University of Glasgow
Institute of Molecular, Cell and Systems Biology
College of Medical, Veterinary and Life Sciences.

Massively parallel next generation sequencing to investigate the *cis*- and *trans*-acting genetic modifiers of somatic instability in Huntington's disease

Asma Mubarak Alshammari

A thesis submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy

2018

Abstract

Huntington disease (HD) is an extremely variable inherited neurodegenerative disorder caused by expansion of an unstable CAG trinucleotide repeat in the huntingtin gene (*HTT*). Somatic instability in HD exhibits an age-dependent, expansion-biased and tissue-specific pattern, and the highest level of somatic instability is found in tissues that are most susceptible to the disease pathology. Therefore, the aim of this project was to test the hypothesis that somatic instability of the HD CAG repeat plays a major role in disease pathology by quantifying somatic instability in the number of CAG repeats by next generation sequencing (NGS) technology in buccal cell DNA.

We developed a method to sequence and genotype *HTT* alleles from blood and buccal swab DNA of the Scottish and Venezuelan populations respectively. A total of 210 individuals from the Scottish general population and 742 HD patients and unaffected individuals from the Venezuelan HD cohort were sequenced on the MiSeq platform. We established that it was possible to sequence and genotype the CAG repeats, the polymorphic CCG repeat and the flanking sequences. Our data highlight the utility of NGS technology as an approach to genotype *HTT* alleles, detect sequence variants and quantify somatic instability of the CAG repeat. Our data emphasise that the somatic instability in HD is age-dependent and expansion-biased, also could be a major factor in disease progression, and could be a potential therapeutic target in HD.

We also investigated the possibility that there are *trans*-acting modifier factors involved in determining the degree of somatic instability in HD patients. We genotyped polymorphisms in candidate mismatch repair (MMR) genes and examined their effect, if any, on the residual variation of somatic instability. Individuals carrying the minor allele of rs3512 in *FAN1* have a higher level of somatic instability than average, suggesting that some of the variations in HD somatic instability could be accounted for by genetic variation in the DNA mismatch repair pathway. The search for modifier genes might have consequences in understanding the pathological process in HD, and may therefore provide therapeutic targets for future investigations.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I would like to thank prof. Darren Monckton for being such an inspiring and positive supervisor. Thanks for your good supervision, encouragement, valuable suggestions, and help through my PhD study. Thanks to my second supervisor Dr. Graham Hamilton for his help throughout my research.

It has been a real pleasure being a part of Monckton's group. I could not have asked for a nicer group, so thank you to all members past and present. Special thanks to Dr. Sarah Cumming and Dr. Marc Cioisi for offering practical and theoretical advices and their help through my research. I also would like to thank Alastair Maxwell for helping me in bioinformatics analysis. Thanks for all members of the Monckton's group for their help and support during my PhD study Dr. Sarah Cumming, Dr. Marc Ciosi, Dr. Khaldah Nasser, Dr. Gayle Overend, Eloise Larson, Mariam Alkhatteb, Tolulope Oyeniyi, Alastair Maxwell, Dr. Afroditi Chatzi, and Vilija Lomeikaite.

I would like to thank my friends at the Davidson building: Mariam Alkhateeb, Sana Alqarni, and Amaal Alrehaili, for their help and support. You girls made my PhD study much more enjoyable.

I am also very grateful to Civil service commission, Kuwait for funding this research.

Finally, I would like to express my sincere gratitude to my parents (Mubarak Alshammari and Moudhi Alkhaldei) and my brothers for their kind words of encouragement and their support through my PhD study. I am especially thankful to my Husband, Abdulrahman Almulla, for being loving, supportive, and always being there for me.

Dedication

To my beloved parents, husband, and little son, Mohammad.

Author's Declaration

I declare that the work presented in this thesis is entirely my own unless stated otherwise. This thesis has not been submitted for any other degree at the University of Glasgow or elsewhere.

Abbreviations

3 ’UTR	Three prime untranslated region
5 ’UTR	Five prime untranslated region
Bp	Base pair
DM1	Myotonic dystrophy type 1
FAN1	FANCD2 And FANCI Associated Nuclease 1
GWAS	Genome Wide Association Study
HD	Huntington’s disease
HTT	Huntingtin protein
HTT	Huntingtin gene
HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
MAF	Minor allele frequency
MMR	Mismatch repair
MSH2	MutS homolog 2
MSH3	MutS homolog 3
MSH6	MutS homolog 6
MLH1	MutL homolog 1
MLH3	MutL homolog 3
NGS	Next generation sequencing
PCR	Polymerase chain reaction
QC	Quality control
QFAM	A family-based quantitative trait association analysis
QTDT	Quantitative Transmission Disequilibrium Tests
SNP	Single-nucleotide polymorphism
SP-PCR	Small pool PCR
SCA	Spinocerebellar ataxia
TP-PCR	Triplet-primed PCR

Chapter 1 Introduction

1.1 Trinucleotide repeat diseases

Nearly 40 inherited human disorders result of expansions of unstable DNA simple sequence repeats (Gomes-Pereira and Monckton, 2006; McMurray, 2010). Most of these disorders involve expansions of a trinucleotide repeat, including expanded CAG repeats at the Huntington disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA) and a group of spinocerebellar ataxias (SCA1, 2, 3, 6, 7, 8, 12 and 17), expanded CTG repeats at the myotonic dystrophy type 1 (DM1), expanded CGG repeats at the fragile X syndrome and expanded GAA repeats at the Friedreich ataxia (FRDA) locus.

These diseases are variable and often have unusual inheritance patterns. The expanded alleles become highly unstable in both germline and somatic tissues (Cleary and Pearson, 2003; Pearson, 2003). Almost all of these disorders exhibit wide symptomatic variability in terms of severity and age of onset, ranging from mild late onset forms to severely congenital forms. Moreover, repeat expansion rates have been linked with increased disease severity and an earlier age of onset (Swami *et al.*, 2009; Morales *et al.*, 2012).

1.2 Huntington's disease (HD)

Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder that affects striatal neurons, resulting in emotional, cognitive and motor disturbances. HD affects 4-8 per 100,000 individuals in Caucasian populations (Harper, 1992). The phenotype is very variable between patients in terms of disease course and range of symptoms. The characteristic symptom of HD is the presence of involuntary movements called chorea. It affects individuals usually in mid-life (approximately 3rd to 4th decade of life), and the disease progresses for 10-15 years after onset to predictable death (Bates, 2005). Subsets of HD cases are associated with the juvenile form of the disease. These patients have a more severe phenotype compared with adult-onset patients and onset before 20 years of age (Nance and Myers, 2001; Andresen *et al.*, 2007). HD shows a trend towards an increase in disease severity in subsequent generations and earlier age

at onset mostly seen through paternal transmission, known as anticipation (Telenius *et al.*, 1993).

Offspring of an affected individual have a 50% chance of inheriting the disease and those who inherit the mutation eventually develop the disease assuming that they do not die of other causes before the age of onset because penetrance is very high in HD. Males and females are equally at risk of HD, and it can be transmitted from mothers and fathers, although the juvenile cases tend to be inherited from fathers (Telenius *et al.*, 1993; Nahhas *et al.*, 2005).

1.3 Genetics of HD

The HD mutation was first mapped to the short arm of chromosome 4 through linkage analysis using polymorphic DNA markers in humans in 1983 (Gusella *et al.*, 1983). The Huntington's Disease Collaborative Research Group (1993) used haplotype linkage disequilibrium to spotlight a small segment of 4p16.3 as the likely location of the HD-causing genetic variant in affected individuals of 75 HD families examined. They identified a polymorphic CAG repeat near the 5' end of the huntingtin (*HTT*) gene, which was expanded and unstable on disease chromosomes (Huntington's Disease Collaborative Research Group, 1993).

HD is caused by a CAG repeat expansion located in exon 1 of the *HTT* gene. The CAG repeat is highly polymorphic, the repeat size in the non HD-causing chromosomes varies in length from 10 to ~35 repeats (Duyao *et al.*, 1993; Huntington's Disease Collaborative Research Group, 1993; Snell *et al.*, 1993). Most non HD-causing CAG repeats are stable through inheritance. Although alleles between 27 and 35 repeats are in the non-disease causing range, they are defined as intermediate alleles (Goldberg *et al.*, 1995). These alleles can be unstable through paternal transmission and can expand into the full mutation upon transmission, resulting in the HD phenotype in the offspring (Semaka, Collins and Hayden, 2010). An allele with 36 to 39 repeats is considered to have a reduced penetrance, i.e. some individuals in that range develop HD and others do not. Most HD patients with adult-onset inherit 40 to 50 repeats in the HD gene that are fully penetrant (Kremer *et al.*, 1994; Rubinsztein *et al.*, 1996). The largest CAG repeat reported in a patient contained approximately 200 repeats (Nance *et al.*, 1999). Patients carrying >50 repeats often have the severe

juvenile form of the disease (Andresen *et al.*, 2007). The expanded CAG repeat is immediately adjacent to a polymorphic CCG repeat of 4 to 12 repeats that is stably transmitted. Most non disease-causing and disease-causing chromosomes have seven CCG repeats (Andrew *et al.*, 1994).

1.4 Molecular pathogenesis of HD

The CAG repeat expansion results in the elongation of a polyglutamine array at the N-terminus of the huntingtin (HTT) protein (Pennuto, 2010). The expanded allele results in greater levels of cell death, dysfunction of a variety of cellular processes, and hence more severe phenotype. This dysfunction is thought to be mediated by gain of function, meaning the mutation result in the mutant protein gaining toxic function (Ross and Tabrizi, 2011). The mutant proteins appear to form aggregate in cells, forming neuronal inclusions bodies (Pennuto, 2010). However, there is evidence of RNA toxicity triggered by expanded CAG repeats through mechanisms involving gene expression (Marti, 2016).

HTT is required for normal development and survival of central nervous system cells. HTT is widely expressed at different tissue and cellular levels. HTT has a role in vesicle transport, regulates gene transcription, regulates protein synthesis and RNA trafficking (Cattaneo et al., 2005; Sadri-Vakili and Cha, 2006). The greatest expression of the protein found in the central nervous system. The mutant expanded polyglutamine leading to the hallmark pathology of HD causing atrophy in the striatum and also other brain regions.

1.5 Genotype-phenotype correlations

The CAG repeat expansion is the only mutation resulting in clinical symptoms of HD in familial and sporadic HD cases (Myers *et al.*, 1993; Kremer *et al.*, 1994). A strong inverse correlation exists between the number of CAG repeats in the expanded alleles and the age at onset for motor signs in HD and severity of the disease (Andrew *et al.*, 1993; Duyao *et al.*, 1993; Gusella and MacDonald, 2009). An increase of a single CAG repeat is associated with a reduction in age at onset by ~2 years (Andrew *et al.*, 1993; Gusella and MacDonald, 2009). Therefore, one inherited repeat length difference in HD patients can have a noticeable effect in

the prediction of age at onset. Most HD patients have age at onset in midlife due to inheriting repeats of 40-50 CAG repeats.

The length of the CAG repeat is the primary determinant of age at onset of clinical symptoms, accounting for approximately 70% of the variation in age at onset (Duyao *et al.*, 1993; The U.S.–Venezuela Collaborative Research Project and Wexler, 2004). The remaining variation and variable range of age at onset associated with each expanded alleles indicate the influence of unknown genetic or environmental modifying factors that are likely to be involved in determining the age of onset. The identification of these factors offers a promising route to finding factors that may delay disease onset in HD patients.

1.6 Genetic instability in HD

The expanded alleles are unstable in both germline and somatic tissues of affected individuals. This instability in HD may explain some features of HD such as the anticipation and progressive and tissue-specific aspects of the disease.

1.6.1 Germline instability in HD

The non HD-causing alleles with <27 CAG repeats are genetically very stable with germline mutation rates that result in a change in repeat length < 1% per generation (Duyao *et al.*, 1993; Kremer *et al.*, 1995). The expanded HD alleles are unstable in about 80% of intergenerational transmissions, showing both decreases and increases in size (Leeflang *et al.*, 1999). However, expansions are more frequent, particularly in the paternal germline (De Rooij *et al.*, 1993; Duyao *et al.*, 1993; Snell *et al.*, 1993; Telenius *et al.*, 1994). Longer alleles have a higher mutation frequency that can approach as high as 98% per generation with an expansion bias in the male germline (Leeflang *et al.*, 1999). As longer alleles are associated with the more severe form of the disease, expansion-biased intergenerational instability often results in a decreasing age of disease onset from one generation to the next. Thus, germline instability explains the anticipation phenomenon observed within HD families and also other trinucleotide disorders such as Myotonic dystrophy type 1(DM1), where an earlier age at onset and increased severity of symptoms is observed in successive generations.

Germline instability also provides an explanation for the parental sex of origin effects associated with HD (Duyao *et al.*, 1993; Kremer *et al.*, 1995). Indeed, the expanded alleles have been shown to behave differently in the male and female germline. Smaller expansions (<7 repeats) occur more frequently in the offspring of affected females than in the offspring of affected males (Kremer *et al.*, 1995).

1.6.2 Somatic instability in HD

The number of CAG repeats have been shown to vary within and between somatic tissues of individuals affected by HD (Telenius *et al.*, 1994; De Rooij *et al.*, 1995). The expanded HD allele is particularly unstable within regions in the brain such as the striatum and cortex that are primarily affected in the disorder when compared to peripheral tissues (Telenius *et al.*, 1994; De Rooij *et al.*, 1995). Somatic instability was most noticeable in juvenile-onset cases of HD, where cells in the striatum and cortex had gained expansions larger than cells in other areas of the brain such as cerebellum (Telenius *et al.*, 1994).

Sensitive small pool PCR (SP-PCR) analyses have been used to analyse the mutation length profiles from different brain regions in HD cases (Kennedy *et al.*, 2003). The results have revealed somatic expansions in some cells in the striatum and cortex that have acquired expansions of up to 1,000 repeats in length, in HD patients who inherited alleles with 40 to 50 repeats. These data support the idea that somatic instability may play a crucial role in the tissue-specific pathology and the progressive nature of the disease. Although such very large expansions appear to occur earlier in the striatum than in the other areas of the brain, in end-stage brain tissue, a high level of mutation instability was still present in the cortex, but not in the striatum (Kennedy *et al.*, 2003). This finding suggests that striatal cells with the largest HD mutation expansions might be lost during the disease process. The fact that the mutation profile in different regions of the brain correlates with the neuropathological involvement of the disease is consistent with the hypothesis that somatic instability may contribute to the tissue specificity and progressive nature of HD.

Analysis of patients with extremely early age at onset relative to the number of repeats inherited has revealed larger expansions in the cortex compared to

patients with later age at onset (Swami *et al.*, 2009). These data further support the hypothesis that somatic instability contributes to the pathogenic process, with higher somatic instability significantly associated with disease progression and earlier age at onset.

While analyses of HD repeat instability in the affected brain regions provides an understanding of the somatic instability preceding the disease onset, they do not allow determination of the contribution of somatic instability to disease onset and HD pathogenesis in most patients. This is because brain tissues cannot be evaluated until death, and it is difficult to analyse the most affected tissues because a high proportion of cells has already been lost during the disease process. However, it is possible to evaluate the level of somatic instability in other peripheral tissues, such as buccal cells. Although buccal cell DNA showed a low level of somatic instability in HD patients, it can be measured, and the degree of instability varied between individuals (Veitch *et al.*, 2007). This variation in somatic instability provided evidence for the major role of inherited allele length in driving somatic instability that accounts for ~70% of the variation in somatic instability observed in individuals with identical age at sampling (Veitch *et al.*, 2007). Thus, if the somatic instability in the brain is reflected in peripheral tissues, these peripheral tissues could be used to further understand the mechanisms and pathogenesis of somatic instability in HD. This study showed a low level of somatic instability using single molecule PCR that is impractical and labour intensive for high-throughput analysis in a large cohort of patients. To overcome the limitation of the previously used method, a novel method should be developed to facilitate high throughput analysis of somatic instability.

Together, these data emphasise that the somatic instability in HD is age-dependent, expansion-biased and tissue-specific. Somatic instability is therefore likely to be a major determinant of disease progression and also can be a potential therapeutic target in HD.

1.7 Molecular mechanisms of expansion

Multiple pathways of DNA metabolism have been implicated in generating repeat expansions in humans, such as DNA replication and mismatch repair.

It has been assumed that replication slippage is linked to the mechanism for generating somatic expansion in HD (Kunkel, 1993; Chi and Lam, 2005). However, there is evidence suggesting that DNA replication is not a major pathway in somatic expansions. The replication slippage model predicts the somatic expansion is cell cycle-dependent. However, the fact that somatic expansions occur in non-replicating cells such as neurons (Shelbourne *et al.*, 2007) argues against the involvement of DNA replication and cell division in generating mutations in somatic tissues in HD and supports the involvement of DNA repair pathways.

1.7.1 Role of mismatch repair (MMR) genes

The critical function of the DNA mismatch repair (MMR) pathway is to correct for misincorporation errors during DNA replication. The MMR can correct for base-base mismatches as well as for insertions/deletions that arise from replication slippage events.

MMR proteins that are present in mammalian somatic cells are composed of two heterodimers: the MutS homologues (MSH2, MSH3 and MSH6) and the MutL homologues (MLH1, PMS1, PMS2, and MLH3). DNA mismatches are recognised by complexes of either MSH2 with MSH3, to form MutSß, or MSH6, to form MutSα, depending on the type of DNA lesion that they recognise (Palombo *et al.*, 1996; Bellacosa, 2001). MutSα is critical for recognition of single base mismatches and MutSß for recognition of small insertions/deletions up to 12 bp. MutS proteins recruit the MutL dimer to the lesion. Likewise, there are two components of MutL: the MutLα complex (MLH1 and PMS2), and the MutLß complex (MLH1 and MLH3) (Li and Modrich, 1995; Lipkin *et al.*, 2000). The MutLα complex recognises single base mismatches, small and large insertions/deletions, while the MutLß complex also recognises small insertions/deletions.

