8.15e^{-4}$**
					4	Q^1-2-2-6-3	0.070	1.073	7.285	3.691	$3.96e^{-1}$
					1	Q^1-4-2-4-3	-0.050	0.951	-4.912	-2.489	$7.43e^{-1}$
3.	Ln (AoD)~ CAG + Haplogroups	0.5867	0.5577	2.054^{-10}	64	**CAG**	**-0.059**	**0.943**	**-5.730**	**-2.903**	**$2.09e^{-11}$**
					9	**A**	**0.156**	**1.169**	**16.878**	**8.551**	**$1.37e^{-2}$**
					18	**C**	**0.115**	**1.122**	**12.241**	**6.202**	**$2.24e^{-2}$**
					4	**C-SA**	**0.209**	**1.232**	**23.195**	**11.752**	**$1.24e^{-2}$**
4.	Ln (AoD)~ CAG + Haplotypes	0.6433	0.5815	$5.412e^{-9}$	64	**CAG**	-0.062	0.939	-6.058	**-3.069**	**3.19e-11**
					1	A2a	0.036	1.037	3.680	1.864	8.11e-1
					1	A2b	0.167	1.182	18.206	9.225	2.75e-1
					2	**A4a**	**0.287**	**1.332**	**33.210**	**16.826**	**1.77e-2**
					5	**A4b**	**0.165**	**1.179**	**17.932**	**9.086**	**2.90e-2**
					1	**B1**	**0.365**	**1.440**	**43.999**	**22.293**	**1.88e-2**
					1	C4	0.028	1.028	2.801	1.419	8.57e-1
					17	**C5**	**0.143**	**1.153**	**15.337**	**7.771**	**6.76e-3**
					4	**C9**	**0.218**	**1.243**	**24.322**	**12.323**	**7.85e-3**
5.	Ln (AoD)~ CAG + Allele structures + Expansion score	0.625	0.5818	$1.296e^{-9}$	60	**CAG**	**-0.072**	**0.931**	**-6.915**	**-3.504**	**$1.56e^{-11}$**
					2	**Q^1-0-0-9-2**	**-0.257**	**0.773**	**-22.679**	**-11.491**	**$3.43e^{-2}$**
					30	**Q^1-2-0-9-2**	**-0.224**	**0.799**	**-20.091**	**-10.180**	**$4.20e^{-5}$**
					4	Q^1-2-2-6-3	-0.017	0.983	-1.659	-0.840	$8.46e^{-1}$
					1	Q^1-4-2-4-3	-0.124	0.883	-11.651	-5.903	$4.11e^{-1}$
						Expansion score	**-0.235**	**0.791**	**-20.922**	**-10.600**	**$1.20e^{-2}$**

The *HTT* repeat tract components (Model 1), disease allele structures (Model 2), background haplogroups (Model 3), background haplotypes (Model 4), somatic expansion (Model 5), relative to the reference. The 'reference' used in the linear model 2 and 5 was the most common disease and non-disease allele structure (Q^1-2-2-P^2-2) and the African 'reference' used in the linear 3 and 4 was haplogroup (B) and haplotype (B2) respectively.

References

Baine, F.K., Kay, C., Ketelaar, M.E., Collins, J.A., Semaka, A., Doty, C.N., Krause, A., Greenberg, L.J., and Hayden, M.R. (2013). Huntington Disease in the South African Population Occurs on Diverse and Ethnically Distinct Genetic Haplotypes. European Journal of Human Genetics 21, 1120-1127.

Baine, F.K., Krause, A., and Greenberg, L.J. (2016). The Frequency of Huntington Disease and Huntington Disease-Like 2 in the South African Population. Neuroepidemiology 46, 198-202.

Bañez-Coronel, M., Ayhan, F., Tarabochia, A.D., Zu, T., Perez, B.A., Khoramian Tusi, S., Pletnikova, O., Borchelt, D.R., Ross, C.A., Margolis, R.L., et al. (2015). RAN Translation in Huntington Disease. Neuron 88, 667–677.