Mouse models have been used to reveal the critical role and involvement of the mismatch repair genes in the somatic instability of the *HTT* CAG repeat. These studies have been carried out by crossing the Hdh knock-in (expanded CAG repeats were inserted at the mouse HD locus) or Htt transgenic mice (insertion of human HD gene, or a fragment of it, randomly into the mouse genome) with mismatch repair gene knockout mice to generate mouse models that have a

targeted mismatch repair inactivating mutation (Manley *et al.*, 1999; Wheeler *et al.*, 2003; Pinto *et al.*, 2013). The effect of the mismatch repair genes (*Msh2*, *Mlh1* and *Mlh3*) knockout is to eliminate somatic expansion. Studies in transgenic R6/1 mice (that were generated by incorporating a human genomic fragment including HTT promoter elements, the entire exon 1 including ~ 116 CAG repeats and a portion of intron 1) which were crossed with *Msh2* deficient mice, have demonstrated that the loss of *Msh2* alleles were required to produce an effect on somatic instability in transgenic mice (Manley *et al.*, 1999). This finding demonstrates that *Msh2* deficiency stabilises the HD repeat in the R6/1 mice model.

The data from HD knock-in mice have shown that the deficiency of *Msh2*, *Msh3*, *Mlh1* and *Mlh3* mismatch repair genes eliminate the somatic instability and delay the phenotype (Wheeler *et al.*, 2003; Dragileva *et al.*, 2009; Pinto *et al.*, 2013). These data support the role of the mismatch repair pathway in somatic instability. This is consistent with the data from transgenic mouse models (Manley *et al.*, 1999).

Similarly, there is also evidence from myotonic dystrophy type 1 (DM1) knock-in mice with expanded 84 CTG repeats placed into the mouse DM1 locus, that MMR components Msh3 and Msh6 have an impact on variation in the degree of somatic instability (van den Broek *et al.*, 2002). Deficiency of *Msh3* resulted in complete suppression of somatic instability in the mouse model, compared with *Msh6* deficient mice, which exhibited a greater somatic instability. It was suggested that competition of Msh3 and Msh6 for binding to Msh2 in MMR complexes might explain the opposite effects of those two MMR proteins. Although Msh2–Msh3 and Msh2–Msh6 repair complexes recognise unusual DNA structures formed by insertions/ deletions, they have slight differences in their specification. The Msh2–Msh3 complex has a preference for binding to large insertion/deletion loops, whereas the Msh2–Msh6 complex prefers to bind to a single base mismatch (Palombo *et al.*, 1996; Bellacosa, 2001).

Transgenic mouse data in DM1 have shown that the MutL protein, Pms2, is also required for somatic expansions (Gomes-Pereira *et al.*, 2004). However, Pms2 does not appear to be absolutely required for expansions, as in deficient mice, there is a reduction of somatic expansions rate by ~50%, relative to the wild-

type mice. Indeed, the requirement of DNA MMR genes, such as *Msh2*, *Msh3* and *Pms2* in mediating repeat expansions strongly implicates the DNA mismatch repair pathway rather than replication slippage. The proposed model that can account for these observations based on the involvement of MMR in an expansion-biased manner is shown in Figure 1-1. This supports the role of the MMR pathway and the possible requirement of other repair pathways as candidate modifiers of somatic instability. MMR proteins are thus potential therapeutic targets in HD and DM1.

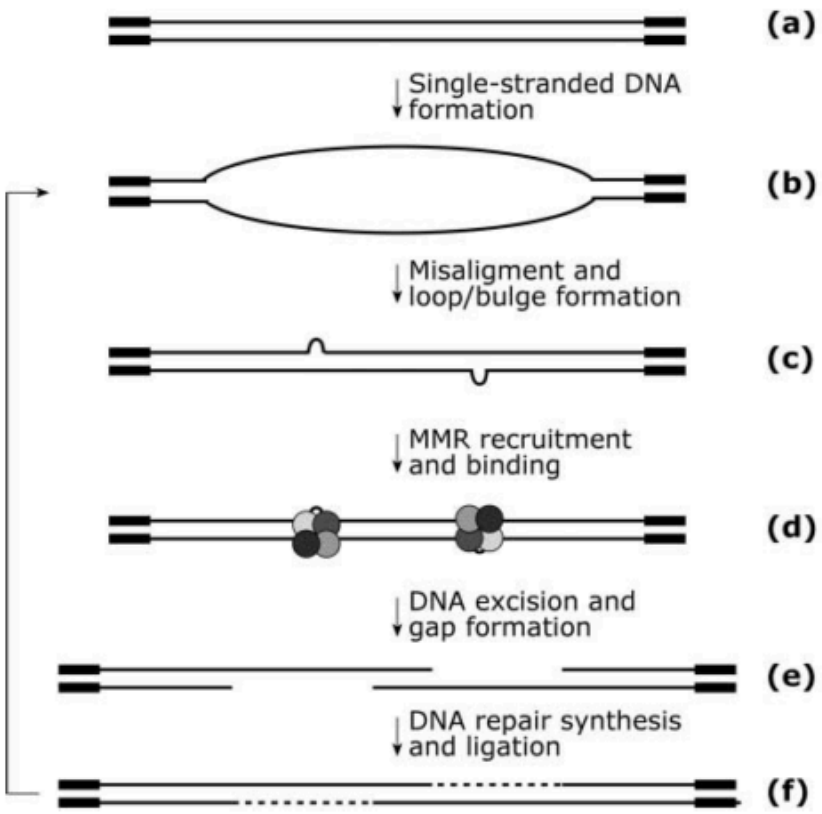

Figure 1-1 A model for the mismatch repair (MMR) involvement in the somatic trinucleotide repeat expansion and deletion in non-dividing cells. The figure represents the process of loop formation and expansion repair by inappropriate DNA MMR. In a double-stranded DNA molecule (a), both strands containing the repeat are separated (b), when the repeats re-anneal, they could cause a misalignment within the repeat region (c). Such misalignment will lead to the formation of loop-out structures. These loop-outs can be targeted by the MMR machinery (d). The correct template strands cannot be distinguished by MMR machinery in the absence of any replication specific signals. Therefore, the MMR machinery might be biased towards loop-outs incorporating DNA rather than deleting it, and that probably requires a single-stranded nick on the opposite strand to create a site of gap extension (e). Then, the single-stranded gaps are filled by polymerase extension and mutant expanded products are generated that are larger by the size of the original loop-outs (f). Multiple rounds of this process may occur leading to large expansions in non-dividing somatic cells. Flanking DNA region are indicated by thick lines, CTG-CAG repeat as thin lines, newly synthesized DNA as dashed lines and MMR enzymes are shaded circles. The image is reproduced from (Gomes-Pereira *et al.*, 2004).

Consistent with the role of MMR genes in mediating somatic instability in mouse models, it also has been shown that polymorphisms in MMR genes modify the expansion dynamics in DM1 (Morales *et al.*, 2016) and the disease progression and age at onset in HD (GeM-HD Consortium, 2015). There is evidence from DM1 patients for an association between a polymorphism in the *MSH3* gene and variation in the degree of somatic instability in blood DNA from Costa Rican patients, although no association with age at onset was detected (Morales *et al.*, 2016). This polymorphism is associated with an increase in somatic expansions in DM1 patients. This finding implicates the role of the DNA repair pathway in the variation of somatic instability in DM1 patients.

In human HD, there is a clear link between the genes involved in DNA mismatch repair and the disease process from Genome-wide association study (GWAS). GWAS are a powerful method of evaluating genetic modifiers that could affect the disease progression. The study involves assessing the entire genome or part of it, to determine genetic variants in the genome, such as deletions, insertions and single nucleotide polymorphism (SNPs) that associate with the disorder understanding. The utility of this unbiased approach has great power to identify naturally occurring genetic variations that modify the disease process.

GWAS was performed by the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. They discovered significant modifiers associated with modification of age at onset of motor signs (GeM-HD Consortium, 2015). GWAS was carried out on 2,131 individuals of European ancestry. Analysis of the GWAS data revealed a genome-wide significant signal in new genetic modifiers of the disease onset that are located in the *FAN1* and *MTMR10* as well as the *RRM2B* genes. In addition, a further interesting locus on chromosome 3 is the *MLH1* gene. Although this locus did not achieve genome-wide significance, *MLH1* was revealed as a potential modifier. This locus is particularly attractive given that *MLH1* is known to regulate somatic instability of CAG repeats and is a modifier of the phenotype in mouse HD knock-in models (Pinto *et al.*, 2013). This study supports the role of MMR genes in modifying HD pathogenesis and implicates DNA mismatch repair as a process of HD modification.

The association between age at onset and those candidate genes in the DNA repair pathway identified by GWAS were replicated and validated in a separate

study of HD and polyglutamine spinocerebellar ataxia (SCA) patients (Bettencourt *et al.*, 2016). The study involved 1,462 patients of HD and SCAs type 1, 2, 3, 6, 7 and 17. SNP analysis, which tested the effect of 22 SNPs on age at onset, was examined. The analysis of the combined samples for all polyglutamine diseases (HD and SCAs) yielded a significant association between age at onset and polymorphism at *FAN1* and *PMS2* and for all the SCAs patients yielded association with only *PMS2*. The results revealed the involvement of DNA mismatch repair genes in modifying age at onset and disease progression in HD patients. In addition, this study suggested that somatic instability might play a role that modifies the age at onset of other polyglutamine diseases.

Although GWAS analysis in HD did not evaluate the correlation between the DNA repair genes and somatic instability, several analyses have shown that DNA mismatch repair genes are required to generate repeat expansions in mice. In addition, the candidate gene study suggested that somatic expansion mechanisms are most likely explained by the genetic variations that alter disease progression and age at onset of polyglutamine disorders (Bettencourt *et al.*, 2016). This evidence supports the hypothesis that somatic expansion is mediated by DNA mismatch repair. MMR proteins are thus credible therapeutic targets in trinucleotide repeat disorders.

1.8 Molecular methods of HD genotyping

1.8.1 Polymerase chain reaction (PCR)

The standard diagnostic method for assessing trinucleotide repeat length is based on fragment length analysis using polymerase chain reaction (PCR). A PCR based method was originally developed using a primer set that amplifies a region containing both the CAG and CCG repeats (Warner, Barron and Brock, 1993). The assessments of CAG repeat length assumed that the CCG repeat did not demonstrate any variation. However, in both normal and affected individuals, the CCG repeat varies in size between 7 to 10 repeats in most cases (Andrew *et al.*, 1994), and also the CCT repeat following the CCG repeat can vary from 2 to 3 repeats in size (Pêcheux *et al.*, 1995). If the CAG repeat size is estimated by this PCR fragment size, there may be errors in the identification of the mutation in individuals carrying borderline CAG repeats, because of the variable size of

the CCG repeat repeats. Therefore, using a PCR assay that amplifies the CAG and the CCG repeats independently was proposed (Andrew *et al.*, 1994). This assay is important for detection of a second allele in cases where only a single allele is detected by the CAG amplification. This may be used to rule out the presence of a rare mutation that can lead to failure of the standard PCR assay for repeat sizing. The PCR protocol consists of PCRs that amplify the CAG tract alone, the CCG tract alone and the whole region containing both the CAG and the CCG repeats (Andrew *et al.*, 1994).

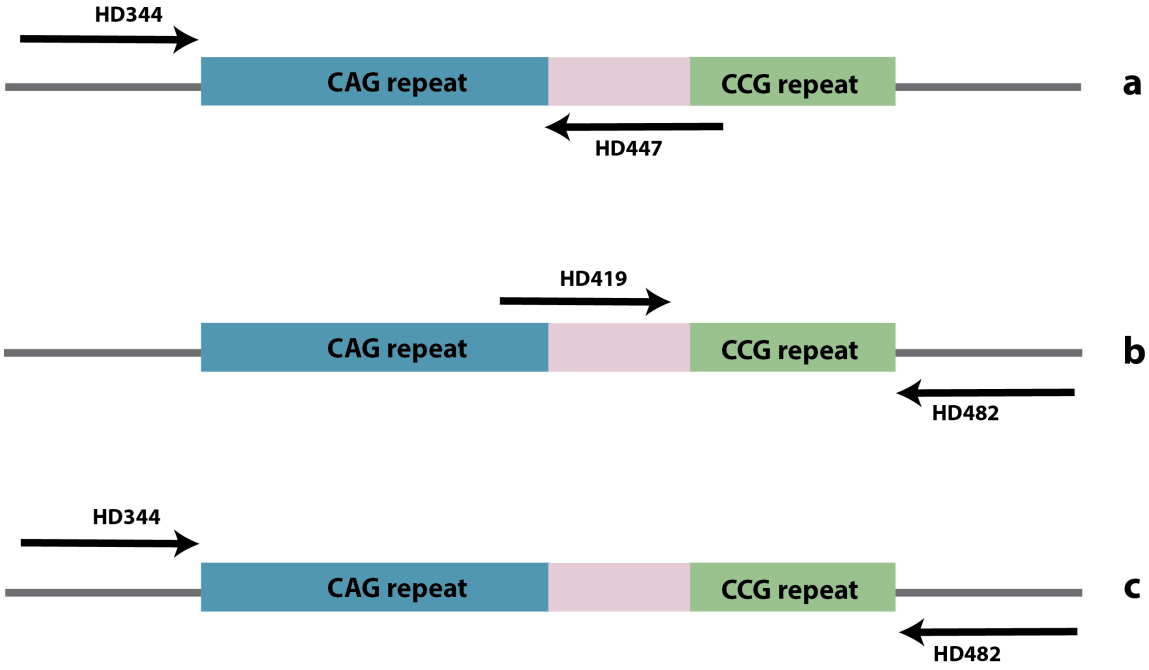

Figure 1-2 Schematic representation of the CAG and CCG polymorphic region in the HD gene and primers used for the determination of estimated CAG and CCG repeats. Both CAG and CCG repeats are shown and also the primer locations are indicated by arrows for the corresponding sequences. The region between CAG and CCG repeats is the intervening sequences: CAACAGCCGCCA. This method involves three PCRs that amplify the CAG repeats alone using HD344 and HD447 primers (a), the CCG repeat region using HD419 and HD482 primers (b) amplify both CAG and CCG repeats region using HD344 and HD482 primers (c). Method and the primers are indicated as in the study by Andrew et al.1994.

The length of the CAG repeat can be determined from DNA samples with primers HD344 and HD447 (Figure 1-2). The adjacent CCG repeat was measured using HD419 and HD482 primers and the entire region containing both the CAG and the CCG repeats using HD344 and HD482 primers as described by Andrew et al. 1994. Then, the amplified products sizes were determined by capillary or gel electrophoresis.

This method is considered as a rapid method for diagnosis of HD. However, large expanded alleles (>100 CAG repeats) do not amplify efficiently, and they cannot be detected using this approach (Warner *et al.*, 1996). Therefore, a true normal homozygous individual cannot be differentiated from heterozygous patients with one normal allele and one large allele of such an expanded allele. Thus, the standard PCR method is not capable of detecting very large expansion and also a new mutation within or flanking the repeats. Such mutations can lead to assay failure, or inaccurate estimation of CAG repeats length.

1.8.2 Triplet repeat primed PCR (TP-PCR)

Use of triplet repeat primed (TP-PCR) is required for the detection of large repeat expansions. This was first introduced by Warner *et al*. 1996. TP-PCR provides an advantage in rapidly identifying large repeat expansions that are difficult to amplify with standard PCR. The TP-PCR assay is based on amplifying the repeats using a fluorescently labelled forward primer (P1) paired with two specific reverse primers (P3 and P4) as shown in Figure 1-3 (Warner *et al.*, 1996). These two reverse primers bind to different positions within the repeats. The P4 primer has a 5' extension with identical sequence to another reverse primer (P3). This technique generates different fragment sizes giving a characteristic ladder on a fluorescence electropherogram, enabling the identification of both the size of smaller alleles and the presence of or absence of expanded alleles. The products obtained are resolved by capillary electrophoresis and visualised using a laser fluorescent sequencer. Traces obtained from the fluorescent sequencer are analysed for detection of the repeat.

Although TP-PCR is a robust and reliable method that also rapidly detects the larger expanded repeats, it does not provide an accurate estimate of the expanded repeat size. The need for very exact measurement of repeat length is important for predicting the risk of the individuals. In addition, this approach does not measure the somatic instability that could underlie disease severity.

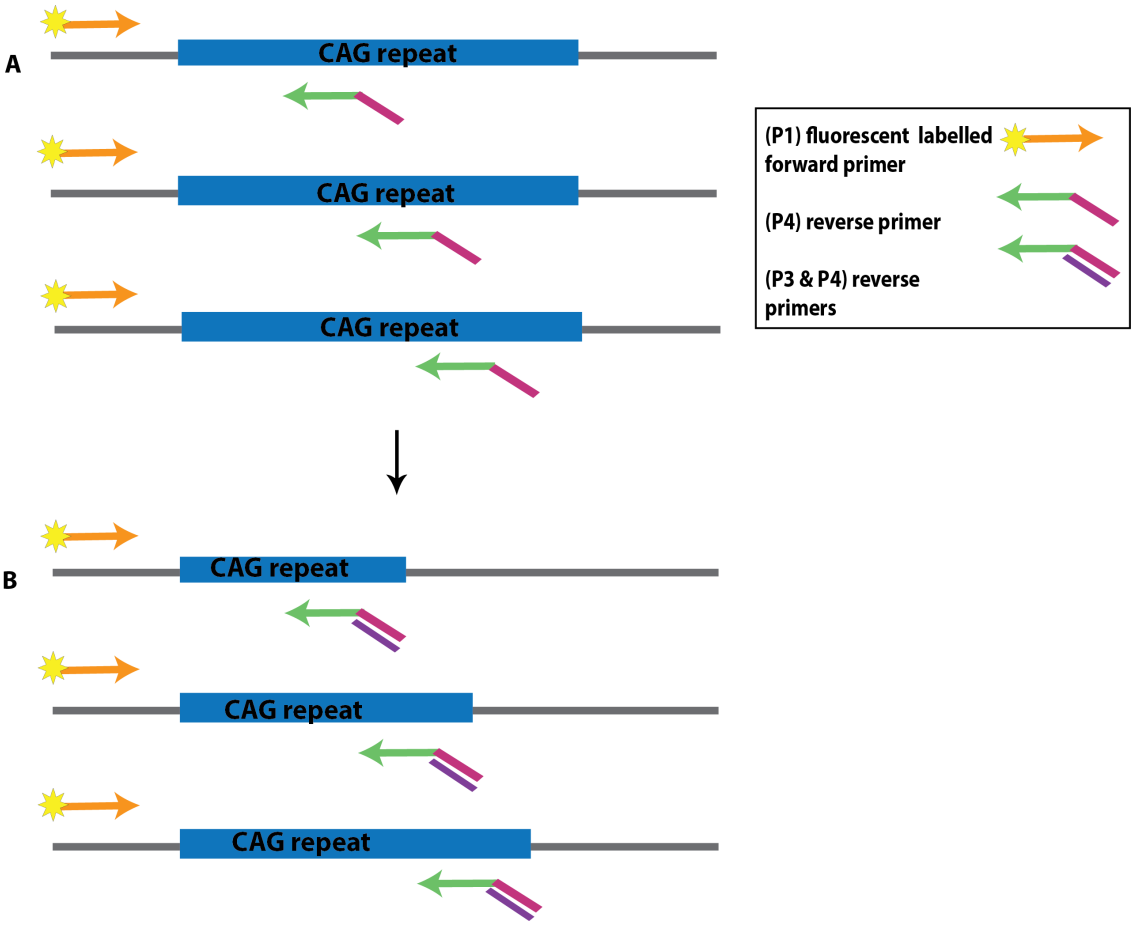

Figure 1-3 The Triplet repeat primed PCR (TP-PCR) method for detecting large repeat expansions. (A) Specific flanking fluorescent-labelled forward primer (P1) and CAG specific reverse primers (P4) are used. P4 is a long primer which is complementary to the CAG repeat and has common 5'sequnce (tail) that is not complementary with the triplet repeat region or any known human sequence. P4 binds at multiple sites within the repeat during the first multiple cycles. This will give rise to a mixture of products of many different sizes. (B) The second reverse primer (P3) binds to the 3' end of products that are complementary to the tail of P4 sequences, and pairs with (P1) forward primer in subsequent cycles to complete the extension of full-length products with variable sizes. This allows for rapid identification of large pathogenic repeats. Image modified from Warner *et al.* 1996.

1.8.3 Small pool PCR (SP-PCR)

Small pool PCR (SP-PCR) is a highly sensitive technique used in analysing mutant alleles and assessing somatic instability (Gomes-Pereira et al. 2004). This method is based on diluting bulk DNA templates to the level of a few DNA molecules in each reaction. This allows the amplification and detection of products derived from single-input molecules (Figure 1-4). The advantage of SP-PCR over the standard PCR method relies on amplifying very few molecules allowing the generation of a detailed distribution of repeat sizes present in any given sample to be assessed.

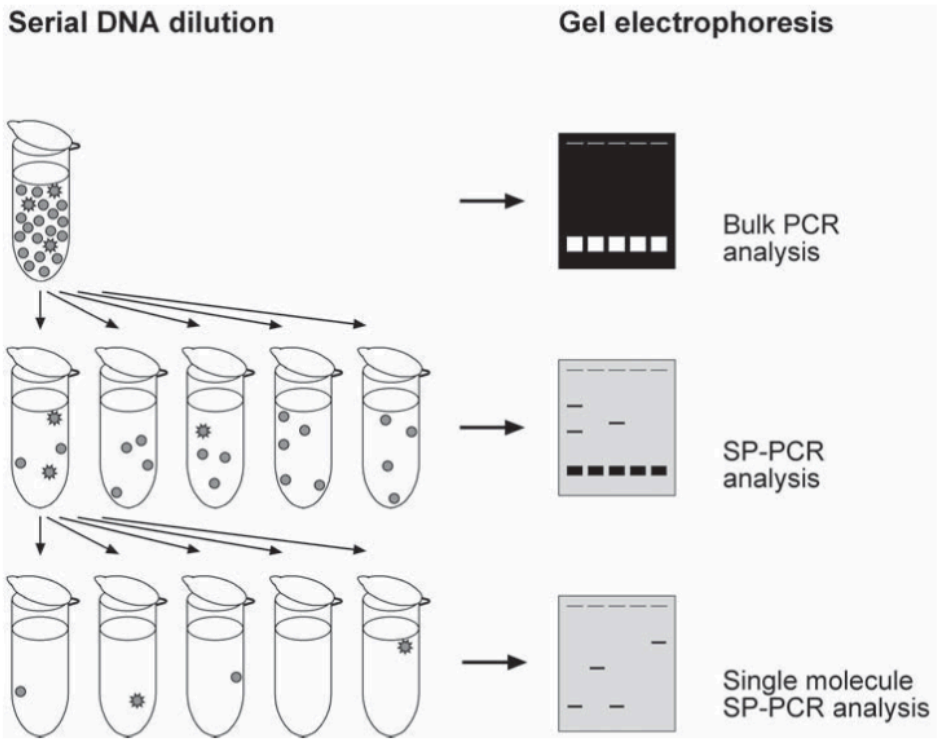

Figure 1-4 Small pool PCR (SP-PCR) method. The top section represents traditional bulk PCR for DNA sample that contains both normal alleles (circles) and a low number of mutant alleles (stars). In bulk PCR analysis, only normal alleles are revealed at the bottom of the gel (white bands). This PCR fails to reveal any mutant alleles in an agarose gel. The middle section represents a serial dilution of the template DNA sample and the expanded alleles are detected by Southern blot hybridization as black bands of variable sizes. The normal allele is also detected at the bottom of the gel (intense black bands). In this step, each SP-PCR amplification has a few DNA molecules as input. The bottom of the figure represents a further dilution of template DNA to the single molecule level per reaction in which the normal and expanded alleles are detected as sharp bands on a gel. This results from amplification of single molecule per reaction generating a single allele per lane. This image is taken from Gomes-Pereira et al, 2004.

A previous HD study had shown SP-PCR and agarose gel-based methods could not quantify somatic instability in the majority of cases (Veitch *et al.*, 2007). Therefore, a sensitive assay based on single molecule PCR was designed for HD to investigate mutation length changes in given samples, which allows for accurate sizing of a single repeat difference (Veitch *et al.*, 2007). For each individual, a DNA sample was diluted to produce an average of 0.5 molecules of template. Then, DNA templates are amplified by PCR. These amplified products are resolved by agarose gel electrophoresis to detect the presence of expanded alleles. Agarose gel electrophoresis is used in this step, as it is more cost-effective than polyacrylamide gel electrophoresis. Then, those products from the expanded HD alleles are re-amplified by PCR with fluorescently nested primers. Then, subsequent detection of both normal and expanded alleles can be sized using the capillary-based system.

The allele length distribution obtained from SP-PCR analysis allows for assessing the level of somatic instability. This approach allows estimating of the repeat size and can reveal the full range of mutation lengths present in a given tissue or cell sample. However, SP-PCR is labour intensive and time-consuming. In addition, this approach cannot detect any variants within or flanking the repeats, which may alter the stability of DNA, and possibly the symptoms. Therefore, new technologies are required that are more rapid and involving higher throughput.

1.9 Massively parallel sequencing technologies

The standard diagnostic and research approaches used in genotyping HD alleles (described earlier in the previous sections of this chapter) have some limitations in terms of sensitivity, accuracy and measuring somatic instability because they are not precise in determining allele sizes and fail to define the extent of somatic instability. In addition, these approaches cannot detect any variant repeats within or flanking the repeats. There is a demand for a cost-effective, rapid and a higher throughput approach, for instance, massively parallel sequencing using next generation sequencing (NGS) technology. The advantage of NGS over conventional approaches lies mainly in the cost per base, analysis duration and accurate quantification of allele length variations. However, technical challenges in sequencing and data analysis must be taken into account when using NGS.

It would be worthwhile to evaluate the utility of this technology as an approach to genotype HD alleles, detect sequence variants and quantify somatic instability of the CAG repeat.

1.9.1 Illumina MiSeq technology

There are many platforms of high throughput sequencing that could be used in NGS that vary in sequencing chemistry and read length (Quail *et al.*, 2012). Several sequencing platforms that could be used in NGS include Illumina (MiSeq and HiSeq), Ion Torrent Personal Genome Machine (PGM) and the Pacific Biosciences (PacBio). Each of these platforms has relative advantages and disadvantages and, the technical specification of each of these platforms are

summarized and compared in Table 1-1. In this project, the main factors for selecting the appropriate platform for HD study are read length and throughput.

NGS platforms	Ion Torrent PGM	PacBio	Illumina HiSeq	Illumina MiSeq
DNA	10 ng	300 ng	50 ng	50 ng
Total output	~ 1 GB	~ 375 MB	50-1,000 Gb	15 Gb
Run time	2-7 hrs	30-180 min	1-6 days	4-55 hr
Read length	up to 400 bp	~ 15 Kb	2x125 bp	up to 600 bp
Total reads	up to 5 million reads	75,000	2 Billion reads	25 million reads
Advantages	−Cheap (cost per base) −Available in many diagnostic Lab	long read > 10 Kb	High capacity	−Median capacity. −Long read lengths up to 600 bp.
Disadvantages	Not efficient in sequencing expanded repeats (> 20 repeats) (our lab's experience of sequencing DM1 alleles)	−Low through put −High error rate −Expensive platform (cost per base)	−Reads are relatively short −Not efficient for repeated sequences	Unable to sequence very large expanded alleles (>600 bp)

Table 1-1 Comparison between next generation sequencing (NGS) platforms. These platforms are Ion Torrent Personal Genome Machine (PGM), the Pacific Biosciences (PacBio) RS and Illumina (MiSeq and HiSeq) (Liu *et al.*, 2012; Quail *et al.*, 2012).

The Illumina MiSeq has a fast turnaround time and outputs intended for targeted sequencing of small genomes. MiSeq has been developed to be used in smaller laboratories and diagnostic clinics and produces low error data for the most complex sequencing samples. The developed Illumina MiSeq platform generates sequence output of 15 GB of high-quality sequence data, which consists of up to 25 million sequencing reads. The read length is up to 600 bp, which should be capable of sequencing a large HD allele. With MiSeq, the unique barcodes must be incorporated into the DNA template to be run on MiSeq platform. MiSeq libraries may be prepared with PCR amplification using locus-specific primers incorporating MiSeq adapters plus a unique dual index bases in order to enable differentiation between each DNA sample (Caporaso *et al.*, 2011). These fragments are added to the flow cell and hybridised to complementary sequences that are located on the surface of the flow cell (Bentley *et al.*, 2008).

The fragments are then amplified to generate millions of clusters through a bridge amplification process to form DNA clusters from each fragment.

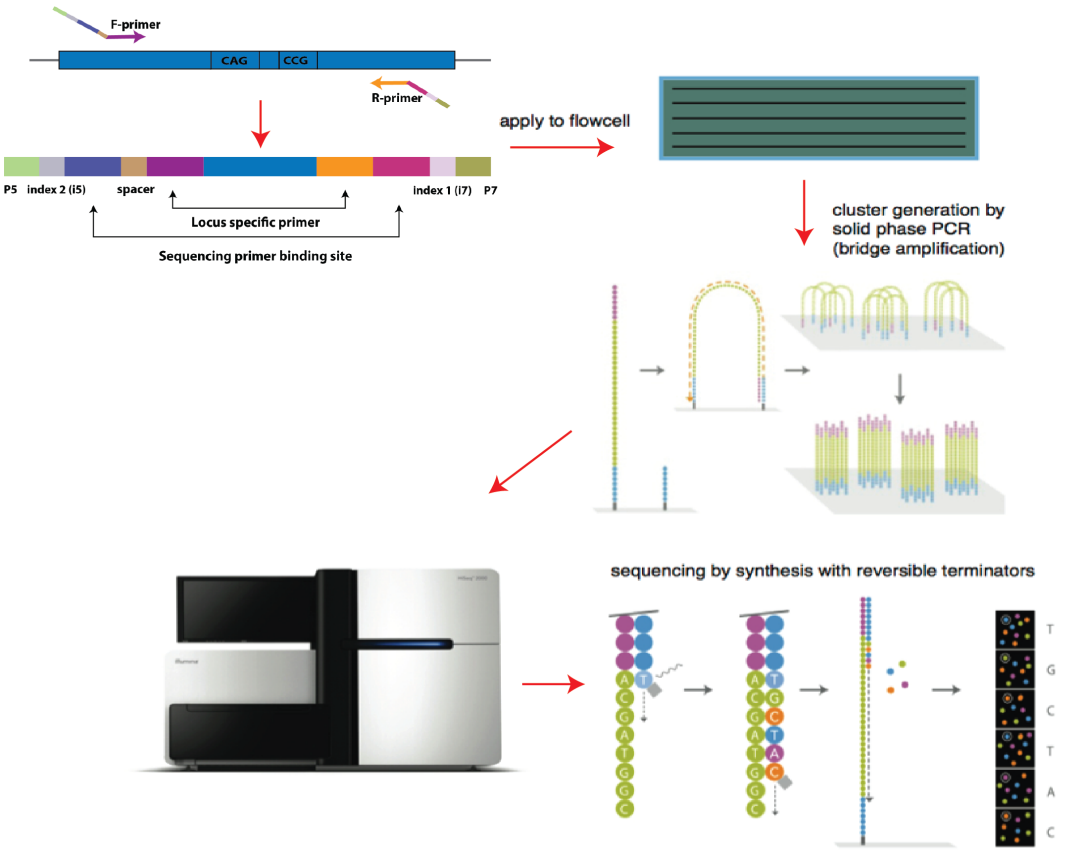

Figure 1-5 Illumina MiSeq protocol for sequencing *HTT* alleles. The locus-specific primers combined with MiSeq adapters that are used for amplifying *HTT* allele and allow the amplicon to be sequenced in the MiSeq platform. The DNA fragment contains CAG and CCG repeats with P5 and P7 sequences, indices, spacer, sequencing primer binding site and locus-specific primers. P5 and P7 sequences are oligos at the end of each library fragment which anneal to their complementary oligos immobilised on the flow cell surface. Index (i7) and (i5) are used to identify each sample. The sequencing primer site is complementary to the sequencing primer on the MiSeq platform. The spacer is used to increase base diversity and for cluster identification on MiSeq sequencing. The MiSeq library is prepared by PCR amplification, and the adapters are incorporated to both fragment ends. These amplicons are loaded into a flow cell and hybridized to the flow cell surface. Each fragment is amplified by bridge amplification and generates clusters from each molecule. Each cluster is sequenced using fluorescently labelled, reversible terminator nucleotides. The signal is generated upon incorporation of the first nucleotide and detected through the imaging system. Before the next cycle proceeds, the fluorophore from each incorporated base is removed to allow the incorporation of the next base. This cycle is repeated to create a read from the whole molecule. Image modified from (https://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/).

The sequencing chemistry of MiSeq is based on the sequencing-by-synthesis approach utilising fluorescently labelled reversible terminator nucleotides for sequencing a single-stranded DNA template (Liu *et al.*, 2012). Identification of

nucleotides is based on a cyclic reversible termination strategy, which sequences one nucleotide from the template strand at a time through multiple rounds of base incorporation, imaging and cleavage followed by cycle termination. These fluorescently labelled ddNTPs are used to stop the polymerization reaction, enabling removal of unincorporated bases and imaging to identify the added nucleotide. This process continues in a similar manner allow the determination of the sequence of the whole DNA template.

Illumina MiSeq can potentially be used for genotyping HD alleles because a very large number of HD alleles may be sequenced simultaneously. This is mainly because it is high throughput, relatively low cost and has a suitable read length (up to 600 bp single read or 2x300 bp paired-end reads).

1.10 PhD hypothesis

Huntington disease is an extremely variable inherited human disorder caused by expansion of CAG trinucleotide repeats that are unstable in both germline and soma. In this project, I will investigate the use of next generation massively parallel DNA sequencing technologies to detect the presence and distribution of additional sequence variants within and flanking the repeat. These data will be used further to clarify the relationship between genetic diversity and symptomatic variation and will lead to the development of improved diagnostic tests.

Previous evidence from studies of repeat instability in mouse models and human HD patients suggested that somatic instability might be a major modifier of HD pathogenesis (Kennedy and Shelbourne, 2000; Kennedy *et al.*, 2003; Veitch *et al.*, 2007; Swami *et al.*, 2009). Therefore, we sought to test the hypothesis that the somatic instability of the HD CAG repeat is driving the disease pathology and to understanding the role of somatic instability in HD by quantifying somatic variation in the number of CAG repeats by NGS technology. In addition to genotyping *HTT* alleles and quantifying somatic instability by NGS technology, we will determine whether sequence polymorphisms in potential *trans*-acting modifier genes are involved in modulating somatic instability in HD patients.

Chapter 2 Materials and methods

2.1 Material

The majority of molecular biology reagents were supplied by Sigma and ThermoScientific. Other materials are mentioned in the appropriate section. All kits and reagents were used according to the manufacturer's instructions unless otherwise stated.

2.2 DNA samples

Blood DNA samples were collected from a subset of Scottish DM1 patients who were recruited from outpatient clinics and sent to our lab for DM1 genetic studies. Patients consented to participate in DM1 genetic variations (DMGV) study led by Professor Darren G Monckton at the University of Glasgow and also consented to participate in other genetic studies in the laboratory. Therefore, we chose to use the DMGV DNA samples as control samples to amplify and genotype non HD-causing alleles. These samples were selected because they were available in our laboratory for research studies and also because they have no known association with HD. Therefore, the DMGV samples can represent a random population in respect to HD locus.

Buccal cell swab DNA samples were collected by the US-Venezuela Collaborative Research Project from 767 unaffected and affected individuals from HD families that are part of the Lake Maracaibo cohort in Venezuela. Samples were air dried and shipped from Venezuela to Glasgow. In Glasgow, the samples were stored at -20°C and then DNA was extracted from buccal swabs.