Becanović, K., Nørremølle, A., Neal, S.J., Kay, C., Collins, J.A., Arenillas, D., Lilja, T., Gaudenzi, G., Manoharan, S., Doty, C.N., et al. (2015). A SNP in the *HTT* Promoter Alters NF-kB Binding and is a Bidirectional Genetic Modifier of Huntington Disease. Nature Neuroscience 18, 807-816.

Brock, G.J.R., Anderson, N.H., and Monckton, D.G. (1999). Cis-acting Modifiers of Expanded CAS/CTG Triplet Repeat Expandability: Associations with Flanking GC Content and Proximity to CpG Islands. Human Molecular Genetics 8, 1061–1067.

Busan, S., and Weeks, K.M. (2013). The Role of Context in RNA Structure: Flanking Sequences Reconfigure CAG Motif Folding in Huntingtin Exon 1 Transcripts. Biochemistry 52, 8219–8225.

Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., et al. (1993). Trinucleotide Repeat Length Instability and Age of Onset in Huntington's Disease. Nature Genetics 4, 387-392.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E.R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M.J., et al. (2017). Whole-Genome Sequencing for an Enhanced Understanding of Genetic Variation Among South Africans. Nature Communications 8, 1-12.

Ciosi, M., Cumming, S. A., Alshammari, A. M., Symeonidi, E., Herzyk, P., McGuinness, D., Galbraith, J., Hamilton, G., et al. 2018. Library Preparation and MiSeq Sequencing for the Genotyping-By-Sequencing of the Huntington Disease HTT Exon One Trinucleotide Repeat and the Quantification of Somatic Mosaicism. Protocol Exchange.

Ciosi, M., Maxwell, A., Cumming, S.A., Hensman Moss, D.J., Alshammari, A.M., Flower, M.D., Durr, A., Leavitt, B.R., Roos, R.A.C., the TRACK-HD team, et al. (2019). A Genetic Association Study of Glutamine-Encoding DNA Sequence Structures, Somatic CAG Expansion, and DNA Repair Gene Variants, with Huntington Disease Clinical Outcomes. EBioMedicine 48, 568-580.

Gellera, C., Meoni, C., Castellotti, B., Zappacosta, B., Girotti, F., Taroni, F., and DiDonato, S. (1996). Errors in Huntington Disease Diagnostic Test Caused by Trinucleotide Deletion in the IT15 Gene. American Journal of Human Genetics 59, 475-477.

GeM-HD Consortium. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. Cell 178, 887-900.

Goldberg, Y.P., McMurray, C.T., Zeisler, J., Almqvist, E., Sillence, D., Richards, F., Gacy, A.M., Buchanan, J., Telenius, H., and 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. Human Molecular Genetics 4, 1911-1918.

Kay, C., Collins, J.A., Wright, G.E.B., Baine, F., Miedzybrodzka, Z., Aminkeng, F., Semaka, A., McDonald, C., Davidson, M., Madore, S.J., et al. (2018). The Molecular Epidemiology of Huntington Disease is Related to Intermediate Allele Frequency and

Haplotype in the General Population. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 177, 346-357.

Kennedy, L., Evans, E., Chen, C.M., Craven, L., Detloff, P.J., Ennis, M., and Shelbourne, P.F. (2003). Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. Hum Mol Genet 12, 3359-3367.

Kraus-Perrotta, C., and Lagalwar, S. (2016). Expansion, Mosaicism and Interruption: Mechanisms of the Cag Repeat Mutation in Spinocerebellar Ataxia Type 1. Cerebellum Ataxias 3, 1-11.

Krauss, S., Griesche, N., Jastrzebska, E., Chen, C., Rutschow, D.s., Achmuller, C., Dorn, S., Boesch, S.M., Lalowski, M., Wanker, E., et al. (2013). Translation of *HTT* mRNA with Expanded CAG Repeats is Regulated by the MID1–PP2A Protein Complex. Nature Communications 4, 1-9.

Lee, J.M., Ramos, E.M., Lee, J.H., Gillis, T., Mysore, J.S., Hayden, M.R., Warby, S.C., Morrison, P., Nance, M., Ross, C.A., et al. (2012). CAG Repeat Expansion in Huntington Disease Determines Age at Onset in a Fully Dominant Fashion. Neurology 78, 690-695.

Liu, J., McFarland, K.N., Landrian, I., Wu, S.S., Bower, M., Hutter, D., Bushara, K., Teive, H.A.G., and Ashizawa, T. (2014). Identifying Novel Interruption Motifs in Spinocerebellar Ataxia Type 10 Expansions. Neurology and Clinical Neuroscience 2, 38-43.

Machiela, M.J., and Chanock, S.J. (2015). LDlink: A Web-Based Application for Exploring Population-Specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants. Bioinformatics 31, 3555-3557.

Neueder, A., Landles, C., Ghosh, R., Howland, D., Myers, R.H., Faull, R.L.M., Tabrizi, S.J., and Bates, G.P. (2017). The Pathogenic Exon 1 *HTT* Protein is Produced by Incomplete Splicing in Huntington's Disease Patients. Scientific Reports 7, 1-10

Panegyres, P.K., Beilby, J., Bulsara, M., Toufexis, K., and Wong, C. (2006). A Study of Potential Interactive Genetic Factors in Huntington's Disease European Neurology 55, 189-192.

Pringsheim, T., Wiltshire, K., Day, L., Dykeman, J., Steeves, T., and Jette, N. (2012). The Incidence and Prevalence of Huntington's Disease: A Systematic Review and Meta-Analysis. Movement Disorders 27, 1083-1091.

Roos, R.A. (2010). Huntington's Disease: A Clinical Review. Orphanet Journal of Rare Diseases 5, 1-8.

The Huntington's Disease Collaborative Research Group. (1993). A Novel Gene Containing a Trinucleotide Repeat that is Expanded and Unstable on Huntington's Disease Chromosomes. Cell 72, 971-983.

Telenius, H., Kremer, B., Goldberg, Y.P., Theilmann, J., Andrew, S.E., Zeisler, J., Adam, S., Greenberg, C., Ives, E.J., Clarke, L.A., et al. (1994). Somatic and Gonadal Mosaicism of the Huntington Disease Gene CAG Repeat in Brain and Sperm. Nature Genetics 6, 409-414

Veitch, N.J., Ennis, M., McAbney, J.P., Shelbourne, P.F., Monckton, D.G., and Project, T.U.-V.C.R. (2007). Inherited CAG·CTG allele length is a major modifier of somatic mutation length variability in Huntington disease. DNA repair 6, 789-796.

Walker, F.O. (2007). Huntington's Disease. The Lancet 369, 218-228.

Warby, S.C., Visscher, H., Collins, J.A., Doty, C.N., Carter, C., Butland, S.L., Hayden, A.R., Kanazawa, I., Ross, C.J., and Hayden, M.R. (2011). *HTT* Haplotypes Contribute to Differences in Huntington Disease Prevalence Between Europe and East Asia. European Journal of Human Genetics 19, 561-566.

Wright, G.E.B., Collins, J.A., Kay, C., McDonald, C., Dolzhenko, E., Xia, Q., Bečanović, K., Semaka, A., Nguyen, C.M., Trost, B., et al. (2019). Length of Uninterrupted CAG Repeats, Independent of Polyglutamine Size, Results in Increased Somatic Instability and Hastened Age of Onset in Huntington Disease. The American Journal of Human Genetics 104, 1116-1126.

Yu, S., Fimmel, A., Fung, D., and Trent, R.J. (2000). Polymorphisms in the CAG Repeat-A Source of Error in Huntington Disease DNA Testing. Clinical Genetics 58, 469-472.