2.3 Molecular methods

2.3.1 Primer design

Sequences of all primer oligonucleotides used in this project were purchased from Sigma. The locus-specific primer sequences for amplifying the HTT CAG/CCG repeat are listed in Table 2-1.

PCR set of primer	Forward primer sequences (5'-3')	Reverse primer seqeunces (5'-3')	Annealing temperature
HD319 / 33935.5	GCGACCCTGGAAAAGCTGATGA	AGCAGCGGCTGTGCCTGC	58.5°C
MS-1F/ MS-1R short	GCCCAGAGCCCCATTCATTG	GCCATCCCCGCCGTAGCC	59°C
31329 /33934	ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC	GGCGGCTGAGGAAGCTGAGGA	59°C

Table 2-1 Sequences for the locus-specific set of primers that were used in our study for amplifying the *HTT* CAG/CCG repeat.

MiSeq primers were designed by incorporating Illumina adapter sequences at the 5'end of the locus-specific primers (HD319/33935.5, MS-1F/MS-1R short and 31329/33934). These locus-specific primer sequences are shown in Table 2-1. The MiSeq primers were designed to include the P5 and P7 sequences, 8 base index sequences (Index 1 and 2), forward and reverse sequencing primer-binding site (Table 2-2). In addition, spacers of different length (0-9 bases) were added between the sequencing primer binding site and locus-specific primer in each forward primer set. These components of MiSeq primers are described in Figure 1-5. An example of MiSeq primers designed is shown in Table 2-2 using the HD319F/33935.5 set of primer as locus-specific primers. The MiSeq primer design was carried out in the same way for the other locus-specific primers (MS-1F/MS-1R short and 31329/33934).

Table 2-2 MiSeq primers designed for amplifying the HD locus. These primers include P5 and P7 sequences, 8 base index sequences (Index 1 and 2), forward and reverse sequencing primer-binding site, a spacer (0-9 bases) and locus-specific primers.

MiSeq Forward primer name	P5	index (i5)	sequencing primer binding site	spacer	PCR primer	bp	MiSeq primer
MiSeq F HD319F D501	AATGATACGGCGACCACCGAGATCTACAC	TATAGCCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT		GCGACCCTGGAAAAGCTGATGA	96	AATGATACGGCGACCACCGAGATCTACACACACTATAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCGACCCTGGAAAAGCTGATGA
MiSeq F HD319F D502	AATGATACGGCGACCACCGAGATCTACAC	ATAGAGGC	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAC	GCGACCCTGGAAAAGCTGATGA	95	AATGATACGGCGACCACCGAGATCTACACATAGAGGCACACTCTTTCCCTACACGACGCTCTTCCGATCTTACGCGACCCTGGAAAAGCTGATGA
MiSeq F HD319F D503	AATGATACGGCGACCACCGAGATCTACAC	CCTATCCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AT	GCGACCCTGGAAAAGCTGATGA	94	AATGATACGGCGACCACCGAGATCTACACCCTATCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCGACCCTGGAAAAGCTGATGA
MiSeq F HD319F D504	AATGATACGGCGACCACCGAGATCTACAC	GGCTCTGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGAT	GCGACCCTGGAAAAGCTGATGA	96	AATGATACGGCGACCACCGAGATCTACACGGCTCTGAACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATGCGACCCTGGAAAAGCTGATGA
MiSeq F HD319F D505	AATGATACGGCGACCACCGAGATCTACAC	AGGCGAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	G	GCGACCCTGGAAAAGCTGATGA	93	AATGATACGGCGACCACCGAGATCTACACAGGCGAAGACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCGACCCTGGAAAAGCTGATGA
MiSeq F HD319F D506	AATGATACGGCGACCACCGAGATCTACAC	TAATCTTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TATTA	GCGACCCTGGAAAAGCTGATGA	97	AATGATACGGCGACCACCGAGATCTACACTAATCTTAACACTCTTTCCCTACACGACGCTCTTCCGATCTTATTAGCGACCCTGGAAAAGCTGATGA
MiSeq F HD319F D507	AATGATACGGCGACCACCGAGATCTACAC	CAGGACGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CTACAT	GCGACCCTGGAAAAGCTGATGA	98	AATGATACGGCGACCACCGAGATCTACACCAGGACGTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTACATGCGACCCTGGAAAAGCTGATGA
MiSeq F HD319F D508	AATGATACGGCGACCACCGAGATCTACAC	GTACTGAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACTATAT	GCGACCCTGGAAAAGCTGATGA	99	AATGATACGGCGACCACCGAGATCTACACGTACTGACACACTCTTTCCCTACACGACGCTCTTCCGATCTACTATATGCGACCCTGGAAAAGCTGATGA
MiSeq reverse primer name	**P7**	**index (i7)**	**sequencing primer binding site**	**spacer**	**PCR primer**	**bp**	**MiSeq primer**
MiSeq R 33935.5 D701	CAAGCAGAAGACGGCATACGAGAT	CGAGTAAT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D702	CAAGCAGAAGACGGCATACGAGAT	TCTCCGGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATTCTCCGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D703	CAAGCAGAAGACGGCATACGAGAT	AATGAGCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATAATGAGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D704	CAAGCAGAAGACGGCATACGAGAT	GGAATCTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATGGAATCTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D705	CAAGCAGAAGACGGCATACGAGAT	TTCTGAAT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATTTCTGAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D706	CAAGCAGAAGACGGCATACGAGAT	ACGAATTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATACGAATTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D707	CAAGCAGAAGACGGCATACGAGAT	AGCTTCAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATAGCTTCAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D708	CAAGCAGAAGACGGCATACGAGAT	GCGCATTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATGCGCATTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D709	CAAGCAGAAGACGGCATACGAGAT	CATAGCCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATCATAGCCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D710	CAAGCAGAAGACGGCATACGAGAT	TTCGCGGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATTTCGCGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D711	CAAGCAGAAGACGGCATACGAGAT	GCGCGAGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATGCGCGAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC
MiSeq R 33935.5 D712	CAAGCAGAAGACGGCATACGAGAT	CTATCGCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	–	AGCAGCGGCTGTGCCTGC	85	CAAGCAGAAGACGGCATACGAGATCTATCGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGCAGCGGCTGTGCCTGC

2.3.2 Amplifying the *HTT* CAG/CCG repeat locus using MiSeq primers

The HD CAG and the adjacent CCG repeats were amplified from genomic DNA using the locus-specific primer pair (HD319F/33935.5, MS-1F/MS-1Rshort or 31329/33934) combined with MiSeq barcoded Illumina adapters that allow the PCR product to be sequenced using MiSeq.

PCR amplification was carried out for 2 µl DNA (10 ng), with 2 µl of each primer (5 µM), 0.2 µl of 1 U of Taq DNA polymerase (Sigma), 1.5 µl 1X Custom PCR Master mix (45 mM Tris-HCl (pH 8.8), 11 mM $(NH_4)_2SO_4$, 4.5 mM $MgCl_2$, 0.113 mg/ml BSA, 0.048% 2-mercaptoethanol, 4.4 µM EDTA, 1 mM each of dATP, dCTP, dGTP and dTTP) (Thermo Scientific, ABgene UK), 1.5 µl of DMSO (BioReagent, for molecular biology, ≥99.9%, Sigma-Aldrich) and 5.8 µl of water in a final volume of 15 µl.

PCR was carried out using a thermal cycler (Eppendorf). The PCR cycle conditions were 96°C for 5 minutes followed by 28 cycles of DNA denaturation at 96°C for 45 seconds, primer annealing at 58.50°C for 45 seconds, polymerase extension for 3 minutes at 72°C, with a final extension step at 72°C for 10 minutes. The HD locus-specific primers used were HD319 and 33935.5.

MS-1F/MS-1R and 31329/33934 primers have the same PCR conditions, as HD 319 and 33935.5, except the annealing temperatures were 59°C for both of them.

2.3.3 MiSeq library preparation

Different types and combinations of indices can be used for Illumina sequencing. Dual indexed libraries were prepared with Illumina TruSeq adapters by adding unique indices to each sample. Index 1 (i7) is an 8 base index adjacent to P7 on the reverse primer. Index 2(i5) is an 8 base index adjacent to P5 on the forward primer. These unique indices act as a unique tag to each DNA sample so sequencing reads can be associated to a DNA sample. Therefore, there are 96 unique combinations generated from the pairwise combination of these indices that are uniquely dual-indexed adapters (these 12 (i7) and 8 (i5) indices). Taking

advantage of these 96 TruSeq barcodes from Illumina, the PCR were manually prepared in 96-well plates (8 forward barcodes x 12 reverse barcodes = 96 barcode pairs) as described in Chapter 3. The PCR plates used was 96-well PCR plate, skirted, low profile (STAR LAB, E1403-5200) and those plates were sealed with 12 Strip PCR Caps, domed (STAR LAB, I1400-1200). After PCR amplification, all samples were pooled.

2.3.4 Gel electrophoresis

Visualisation of PCR products was conducted by running the products on 2% agarose gels allowing for separation of the DNA fragments by size. 2% agarose gel was prepared by adding 2 g of agarose to 100 ml of 0.5x TBE buffer. This mixture was heated until the agarose completely dissolved. Once cooled, ethidium bromide was added to give the final concentration of 0.2 μM that is widely used to visualise the DNA bands by staining the agarose gel. This mixture was poured into a gel cast containing the appropriate comb and left to set on an even surface for 15-30 minutes at room temperature. Once set the comb was removed, and the gel was placed in a tank and was immersed in 0.5x TBE buffer.

The PCR products were mixed with loading dye and loaded into the wells. The molecular weight marker used was a 1 kb ladder (60 ng/μl 1 kb ladder, 1X DNA loading dye in 1X TBE) that was added to the end lanes of each PCR agarose gel. Half of each PCR (7.5 μl) was mixed with 10x loading dye (0.6% (w/v) Orange G, 50% glycerol in H_2O) and loaded in each of the lanes. The no DNA controls were also loaded onto the gel. The loaded gel was run at 100-120 V at room temperature for the appropriate time to allow the separation of PCR products.

For PCR product visualisation, the gel was removed from the tank and placed on a transilluminator (UVP image store 7500 system). A gel documentation system with high sensitivity camera, which was placed above the transilluminator, was used to photograph the gel.

2.3.5 DNA purification

2.3.5.1 Gel extraction

QIAquick gel extraction kit (250, Qiagen) was used to clean-up DNA from gels. The kit is used for the clean-up of PCR products and provides spin columns, buffers and collection tubes. A gel purification method was used for some of the libraries after PCR amplification to purify PCR products from a gel and get rid of primer dimers. This purification procedure removes agarose, excess primers, nucleotides, enzymes and other impurities from DNA samples. Specialised binding buffers are utilized with this method to promote selective binding of DNA fragments within a particular size range. Then, the purified DNA for the library can be used for sequencing on MiSeq platform.

2.3.5.2 PCR clean-up by AMPure beads

PCR products from some of the libraries were also purified with Agencourt AMPure XP PCR purification system using a paramagnetic bead technology (Beckman Coulter). This AMPure purification step was performed to remove primer dimers and to concentrate the sequencing library. The PCRs for the whole pooled library were split into 2 tubes. One tube of the library was stored and saved to allow repeat of library preparation or sequencing if required. The remaining half of the 96 PCRs were also pooled into 2 pools of 48 samples. Each pool was then purified with AMPure XP beads (0.6 µl of beads/ µl of PCR product, elution in 100 µl) using a magnetic stand for AMPure XP beads for 1.5 ml tubes. This step was performed to get rid of the primer dimers and to concentrate the library. The two purified pools were then pooled together, and another round of AMPure XP bead clean-up was performed (1.6 µl of beads/ µl of PCR product, elution in 55 µl) to concentrate the library further.

An optimized buffer was used with Agencourt AMPure XP technique to promote a DNA size selection of 100 bp and larger to bind to the paramagnetic beads. Agencourt AMPure XP method is based on utilizing magnetic separation and requires no centrifugation or filtration. Excess primers, nucleotides and salts can be removed using a simple washing procedure. The resulting purified PCR products can be used for sequencing.

2.3.6 DNA quantification

The concentration of DNA samples and libraries were determined by Qubit fluorometer using the Qubit dsDNA HS (High Sensitivity) Assay Kit (ThermoFisher). The Qubit fluorometer method allows measuring double-stranded DNA concentration in the range of 10 pg/μL to 100 ng/μL.

In addition, Bioanalyzer (Agilent) analysis was used to evaluate the quality and quantity of the sequencing libraries, to check the fragment had the expected size, that primer dimers were absent and to estimate the molarity of the library. This Bioanalyzer analysis was performed at Glasgow Polyomics facilities at the University of Glasgow. Two DNA fragments of known size and molarity are run with each sample on the Bioanalyzer, which are lower and upper markers. These markers are used for estimating the DNA size of the tested samples. DNA quantification was done using the upper marker in which the area under the peak for this marker was compared with the sample peak areas. Therefore, the concentration of each sample can be estimated by comparing to the known concentration of the upper marker. At that stage, the aim is to have at least 25 μl of a sequencing library at 10 nM.

2.3.7 MiSeq run

The libraries were submitted to the Glasgow Polyomics facilities for sequencing using MiSeq platform following the standard guidelines for a MiSeq run. MiSeq Reagent Kit v3 (600 cycles) was used with the 600 cycles of sequencing being performed by 2x300 bp or 400 bp forward and 200 bp reverse and 5% PhiX spike-in. After the run (~56 hours) the MiSeq reporter software was used to demultiplex the reads based on the 96 TruSeq barcodes pairs. The MiSeq reporter software outputs the sequencing reads in fastq files, two (forward and reverse reads) for each of the 96 barcodes with a maximum output of 15 GB per run.

2.3.8 KASP assay library preparation

A pilot study for KASP (Kompetitive Allele Specific PCR, www.lgcgenomics.com) assay was conducted for genotyping 31 SNPs in 24 DNA samples including 22 HD patients from the Venezuelan cohort and two positive controls. These positive

controls were added into each plate for assay validation. These controls had good DNA quality and high concentration compared to the buccal swab DNAs from the Venezuelan cohort, but these were of unknown genotypes. Two no template controls were included into each plate to improve confidence in the validity of the genotyping results.

DNA aliquots of 95 µl were assembled in PCR plates for each sample and sent to the LGC Genomics laboratories in Hoddesdon, UK. Thermo Scientific 96-well PCR plates (AB-0800) and adhesive PCR plate sealing sheets (AB-0558) were used to send these samples. The KASP assay protocol was carried out in the LGC Genomics laboratories according to their protocol. The submitted DNA sample validation, primer design and optimisation of the assay conditions were also conducted by LGC Genomics laboratories.

From the pilot study result, 19 SNPs for the KASP assay were validated by the LGC laboratory. Therefore, these 19 SNPs were tested in 412 HD patient samples; an additional 59 unaffected individuals and 20 controls from the Scottish population were also genotyped. DNA volume of 60 µl was assembled in 96 PCR well plates and sent to the LGC Genomics laboratories.

2.4 Statistical data analysis software

Most of the graphs were prepared using GraphPad Prism 6. Some of these graphs were prepared using Microsoft Excel. Adobe Illustrator was used to creating most of the figures and diagrams. IBM SPSS (version 22) was used for the regression analysis and for most of the data analysis. IBM SPSS was used to perform linear regression analysis between somatic mosaicism and age at sampling, inherited allele length and their interactions for HD patients from the Venezuelan cohort. This software was also used for the genotype/phenotype correlation among HD patients by performing linear regression analysis between age at onset and inherited CAG repeats. An independent sample T-test was applied to evaluate the significant difference in the mean age at onset between homozygous and heterozygous individuals for the HD mutation, using IBM SPP. Similarly, the differences in mean age at onset were evaluated between males and females using an independent sample T-test (using SPSS).

Genotype/phenotype data analysis was performed by gPLINK (version2.050, http://zzz.bwh.harvard.edu/plink/gplink.shtml), graphical user interphase for the PLINK software. gPLINK performs a wide range of basic and large-scale genetic analyses. Allele frequencies, Hardy-Weinberg equilibrium tests, linear regression analyses and family-based association study were performed by gPLINK. gPLINK was also used for regression analyses to test the association of the residual variation of somatic instability with each SNP genotype. Also, a family-based quantitative trait association analysis (QFAM) was tested, that combines a linear regression of phenotype on genotype with a permutation test, which accounts for relatedness among family members. gPlink was used to test an association of residual variation in somatic instability across all SNPs. For multiple corrections for SNPs association with somatic instability, the Bonferroni multiple significance tests was used through gPlink software.

2.5 Web-based bioinformatics resources

Genetic databases used in this project to check for variant types, population allele frequency were: Ensemble genome database (https://www.ensembl.org/index.html), the 1000 Genome database (http://www.internationalgenome.org) and dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/).

2.6 Bioinformatics data analysis tools

2.6.1 CLC genomic workbench

This software was used for the alignment of sequencing reads generated by the Illumina MiSeq platform. The sequencing reads obtained were mapped with CLC Genomics Workbench Version 7.5 to reference sequences including the region flanking the repeats and various numbers of CAG and CCG repeats: 1 to 100 CAG repeats and 1 to 20 CCG repeats in order to facilitate alignment for the sequence and genotyping the *HTT* alleles. Also, CLC was used to apply quality control analysis on reads, read filtering (demultiplexing and adapter trimming), alignment parameter optimisation, visualisation of the aligned reads and genotype identification.

Demultiplexing was required in some MiSeq runs, if the library was sequenced with other samples from our group and barcodes were shared with other users. Also, trimming adapter step is essential for samples for which the insert size is below the specified sequencing length (if the insert size is 200 bp, and the MiSeq is sequencing 300 bp), the adapter will be sequenced. In this case, the sequences of corresponding library adapter can be present in the output files at the 3' end of the reads. Therefore, removing the adapter sequencing from these samples is detrimental to obtain a better read alignment.

2.6.2 Tablet

Visual representation of obtained sequencing read alignments (from CLC genomic workbench) was performed using Tablet. Tablet is a graphical viewer for NGS alignment (Milne *et al.*, 2013). Tablet displays read coverage, read names, and it allows searching for specific coordinates across the dataset.

2.6.3 KASP software (SNPviewer)

The SNPviewer software enables genotyping KASP data and for all SNPs to be viewed as a cluster plot for each plate (https://www.lgcgroup.com/products/genotyping-software/snpviewer/#.WmcLrhS3mOo).

The SNP genotyping data were analysed after being received from the genotyping service at LGC genomics laboratory. LGC genotyping results files are provided as SNPviewer file that is designed to be viewed with SNPviewer software. This allowed for visualising the genotyping clusters plate by plate graphically. SNPviewer was used in our project to genotype SNPs from candidate modifier genes of genetic instability in the Venezuelan cohort of HD patients.

Chapter 3 Non-disease associated allele genotyping using MiSeq sequencing

3.1 Introduction

The CAG size ranges were defined as follows: normal alleles have 9 to 26 CAG repeats and intermediate alleles have 27 to 35 CAG repeats (Goldberg *et al.*, 1995). Alleles with 36 to 39 CAG repeat show reduced penetrance, while alleles with 40 or more repeats are fully penetrant (Kremer *et al.*, 1994; Rubinsztein *et al.*, 1996). Most normal size CAG repeats are stable through inheritance, while several studies have shown that intermediate alleles are unstable and can be transmitted as an expanded allele upon transmission to offspring (Goldberg *et al.*, 1993; Semaka, Collins and Hayden, 2010). Nevertheless, individuals who carry intermediate alleles have a normal phenotype. There is a high frequency of people carrying intermediate alleles in the general population with no known association with HD (5.8%) (Semaka *et al.*, 2013). Small numbers of individuals with incomplete penetrance alleles manifest HD symptoms with later onset. The HD allele, in individuals carrying 40 CAG or more, demonstrates complete penetrance that gives rise to disease within the average lifespan. Also, disease alleles may either increase or decrease in length during transmission to the next generation (MacDonald *et al.*, 1993). CAG repeat length changes occur in transmissions from either parent, although large increases are mainly seen in paternal transmission (Leeflang *et al.*, 1995).

In patients with HD, the CAG repeat length is inversely correlated with the onset of clinical symptoms (Andrew *et al.*, 1993; Duyao *et al.*, 1993; Gusella and MacDonald, 2006; Andresen *et al.*, 2007), most juvenile cases of HD are transmitted from affected fathers and are associated with a large CAG repeat size (60 CAG or above) and a more severe phenotype (Telenius *et al.*, 1993; Nahhas *et al.*, 2005). The observation of increasing disease severity and decreasing age at onset with each successive generation, known as anticipation, is seen in HD and other trinucleotide repeat disorders (Telenius *et al.*, 1993). This phenomenon of anticipation arises mainly from CAG repeat instability during male transmission.

Although the molecular basis of HD is the expansion of CAG repeats, there is another polymorphic CCG triplet sequence identified downstream to the CAG tract (Barron *et al.*, 1994). CCG repeats vary from 6 to 12 repeats, but 7 and 10 CCG repeats are predominant in most populations (Andrew *et al.*, 1994; Squitieri *et al.*, 1994; Pêcheux *et al.*, 1995). The smallest CCG allele identified contains 4 CCG repeats in the Indian population (Pramanik *et al.*, 2000). CCG 7 is strongly associated with affected chromosomes in most populations where HD has a high prevalence (Europe and Venezuela)(Squitieri et al. 1994; Barron et al. 1994; Costa et al. 2006). In Japan, where HD has a low prevalence, HD chromosomes are strongly associated with an allele of CCG 10 (Morovvati *et al.*, 2008).

In an HD study, the main factors for selecting the appropriate platform are read length and throughput. The developed Illumina MiSeq platform generates sequence output of up to 15 GB of high quality sequence data, which consists of up to 25 million sequencing reads. The read length is up to 600 bp, therefore it should potentially be capable of sequencing a large HD allele (Illumina, 2006).

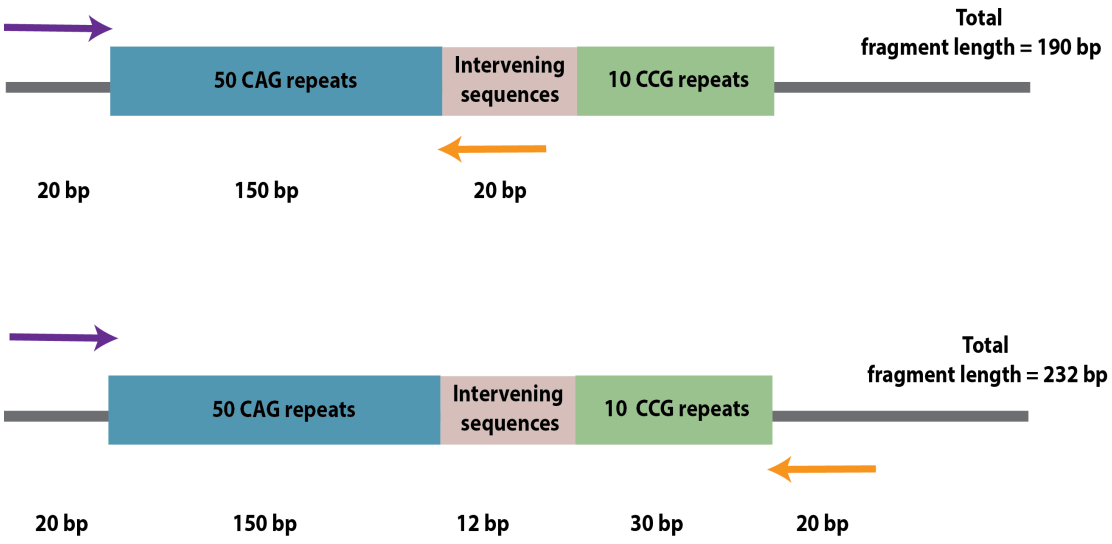

Figure 3-1 The minimum fragment length required for sequence based genotyping. The top allele structure indicates the minimum amplicon and fragment length that is necessary to be sequenced to genotype the CAG repeats. Assuming it is 50 CAG repeats, that would be 190 bp, if the primers were positioned immediately flanking the 5' prime of the repeat and spanning the intervening sequence on the 3' prime side. The intervening sequence is: CAACAGCCGCCA and is located between the CAG and CCG repeats. The bottom diagram shows that the minimum fragment size required to genotype 50 CAG and the CCG repeats assuming 10 CCG is 232 bp, if the primers were located immediately flanking the 5' prime of the CAG repeat and the CCG repeats on the 3' prime side.

For sequencing expanded alleles using MiSeq sequencing, there are several kits that can be used. The standard sequencing run for MiSeq is 2x300 bp. Most HD

patients have CAG repeats between 40 to 50 repeats. For instance, the length of an HD allele of 50 repeats will be 150 bp, and 20 bp for each primer pair positioned immediately flanking the 5' prime of the repeat and spanning the intervening sequence on the 3' prime side (190 bp) (Figure 3-1). Therefore, a minimum 190 bp reads are required to sequence the CAG repeats alone. For determining the intervening sequences and CCG repeats to identify any atypical allele structures, 232 bp read length are necessary for a patient with 50 CAG repeats and 10 CCG repeats (Figure 3-1). Therefore, a 2x300 bp run can be used to sequence most HD patients' DNA. However, theoretically sequencing longer CAG repeats (170 repeat alleles) could be done using a 600 bp run. Therefore 400 bp and 600 bp runs were tested to determine whether longer alleles could be genotyped.

Sequencing the repeat is challenging and can pose a problem for the assembly of short reads. Therefore, paired-end reads have proven to be a valuable solution to detect repetitive sequences by including part of flanking region in both the 5' and 3' end of the sequence. Paired-end sequencing allows users to sequence both ends of a fragment and generate high quality sequencing data. For sequencing HD alleles, paired-end might be useful to sequence CAG repeat along with flanking region and CCG repeats from the same direction. For longer alleles, sequencing CAG repeats can be done from one direction, and the CCG repeats from the other direction because we cannot match or merge the reads in the middle of the repeats.

In our study, we chose to use DM1 patient DNA samples as control samples to amplify and genotype HD normal alleles. These samples were selected as they are available in our Laboratory for research studies and also they have no known association with HD. Also, DNA samples collected from HD families from the region of Lake Maracaibo in Venezuela were sequenced to genotype HD chromosomes from both unaffected and affected individuals (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004). The US-Venezuela Collaborative Research project team collected buccal cell swabs from each individual in HD families in this area. The Venezuelans represent a large HD population and also have the highest prevalence in the world. These cohort members have been involved in over 20 years of studies including genetic,

cognitive and neurological studies. All the affected individuals of these kindreds are descendants of one woman who lived on Lake Maracaibo in the early 1800s. She had the disease, and her affected chromosomes have been passed through generations.

The aim of the chapter is to develop massively parallel sequence approaches to genotype HD CAG repeats. Genotyping by sequencing has the potential to detect insertions, deletions or other sequence variants within or flanking the repeat. Also, investigation of the use of NGS to detect the distribution and pattern of additional sequence variants. We describe an optimised library preparation for MiSeq sequencing to genotype HD alleles. In this study, we report the analysis of the normal allele in a large sample. The aims were to characterise the distribution of the CAG repeat alleles in Scottish and Venezuelan HD families and to test a potential association of the CCG repeat size with CAG length.

3.2 Results

3.2.1 Optimisation of MiSeq library preparation protocol

The sequencing platform of choice was the Illumina MiSeq platform with which a DNA molecule can be sequenced up to 300 bp in both directions or up to 600 bp in one direction, or in any other combination up to 600 bp in total. Using 600 bp run has the potential to allow the sequencing of relatively long expansions (>170 repeats)(Table 3-1).

	Maximum number of CAG repeat that could be sequenced using different runs		
Primer pair	**2x300 bp**	**400 bp/200 bp**	**600 bp**
31329/33934	80	113	180
HD319 F/33935.5	66	99	166
MS-1F/MS-1R	42	76	142

Table 3-1 The expected maximum number of CAG repeats that could be sequenced using the three different primer pairs and different type of MiSeq run. The three different primer pairs: 31329/33934, HD319/33935.5 and MS-1F/MS-1R short can be used for sequencing the CAG repeats, when aiming at sequencing 30 bp before the CAG repeat and 30 bp beyond the CAG repeat and also using a different type of MiSeq run. These sequencing runs are sequencing 300 bp in both directions, sequencing 600 bp from one direction and 400 bp from one direction and 200 bp from the other direction.

There are two sets of HD primers routinely in use in our laboratory (31329/33934) and (MS-1F/MS-1R) either of which could be used to design MiSeq primers for direct sequencing of the product including adapters and patient-specific barcodes. The first experiment was to test both pairs of locus-specific primers. PCR products of 143 bp and 417 bp were obtained from two different DNA templates using HD primers (31329/33934) and (MS1F/MS1R) respectively (Figure 3-2). This is consistent with the expected sizes of flanking region of 86 and 360 bps.

Primers (31329/33934) were chosen to design MiSeq primers as it has a shorter flanking region of 86 bp whereas (MS-1F/MS-1R) has a flanking region of 360 bp. (31329/33934) primer pairs were chosen for designing the sequencing primers.

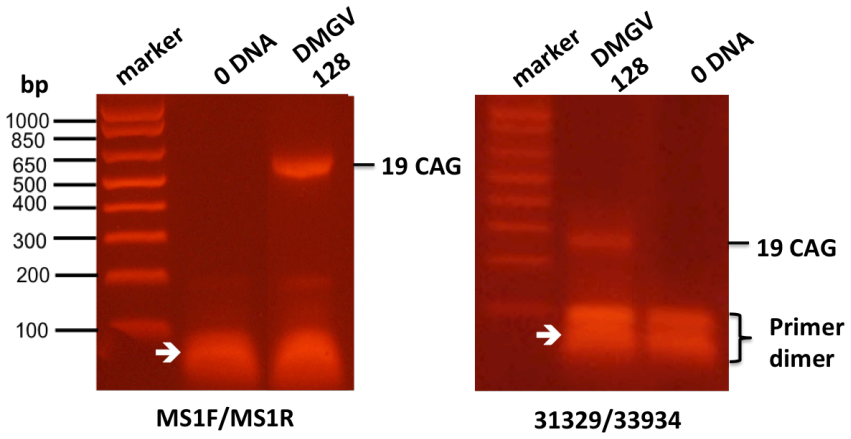

Figure 3-2 Agarose gel electrophoresis of the same DNA sample using two different combination MiSeq primers (31329/33934 and MS-1F /MS-1R short primers) for the target sequence. Lanes show PCR products obtained from sample (DMGV 128) using two different sets of MiSeq primers. The band size was obtained for the sample as expected, primer dimer is seen here at the bottom of the gel.

The principle for designing the MiSeq primer is to generate primers with additional Illumina adapters already incorporated at the 5′ end. These primers described in Figure 1-5 and Figure 3-3 were designed to include the P5 and P7 sequences, 8 base index sequence (index i5 and i7), and forward and reverse sequencing primer-binding site. Also, spacers of different length (0-9 bases) were added between the sequencing primer binding site and the locus-specific primer in each of the 8 forward primer sets.

The P5 and P7 sequences are oligos at the end of each library fragment were used to allow PCR products to anneal to their complementary oligos immobilised on the flow cell surface and to generate clusters. The forward and reverse primer binding sites are gene specific sequences used to amplify the HD fragment. The short spacers have been used to shift sequences in template DNA increasing base diversity across the length of the entire sequence reads, which are produced using the same pair of locus-specific primers and amplifying a highly repetitive region. Therefore, the sequencing for forward reads starts at the different spacers before each forward gene specific primer, which is critical for cluster identification in the first cycle of MiSeq sequencing. PCR was carried out using modified primers with a 5' extension with the adapter sequence. The resulting PCR products were then pooled and sequenced using the Illumina MiSeq platform.

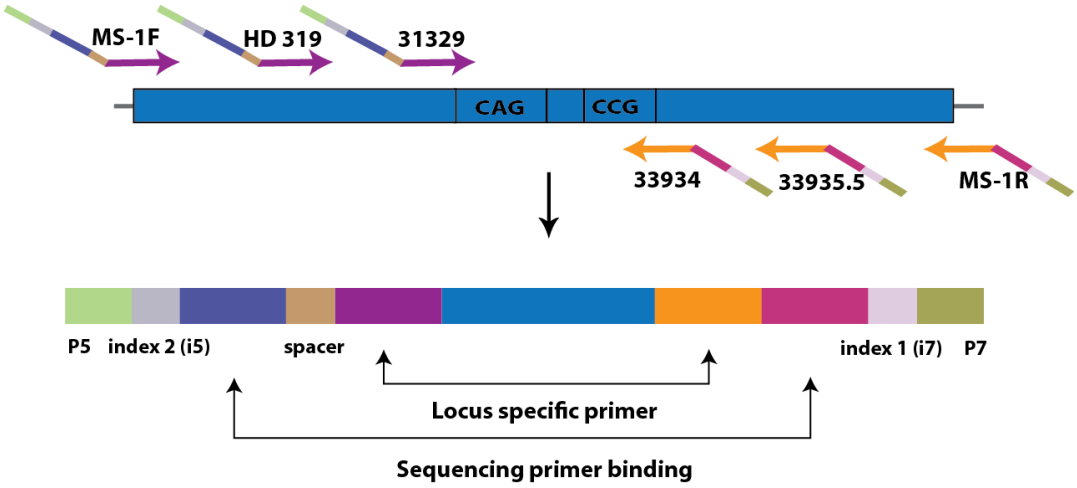

Figure 3-3 Primers designed for next generation sequencing using MiSeq platform. The location of PCR primers used for PCR amplification of *HTT* CAG repeat associated with HD relative to the DNA fragment that has CAG and CCG repeats. Three different locus-specific primers: (31329/33934), (HD319/3935.5) and (MS-1F/MS-1R short) combined with MiSeq adapters that are used to amplify *HTT* allele in different libraries and to allow the amplicon to be sequenced in the MiSeq paltform. DNA fragments that contain CAG and CCG repeats with P5 and P7 sequences, indices, sequencing primer binding site and locus-specific primers. These MiSeq primer components are described in Figure 1-5. The bottom of the figure shows the final amplicon ready to be sequenced on the MiSeq sequencing platform.

Different combinations of indices can be used using MiSeq primers. Dual indexed libraries were prepared with Illumina TruSeq adapters by adding unique indices to each sample. Index 1 (i7) is an 8 base index adjacent to P7 on the reverse primer. Index 2 (i5) is an 8 base index adjacent to P5 on the forward primer. These unique indices act as an ID tag to each sample so sequencing reads can be

associated with an individual. There are 96 unique combinations of indices generated using 8 forwards and 12 reverse indices that are uniquely dual indexed adapters. The 96 combinations can be used for this experiment.

To optimise Miseq library preparation, several experiments were performed, including selecting primers, sequencing run type and other slight modifications. The data analysis using CLC Genomics workbench was also optimised.

3.2.1.1 MiSeq sequencing for HD normal alleles of 14 individuals using a 2x300 bp run

A library was prepared for sequencing the HD normal allele using the MiSeq platform. The HD CAG and the adjacent CCG repeats were amplified from 14 DM1 patients using the locus-specific primers combined with MiSeq barcoded Illumina adapters that allow the PCR product to be sequenced using MiSeq. Different combinations of 14 indices were used in this experiment. Then, the PCR products were resolved by agarose gel electrophoresis. This step was performed to check the amplification using half of each PCR. The negative controls were checked for absence of contamination in the same way. The aim was to check that PCR amplification could be confirmed on the gel for the vast majority of the samples.

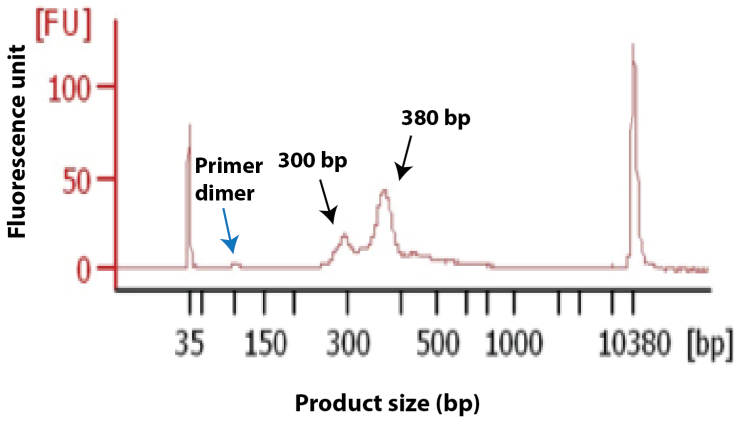

Figure 3-4 Bioanalyzer trace of the final library after pooling individual samples. The X-axis represents the product size in bp, and the Y-axis is the fluorescence intensity units (FU). The blue arrow points at a peak that potentially corresponds to primer dimers, black arrows point to the expected sizes for normal and expanded alleles in the library. The two peaks at 35 and 10,380 bp are the lower and upper markers that are run with each of the samples for estimating the size range and for analysing the data correctly.

The use of these primers was associated with the production of primer dimers during the PCR. Such primer dimers made it necessary to purify the library from a gel. Agarose gel purification is efficient to recover primer dimer-free PCR products, and it is inexpensive and relatively simple. However, this method is a low throughput and multistep procedure. PCR was followed by gel purification using a gel extraction kit supplied by Qiagen.

Then, the DNA concentration was measured using a Qubit fluorimeter using the Qubit dsDNA HS Assay Kit (Life Technologies) and a Bioanalyzer (Aligent). The Bioanalyzer analysis was used to evaluate the quality and quantity of the sequencing libraries, to check the fragment had the expected size, the primer dimers were absent and to estimate the molarity of the library. An example of a bionalyzer result is shown in Figure 3-4 for one of the prepared libraries. After that, the library was prepared by pooling the different samples at equimolar concentration.

Those sequencing libraries were sent to Glasgow Polyomics Facility for sequencing using MiSeq platform with the 600 cycles of sequencing being performed by 2x300 bp in both forward and reverse directions. Libraries were sometimes sequenced with other samples from our group. CLC genomic workbench software was used to analyze sequencing data for the *HTT* locus allele. This software allows the user to analyze, visualize, and compare the data from all major high throughput-sequencing platforms such as Illumina.

Sequencing data were received for all samples. From this data, we aimed to determine the genotype for each sample. Genotyping for these samples and the rest of the libraries are explained in more details in section 3.2.2.4. We established that it was possible to sequence and genotype normal *HTT* CAG alleles and the adjacent CCG repeats using the MiSeq platform by amplifying the region using the internal (31329/33934) locus-specific primers combined with MiSeq sequencing adapters (Figure 3-3). Data from similar studies in DM1 have shown that the region immediately flanking the repeat is a hotspot for sequence variants (Braida *et al.*, 2010). Thus, using PCR pair of primer 31329 and 33934, which are very close to the CAG and CCG repeats, might not be the best approach, as these primers may not detect any additional sequence variants in this region.

To address these issues, we considered the use of additional primer pairs flanking the HD CAG repeats. Primers (MS-1F and MS-1Rshort) were used instead (Figure 3-3), which are located further apart from the repeats and yielded good amplification of the CAG repeats. There are not many primers we can use for amplifying the CAG repeats because of the extreme GC content in the flanking region of the repeats.

3.2.1.2 MiSeq sequencing for HD normal alleles of 19 individuals using 2x300 bp and 600 bp run

Another library was prepared for sequencing the HD normal allele. The HD normal alleles were amplified from 19 DM1 patients and were allocated different combinations of indices using MS1F/MS1R short MiSeq primers. By using (MS1F and MS1Rshort) primers, it was possible to purify the sequencing library without the use of gel purification. This can be done using the paramagnetic beads (Agencourt AMPure XP reagent, supplied by Beckman coulter) based approach to clean up the PCR, as there is a greater size difference between the full-length amplification products and primer dimers. The PCR Products were cleaned up using a ratio of AMPure beads allowing size selection to remove the primer dimer using (0.6 µl of the bead solution for each µl PCR). The second round of AMPure clean up was necessary to concentrate the library (using 1.4 µl of bead solution for each µl of sequencing library).

Single (1x600 bp) and also paired-end reads (2x300 bp) were used to compare between both runs in terms of sequencing quality and a number of reads obtained for each run (Figure 3-5). Dual indexed libraries were prepared with adapters containing two indices for both libraries (as the previous library).

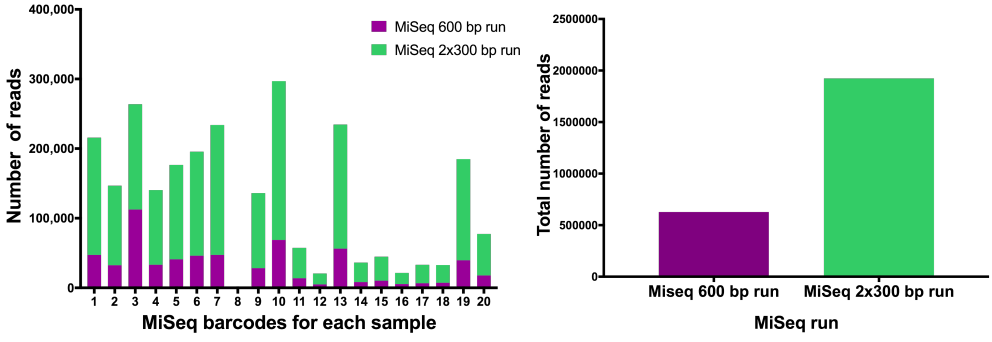

Figure 3-5 Number of reads obtained for each sample in two different MiSeq runs. Total number of reads obtained for MiSeq 2x300 bp and 600 bp run.

The accuracy of DNA sequencing is assessed by a Phred score, which represents base call accuracy (Ewing and Green, 1998; Ewing *et al.*, 1998). The Phred score provides the probability that the base called is correct. For each base call, parameters of the cluster, including signal-to-noise ratio and intensity profile are measured. These are then compared against a quality table of standard data obtained by sequencing a range of well-characterised loci and alignment against reference genomes. The Phred score is presented on a logarithmic scale that provides a quality score for each base. The assigned quality scores are a measure of the accuracy of the base call of the sequencing platform. This results in a graph showing the base position and quality score. These scores ranged from 0 to 63 with higher scores corresponding to higher quality. Phred scores of 20 and above are considered high quality bases. The sequencing quality score of 20 represent an error of 1 in 100, which means there is a 99% chance the base is called correctly. High and low quality reads can be identified by looking at individual sequences. It is normal to see the sequencing quality dropping off near the end of reads, as many errors are more likely to occur at the 3' end of the reads. Such low quality ends can be trimmed off using the trim sequences tool. The quality assessment for the 2x300 bp run, more than 50% of the reads have a Phred score of more than 30. This quality score of 30 represents an error of 1 in 1,000, indicating a smaller probability of error. However, for the 600 bp run, by 400 bp the Phred score goes down to 6 representing an error of 1 in 10. That means there is only a 90% chance the base is called correctly. This low score can result in a significant proportion of the reads not being useable and may also lead to increased false positive variant calls leading to inaccurate genotyping. We found that the sequencing quality drops down significantly after 400 bp(Figure 3-6), therefore, the 2x300 bp or 400x200 bp run was used to sequence HD alleles in order to sequence the CAG repeats from the forward strand and the CCG repeats from the reverse strand in longer CAG repeats (> 113).

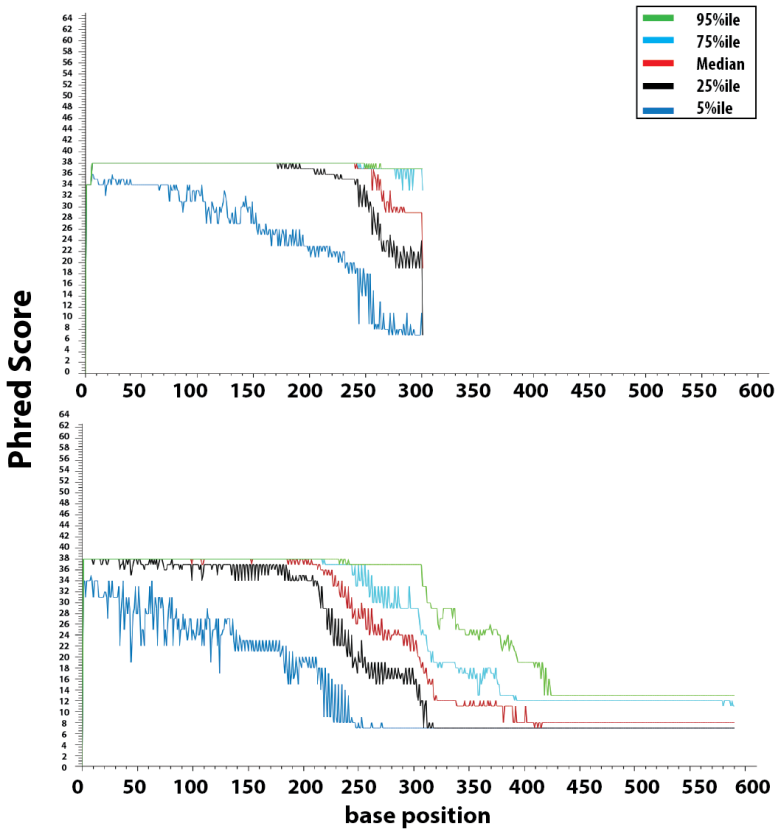

Figure 3-6 Quality assessment for each run showing the quality values per base position for 2x300 bp and 600 bp runs. The top graph shows the quality distribution for 2x300 bp run and the bottom graph shows the quality distribution the for the 600 bp run. (The graphs obtained from CLC workbench software for creating quality control report).

3.2.1.3 MiSeq sequencing for HD normal alleles of 96 samples using 2x300 bp run

Sequencing the HD normal allele in 96 unaffected individuals was carried out in this experiment. The PCRs were manually prepared in 96 well plates (8 forward barcodes x 12 reverse barcodes = 96 barcode pairs) using 96 combinations of Illumina TruSeq barcode set from Illumina as indicated in Figure 3-8. Therefore, 96 different combinations of indices were used in preparing the library using MS1F/MS1R short primers. Paired-end reads (2x300 bp) were used in this experiment.

After PCR amplification was checked by gel electrophoresis, the vast majority of the PCR amplifications were successful. The remaining half of the 96 PCRs were then pooled (2 pools of 48 samples). Each pool was then purified with AMPure XP beads (using 0.6x ratio of beads / µl of PCR product) allowing the size selection to remove of the primer dimers. Then, the two purified pools were pooled and

another round of AMPure XP beads (1.4 µl of beads / µl of PCR product) to concentrate the library further. After that, the DNA concentration was measured using a Qubit fluorimeter using the Qubit dsDNA HS Assay Kit and was then evaluated by capillary electrophoresis on a Bioanalyzer (Aligent) to check the fragment had the expected size.

3.2.1.4 Further optimisation of the library preparation protocol

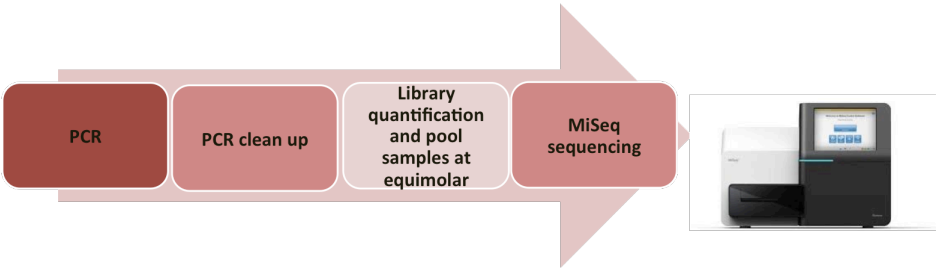

Figure 3-7 Strategy to sequence the HD normal allele. A library was prepared by pooling different samples at equimolar concentration. The completed library was sent to Glasgow Polyomics for sequencing on MiSeq platform.

After a few sequence runs, our laboratory protocol was modified to do PCR for 90 test DNA samples and 6 controls in one 96 PCR well plate without the need to use agarose gel electrophoresis to confirm amplification. The overall sequencing strategy for sequencing *HTT* alleles is illustrated in Figure 3-7. The figure illustrates our current method in the lab. The PCRs were manually prepared in 96 well plates. This library included 6 controls instead of 96 for each primer combination because sequencing no template DNA controls are more sensitive than gel electrophoresis (Figure 3-8). Previously we discovered that some samples did not have visible bands on an agarose gel, but they still gave sufficient amount of reads to allow genotyping by MiSeq sequencing. Therefore, checking the PCR products by gel electrophoresis is not necessary. After that, we evaluated the quality and quantity of the sequencing libraries, by Bioanalyzer analysis to check the library had the expected size and that the primer dimers were absent.

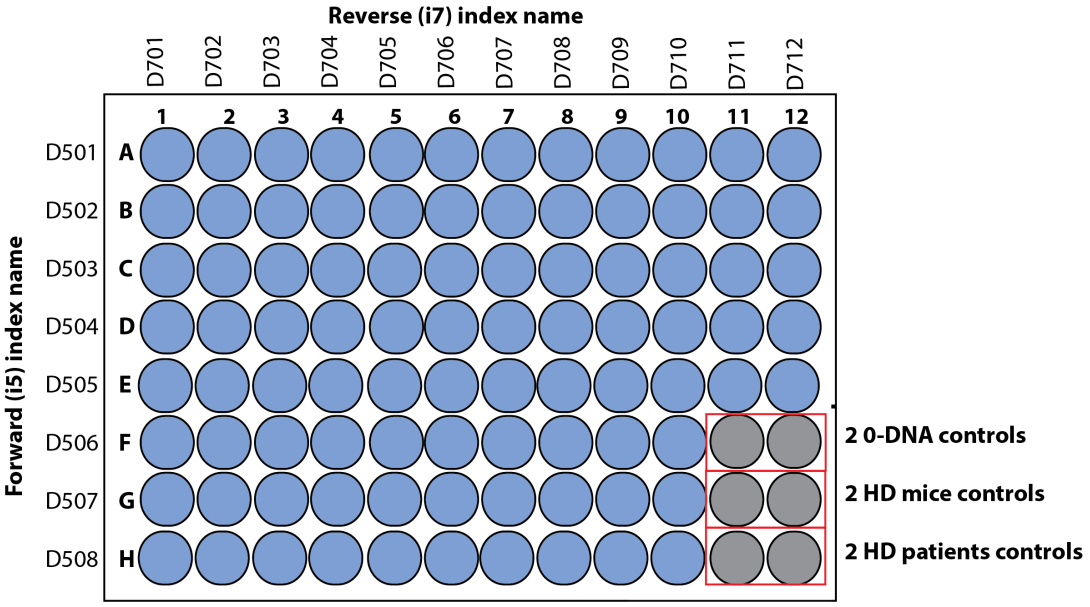

Figure 3-8 96 PCR well plate includes 90 test DNA and 6 controls including 0 DNA (no DNA template), 2 positive HD patients and HD mice controls. Blue wells represent DNA samples with a different combination of indices including forward (i7) index and reverse (i5) index. Grey wells represent the control samples used with each plate. These controls can be placed in any wells with a different combination of indices.

Although we obtained genotypes of all samples using MS1F/MS1R short, long CAG repeats (> 76) cannot be identified using these outer primers because those primer pairs lie quite a distance from the repeat. Therefore, Dr. Marc Ciosi in our group managed to optimise another primer pair (HD319 F/33935.5) for MiSeq sequencing (Figure 3-3). The new PCR primers were designed to minimize the length of the flanking region as well as to avoid the region immediately flanking the repeat, which could be polymorphic. In addition, reduced flanking region would potentially allow the sequencing of longer CAG repeats (Table 3-1). The intermediate primer pair (HD319 F/33935.5) was selected, as it was associated with good amplification of the CAG repeats and quality sequencing data. Also, it is still possible to purify the sequencing library using AMPure beads based approach to clean up the PCR rather than gel purification.

We have developed an optimised assay for the detection of CAG repeat length in the *HTT* gene. After that, further libraries were prepared for genotyping the normal alleles of the unaffected and affected individuals from the Venezuelan cohort.

3.2.2 Analysis of the NGS data using CLC Genomics Workbench software

To determine the genotype for samples, CLC Genomics Workbench Version 7.5 was used to visualize and analyse NGS sequencing data. CLC Genomics Workbench provides a user-friendly interface and is capable of performing multiple analyses. This software has a major advantage in flexibility in the parameters option for read assembly. CLC is relatively expensive, and its licence is no longer available for use by University of Glasgow staff and students. Therefore, our group has moved to using Galaxy software that has been widely used for NGS data analysis, and it is free software.

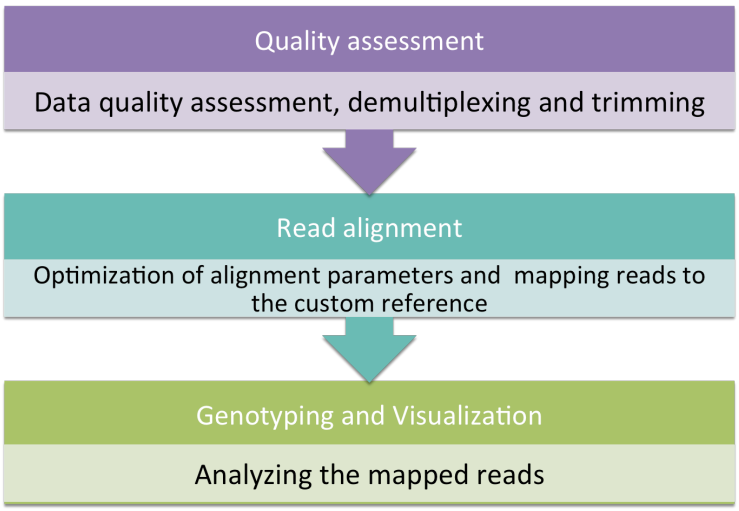

Figure 3-9 Workflow for analysis of MiSeq sequencing data using CLC genomic workbench. After library preparation, samples are sequenced on MiSeq platform. The next steps are quality assessment, read alignment, followed by genotyping the samples and visualisation of the data.

The analysis of our data using CLC genomic workbench was composed of: quality assessment of the raw data, alignment parameter optimisation, read alignment to the reference and genotype identification and visualization of the data (Figure 3-9). In the following sections, each of these steps will be explained briefly.

3.2.2.1 Quality assessment

After completing the sequencing run, it is important to evaluate the quality of raw reads and to remove or trim reads that might have multiple errors and low

quality reads. The raw data generated by MiSeq sequencing might include base calling errors, insertions/deletions, poor quality reads and adapter contamination. Therefore, it is necessary to perform a base quality assessment, read filtering and trimming of reads before performing any downstream analysis for enhancing mapping accuracy.

To assess the sequencing data quality, CLC provides various tools including visualization of sequence read lengths, base coverage, nucleotide contribution, GC-content distribution, ambiguous base content and quality distribution. Creating a quality control (QC) report provides quality control checks on raw sequence data that gives a clue of whether your data has any problems before doing any further analysis. The quality analysis examines quality scores of bases and sequences that feature individual Phred score in 64 bins from 0 to 63 of its base qualities as shown in (Figure 3-6). Running QC report yields summary graphs and tables to assess the data quality.

For some MiSeq libraries, indices were shared with other users sequencing different loci. In this case, demultiplexing of the sequencing data was required to separate the reads into their corresponding samples before further analysis. The sample-specific tag, also called the barcode can be used to distinguish between the different samples when using the demultiplexing tool in CLC. In our data, the sample specific spacer and locus-specific forward primer can be used to distinguish between different loci when analysing the sequencing data. The forward reads for the shared index libraries were demultiplexed using the spacer plus 10 bases of the locus-specific forward primer as a barcode. Demultiplexing was carried out on the sequencing reads if the sequencing reads contain an exact match to any of the barcode sequences provided. The demultiplexed reads will be generated as output for each specific barcode and also the demultiplex reads report showing the number of reads with and without a barcode.

The trimming tool offers a number of ways to trim sequence reads prior to mapping, including adapter trimming, quality trimming and filtering on length. If the insert size is below the specified sequencing length (if the insert size is 200 bp, and the MiSeq is sequencing 300 bp), the adapter will be sequenced. In this case, the sequences of corresponding library adapter can be present in the output files at the 3’ end of the reads. To remove these sequences and prevent

issues with alignment, adapter trimming can be performed with CLC software by selecting the sequence of the reverse sequencing primer-binding site as an adapter sequence to be trimmed. When an adapter sequence is specified for trimming, the adapter and sequence of the 5' end of the adapter will be removed. Trimming these sequences would remove the insert because it is at the 5' end of the adapter sequence. Therefore, we must remove the reverse compliment of the adapter sequence that is found in the minus strand. The adapter sequence that is used as recognition site and 3' of the adapter is removed from the raw reads data in relation to the positive strand. The trimming parameters were used in our data: Mismatch cost = 1; gap cost = 2; minimum score for internal matches = 6; minimum score for end matches = 3. These trimming parameters were optimised by my colleague Dr. Marc Ciosi. After trimming, a list of sequences that have passed the trim will be produced for each input sequence list when the sequences are selected. Also, the trimmed report shows a number of nucleotides after trimming and read length distribution before and after trimming (Figure 3-10).

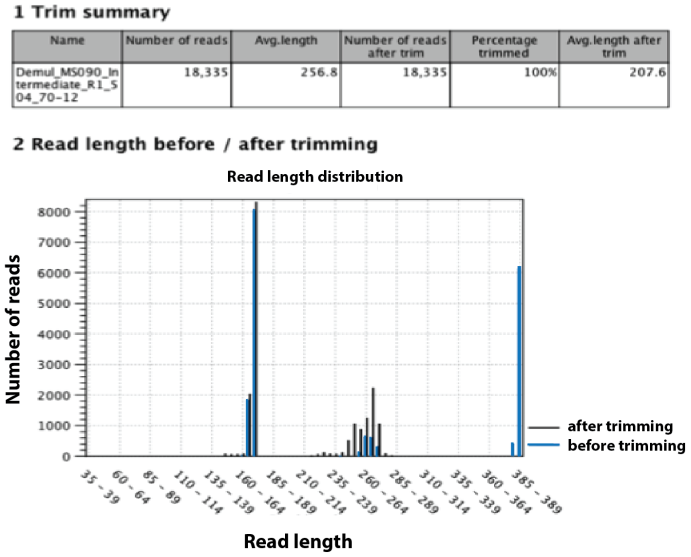

Figure 3-10 Trim summary with statistics and graph are generated from trimming the specified adapter sequences for MiSeq sequencing data by CLC workbench software. The trim summary shows the number and the average length of the reads in the input file, the number and percentage of reads after trimming and the average length of trimmed reads. The graph shows the number of read length before and after trimming to see how the trimming has affected the read lengths.

3.2.2.2 Mapping

After reads were processed, the reads were mapped to custom references including the region flanking the repeats and various numbers of CAG and CCG repeats. Four thousand reference sequences were considered (1 to 200 CAG repeats and 1 to 20 CCG repeats) in order to obtain the genotype. This choice of CAG and CCG ranges were included in the reference to represent most possible repeat sizes found in the normal and expanded alleles in unaffected and affected individuals. As most HD patients have CAGs from 40 to 50 repeats and CCG repeats varying from 4 to 12 repeats. This customised reference will yield greater alignment by allowing the alignment software to map reads to the closet match in terms of CAG and CCG length for each sample. To start read mapping, the sequences containing the sequencing data and the reference sequences should be selected. Single sequences or a list of sequences can be used as a reference. The software can generate a summary report about the mapping process.

The first library was prepared by amplifying the repeats by inner primers (3139/335), so the reference was designed to have a flanking region of the pair of primers. Then, we used the outer primers (MS1F/MS1R short) to amplify *HTT* allele in few libraries. After that, another primer pair HD319 F/33935.5 (intermediate primers) was used for library preparation for MiSeq sequencing. The new PCR primers were used to minimize the length of the flanking region from the outer primers as well as to avoid the region immediately flanking the repeat using the inner primers. Therefore, the rest of libraries were prepared using intermediate primers and the custom references were designed to have the flanking region of these primers.

For each allele, the CAG/CCG genotype was defined to be the genotype of the reference against which the highest number of reads had aligned. Both CAG and CCG repeat sizes were determined for all individuals. By alignment of the obtained reads to the reference flanking regions, the start and end of both repeat regions can be determined, which allows the exact length and sequence of the region to be obtained. Mapping parameters will be discussed in detail in the next section.

3.2.2.3 Alignment parameter optimisation

The alignment parameters in CLC allow mismatch and gap costs to be adjusted. The mismatch cost, insertion cost and deletion cost parameters refer to the score for a match between the sequencing reads and reference for the mismatch, insertion and deletion. Adjusting these parameters can improve the alignment quality.

Mismatch, insertion and deletion cost ranged from 1 to 3. We selected a mismatch cost of 2 for our data because we expect the PCR errors and possible atypical alleles to generate mismatches in the sequencing reads. Therefore, we lowered the mismatch cost to allow for such errors in reads, as the reference is sometimes expected to differ from the sequenced reads. Increasing parameter scores for insertion and deletion cost to 3 results in the reads mapping to a reference with an exact match without insertions and deletions. However, it is important to note that the custom reference includes a variable number of CAG and CCG repeats to allow the obtained reads to map to the closest length of repeats.

To discover if these selected parameters are optimal for alignment of the obtained sequence reads, a single sample was selected to compare and find the optimal parameters for the reads based on the cost parameters. The optimal alignment parameters of the reads were determined. The mapping process involved the following mapping parameter values: mismatch cost =2, insertion cost =3, deletion cost = 3, length and similarity fractions = 0.9. Those were identified based on the accuracy and the highest proportion of reads that were aligned to the reference.

After the optimal alignment of the read is found, a filtering process determines whether this match between the reads and reference is suitable for the read to be included in the alignment output. The filtering process is determined by length fraction and similarity fraction. Length fraction refers to how much of the read must match the reference sequence to the level of similarity for this read to be mapped. The selected threshold was 0.9; that means at least 90% of the read needs to align to the reference. Similarity fraction specifies how similar the sequence reads must be for the reference to be mapped. We used 0.9 for

similarity fraction and 0.9 for length fraction; it means that at least 90% of the read should align with 90% similarity in order to be mapped. The mapping process involved the following mapping parameter values: mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.9 and similarity fractions = 0.9.

3.2.2.4 Genotyping identification and visualisation

Validation and visualisation of the generated results are important steps in NGS data analysis. Visual representation of obtained data is very useful for the interpretation of the result. CLC workbench supports the NGS visualization tool by displaying aligned reads and identified any mismatches. We also used another graphical viewer for NGS alignment, Tablet, to visualize the read mappings (Milne *et al.*, 2013). Tablet displays read coverage, reads names and it allows searching for specific coordinates across the dataset.

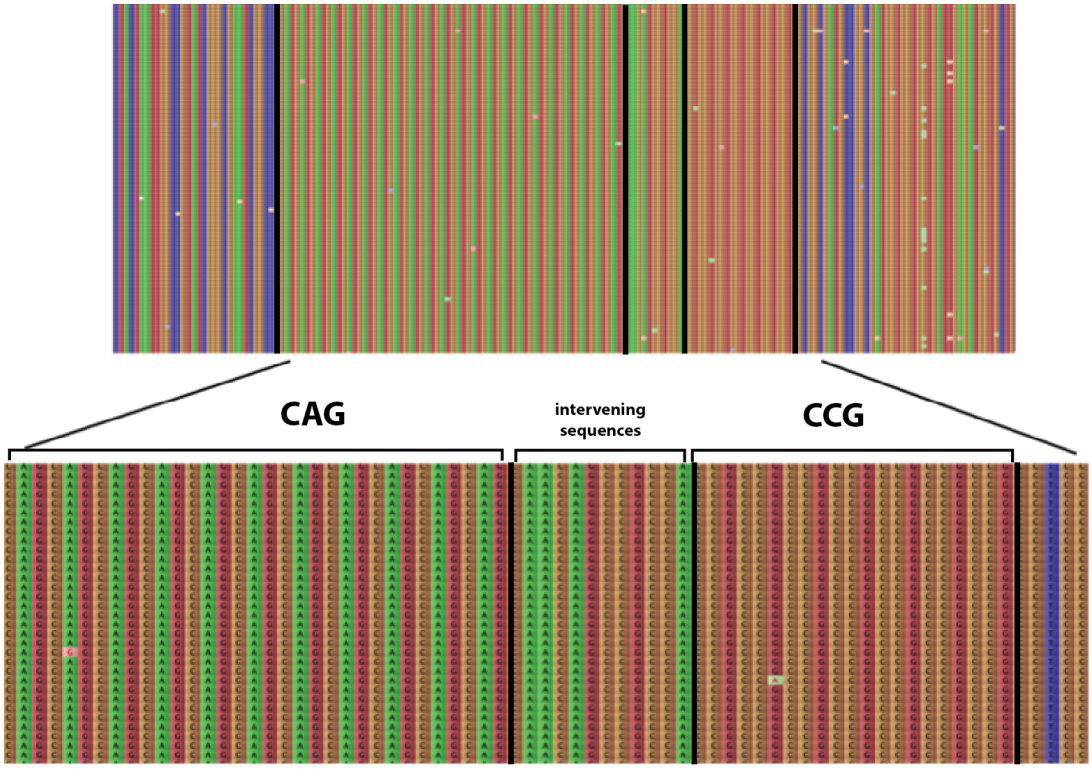

Figure 3-11 Sample reads were aligned against a set of reference sequences with $(CAG)_{1\text{-}200}$ and $(CCG)_{1\text{-}20}$ using CLC genomics workbench software. The target CAG repeats are seen and also CCG and the intervening sequences. The intervening sequences are CAACAGCCGCCA. This is the alignment for an allele that has 11 CAG repeats and 7 CCG repeats. Visual representation of obtained read mapping was performed using Tablet.

Analysing the mapped reads confirms the presence of both flanks to the repeats. An example of mapped reads with 11 CAG repeats and 7 CCGs with flanking sequences, is shown in Figure 3-11. Normal allele length distributions are presented in Figure 3-12 for one individual. That individual is heterozygous for 11, and 17 CAG repeats normal alleles as well as heterozygous for 7 and 10 CCG repeats. One of the HD alleles has 17 CCG repeats with 7 CCG repeats, and other allele has 11 CAG repeats and 10 CCG repeats. The mapped reads of the normal allele with 11 CAG repeats showed putative PCR slippage where multiple reads were mapped against references of different lengths, resulting in multiple peaks (Figure 3-12). Putative slippage results in reads mapping against 9, 10 and 12 repeats. Also, the alignment for the 17 repeat allele shows PCR slippage that results in reads mapping against 13, 14, 15, 16 and 18 repeats (Figure 3-12). This is also seen for CCG slippage in which multiple reads were mapped against references of different lengths. PCR slippage clearly occurs and becomes more noticeable as the allele gets bigger.

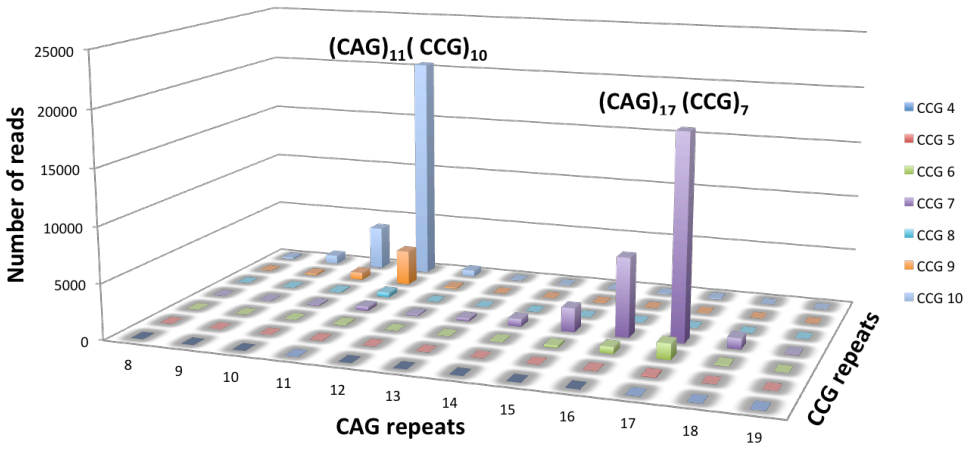

Figure 3-12 Normal allele length distribution of 11 and 17 CAG repeats against a number of reads for one individual. The reads were aligned against the references. Most of the reads that aligned against the 10 CCG reference mapped against 11 CAG repeats, represented as the heighest blue peak. Also, a large number of the reads that aligned against the 7 CCG references aligned against 17 repeats, represented as the highest purple peak.

3.2.3 Analysis of CAG and CCG repeats in normal alleles

The analysis of the *HTT* alleles was carried out on 210 blood DNA samples from unaffected Scottish individuals and 742 buccal swab DNAs from the US-Venezuelan Collaborative Research Project consisting of unaffected and affected individuals from HD families. Our samples from Venezuela compromised 131

families. The family size ranged from 1 to 20 individuals with an average of 3.6. Of the 210 Scottish samples, 96 samples were prepared by my colleague Dr. Marc Ciosi working on a CHDI funded project. Also, he prepared one of the libraries that included 91 samples to amplify HD alleles from affected individuals from the Venezuelan cohort, and I carried out the data analysis for those libraries.

3.2.3.1 Analysis of CAG repeats in normal alleles of unaffected Scottish individuals

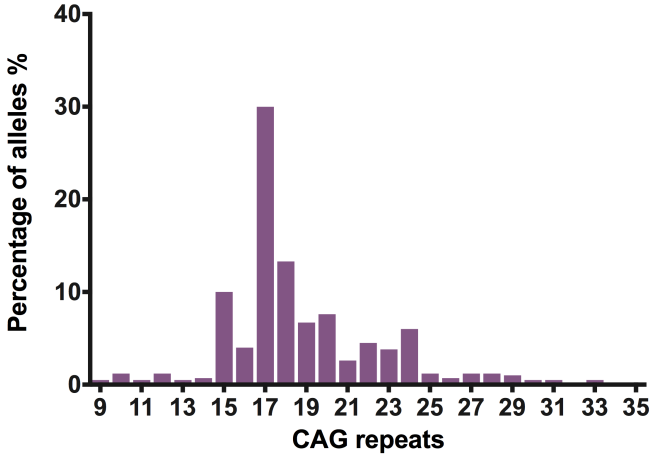

Figure 3-13 Distribution of CAG repeats of HD alleles from 210 individuals from the Scottish population. The distribution is shown for the number of CAG repeats observed on 420 chromosomes from the unaffected Scottish population.

Among the Scottish population with no association to HD, CAG sizes were determined for 420 chromosomes. CAG sizes are considered the length of pure CAG repeats in the *HTT* gene. The CAG size of the chromosomes ranged between 9 and 33 repeats, with the most common size being 17 CAG. The mean CAG size of normal alleles was 24. The CAG size distribution of these general population chromosomes is illustrated in Figure 3-13. In this study, 21 intermediate alleles (27 to 35 CAG repeats) were identified representing an allele frequency of 5% in the Scottish population. The mean CAG size of intermediate alleles was 29.1, and the median was 29 CAG. No reduced penetrance HD alleles (36 to 39 CAG) were identified in this population (Table 3-2).

3.2.3.2 Analysis of CAG repeats in normal alleles of unaffected and affected individuals from the Venezuelan cohort

For this study, Venezuelan kindred members can be separated into four groups according to the length of CAG repeats in their *HTT* alleles: unaffected individuals who have normal sized alleles on both chromosomes, individuals with intermediate alleles who inherited CAG between 27 to 35 repeats, individuals who have low penetrance alleles (36 to 39 CAG repeats) and affected individuals who have full penetrance alleles (40 CAGs or more).

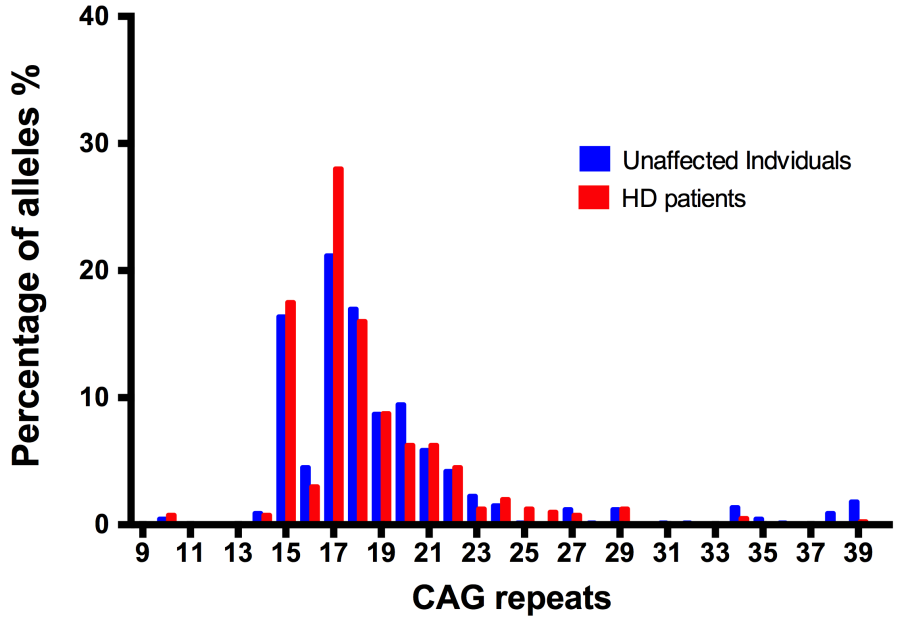

Figure 3-14 Distribution of CAG repeats in unaffected individuals and affected of HD families from the Venezuelan cohort. The CAG allele frequencies of normal alleles were estimated from 333 unaffected individuals (666 alleles) and from 400 HD patients (400 alleles).

We assessed CAG repeat length in the 742 unaffected and affected individuals of HD families of the Venezuelan cohort. The distribution of CAG repeats sizes on the non-disease associated alleles are presented for 333 unaffected individuals with two normal chromosomes (Figure 3-14). This gave a total of 666 normal alleles in unaffected individuals of HD families.

		Normal allele			Intermediate allele			Reduced penetrance allele		
Population	Individuals (alleles)	No.	Mean	Median	No.	Mean	Median	No.	Mean	Median
Unaffected indivduals from Scottish population	210 (420)	399	18.2	17	21	29.1	29	0		
Unaffected indivduals from Venezuelan population	333 (666)	616	17.96	18	31	30.6	29	19	38.5	39
HD patients from Venezuelan population	400 (400)	389	17.97	17	10	29.4	29	1	39	39
Total	943 (1486)	1404			62			20		

Table 3-2 CAG allele frequency in the normal allele, intermediate allele and reduced penetrance allele ranges in unaffected individuals from Scottish population, unaffected and affected individuals of HD families from the Venezuelan cohort. The number of alleles, mean and median were estimated for each CAG allele range for all populations. No.= number of alleles.

The range of the CAG number on the normal chromosomes was 10 to 39 repeats in unaffected individuals. The 17 CAG repeats were the most common length for the normal allele. Out of a total of 333 individuals (666 alleles), 31 individuals had CAG repeat lengths in the intermediate range with a mean length of 30.6 CAG and median of 29 CAG (Table 3-2). Approximately 4.65% have intermediate alleles. Nineteen (2.85%) individuals have alleles of reduced penetrance with a median of 39 (Table 3-2).

We determined the size of CAG normal alleles of 409 affected individuals with HD. In 400 patients, one normal and one expanded allele were found, while 9 subjects were homozygous for two expanded alleles. The expanded alleles genotyping and distribution will be discussed in detail in Chapter 4. In the 400 heterozygous subjects, who carried one expanded allele, the normal alleles contained 10 to 26 CAG repeats with a mean of 17.97 and a median of 17. The 17 CAG repeats were the most common allele in the shorter allele of HD patients, which is similar to the unaffected individuals' result. Ten (2.5%) alleles were found in the intermediate allele range. One allele was found in the reduced penetrance range (0.25%). The distribution of the normal alleles in the HD patients was similar to the distribution in the normal chromosomes analysed in the unaffected individuals. However, there are a few differences in which 17 CAG frequency was slightly higher in affected than in unaffected individuals, and

intermediate alleles occur at a lower frequency in affected individuals than in unaffected individuals.

3.2.3.3 Analysis of CCG repeats in normal alleles of unaffected Scottish individuals

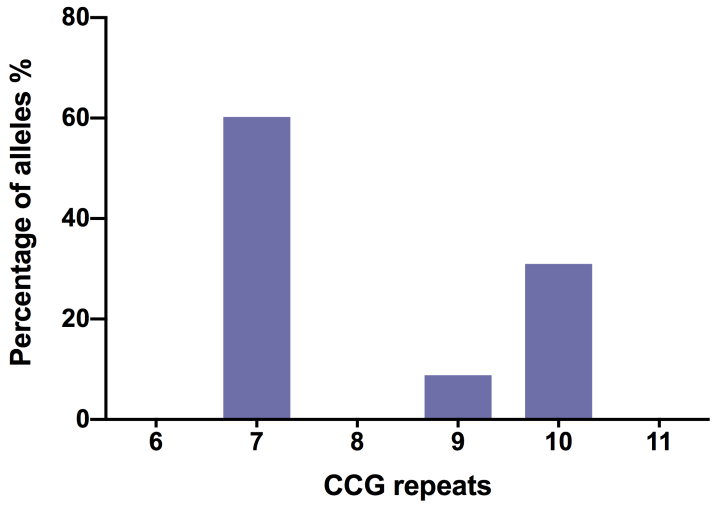

Figure 3-15 The distribution is shown for the number of CCG repeats observed on the *HTT* normal chromosomes from 120 unaffected Scottish population.

The CCG repeat alleles in 121 Scottish individuals were assessed. Three different CCG repeats were identified 7 (60.23%), 9 (8.8%) and 10 (30.95%) repeats (Figure 3-15).

3.2.3.4 Analysis of CCG repeats in normal alleles of unaffected and affected individuals from the Venezuelan cohort

In 333 unaffected individuals, we identified five CCG repeat allele lengths, repeat numbers 6 (0.45%), 7 (55.7%), 9 (15.6%), 10 (28%) and 11 (0.15%). We also identified five CCG alleles in the 409 HD patients. HD patients inherited 6 (0.97%), 7 (57.9%), 9 (13.9%), 10 (25.6%) and 11 (1.46%) in their normal allele.

The 7 repeat allele was the most frequent in both unaffected individuals and HD patients. The 10 repeat was the next most frequent. The result shows that five HD patients and one unaffected individual were associated with 11 CCG repeats. Out of five HD patients who inherited 11 CCG repeats, three of them were siblings. We also identified 6 CCG repeats in four HD patients and three unaffected individuals. Among individuals with 6 CCG repeats, two patients were

siblings, and two other HD patients and one unaffected individual were siblings as well.

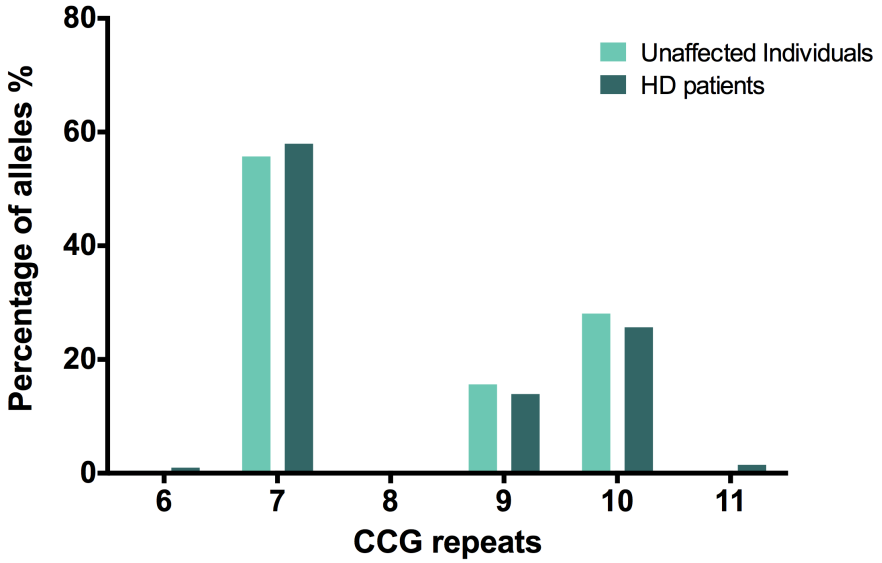

Figure 3-16 The distribution is shown for the number of CCG repeats observed on 1,066 *HTT* normal chromosomes from 333 unaffected and 400 affected individuals from the Venezuelan population.

3.2.4 Haplotype analysis of normal chromosomes

We have constructed haplotypes based on the length of both CAG and CCG repeats for all individuals who were either homozygous or heterozygous at the CAG or CCG locus. We determined the haplotypes for normal alleles of the Scottish and Venezuelan population from the NGS data for each individual.

3.2.4.1 Haplotype analysis of normal chromosomes in the Scottish population

In the unaffected Scottish population, we determined the haplotype for those individuals for the CAG-CCG locus. This method helped us to determine 40 haplotypes in 420 normal chromosomes using NGS sequencing. Haplotype CAG 17-CCG 7 has the maximum frequency (18.8%), and CAG 17-CCG 10 has the second highest frequency (9.8%) among all haplotypes (Figure 3-17).

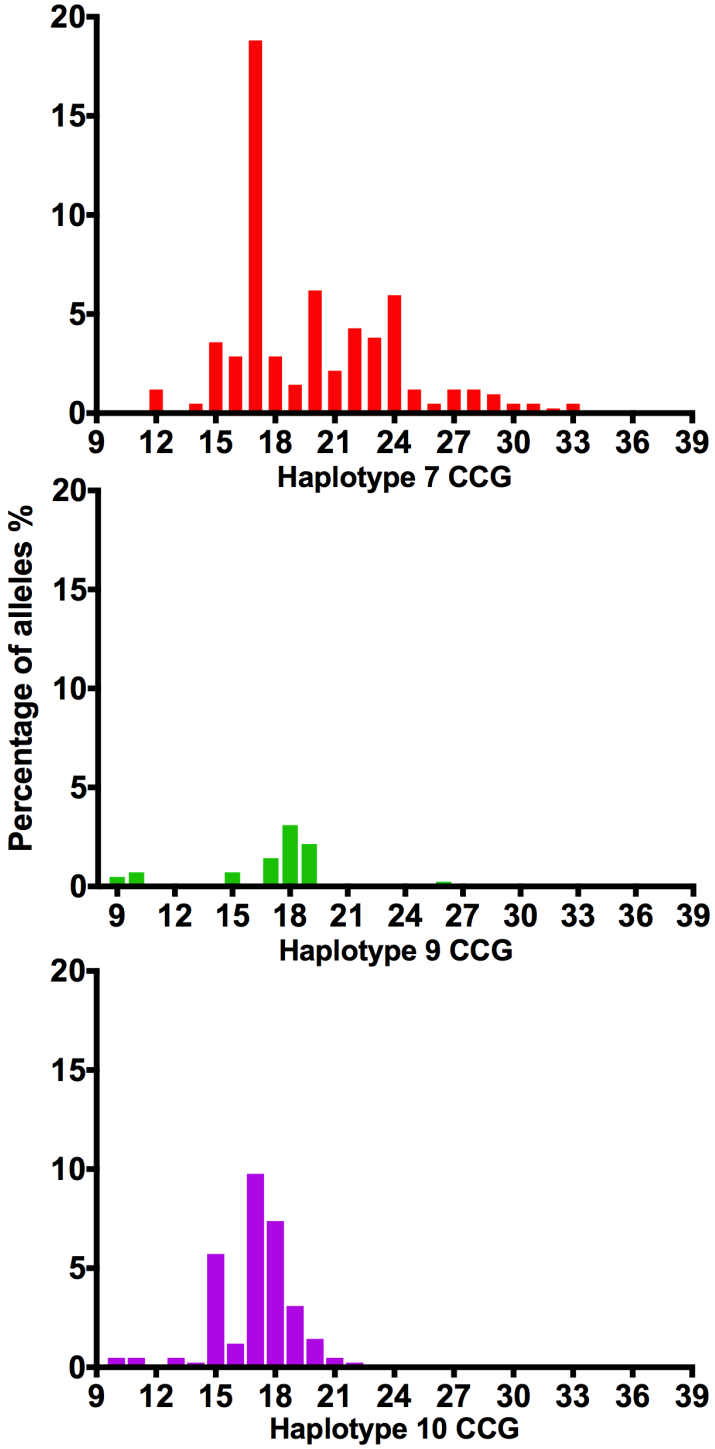

Figure 3-17 Haplotype analysis of *HTT* in 210 unaffected Scottish population. The graphs show the distribution of CAG sizes on normal chromosomes for individuals carrying 7 CCG, 9 CCG and 10 CCG repeats.

With haplotype 7 CCG, CAG 17 has the highest frequency (18.8%) followed by 20 CAG (6.2%). CAG repeats from 12 to 33 were represented in this haplotype. Also, 17 CAG (9.8%) was the most common in haplotype 10. The second most frequent allele in haplotype 10 was 18 CAG repeats (7.4%), which is different than the 7 CCG haplotype. CAG repeats from 10 to 22 were represented in the haplotype

10. Haplotype 9 CCG was associated with a lower number of samples. The 18 CAG (3.1%) was the most common in haplotype 9. Most CAGs ranged from 9 to 19 with haplotype 9 CCG, though one allele had 26 CAG repeats.

We calculated the percentage of *HTT* chromosomes, mean CAG size and ranges of CAG repeat length associated with different CCG repeats (Table 3-3). The mean CAG size and range of CAG repeats were different with 7 and 10 CGG. HD chromosomes with 7 CCG were associated with a larger range of CAG repeat lengths and higher mean CAG length than 9 and 10 CCG alleles. The mean CAG repeats sizes on chromosomes with 7, 9 and 10 CCG repeats were 19 ± 0.26, 16.92 ± 0.56 and 16.98 ± 0.17 respectively. The mean CAG size was similar for 9 and 10 CCG repeats.

	Number of chromosomes (%)		Mean CAG repeat size		Range of CAG repeat length	
CCG	**Scottish**	**Venezuelan**	**Scottish**	**Venezuelan**	**Scottish**	**Venezuelan**
6	0	7 (0.66%)	0	15.14 ± 0.14	0	15 –16
7	253 (60.23%)	599 (56.19%)	19.92 ± 0.26	19.78 ± 0.21	12 – 33	14 –39
9	37 (8.81%)	161 (15.1%)	16.92 ± 0.56	17.64 ± 0.13	9 – 26	15 –26
10	130 (30.95%)	292 (27.39%)	16.98 ± 0.17	17.74 ± 0.15	10 – 22	10 – 27
11	0	7 (0.66%)	0	13.86 ± 1.03	0	10 –17
Total	420	1066			9 – 33	10 – 39

Table 3-3 CCG repeat frequency in normal chromosomes among 210 unaffected Scottish individuals and 333 unaffected and 400 affected of HD families from the Venezuelan population. The mean CAG repeat size and range of CAG repeat lengths were estimated for each CCG allele length from 6 to 11 repeats.

3.2.4.2 Haplotype analysis of normal chromosomes in the Venezuelan population

Analysis of CCG repeats revealed the HD normal alleles presented five different haplotypes with a size of 6, 7, 9, 10 and 11 CCG repeats in the unaffected and affected individuals from the Venezuelan population. We determined the haplotype in 1,066 normal chromosomes of individuals from HD families. We found 45 different haplotypes in the normal chromosomes of HD families. Haplotype CAG 17-CCG 7 has the maximum frequency (14.1%), and CAG 18-CCG 9 has the second highest frequency (9%) among all haplotypes (Figure 3-17).

CAG 17-CCG 7 has the maximum frequency with haplotype 7 CCG. The second most frequent allele in haplotype 7 was 15 CAG repeats (8.3%). The haplotype analysis for 10 CCG revealed CAG 17-CCG 10 has the highest frequency (8.3%), followed by CAG 19-CCG 10 that has the second most frequent allele (6.1%) in haplotype 10. With haplotype 9 CCG, CAG 18 has the highest frequency (9%). For haplotype 6 and 11 CCG, 15 CCG has the highest frequency of 0.1% and 0.38%, respectively. These haplotypes (6 and 11 CGG) were least common and were associated with small numbers of chromosomes.

The mean CAG length was calculated for the Venezuelan population for the five haplotypes as shown in Table 3-3. A mean CAG repeat size of 19.78± 0.21 was observed in individuals carrying 7 CCG haplotype, that is higher than the mean CAG repeat size in individuals with 9 and 10 CCG haplotype. The 9 and 10 CCG alleles were associated with a similar mean CAG repeat size. This is similar to haplotype analysis for the Scottish population in which 7 CCG has the highest CAG repeat mean size and the 9 and 10 CCG alleles were associated with a similar mean CAG repeat. The least common haplotypes (6 and 11 CCG) were associated with a smaller mean of CAG repeat lengths.

In addition, no effects of the size of the CCG repeat in normal alleles for haplotype 9 and 10 were observed for the size of CAG repeats, since individuals have no significant differences in the mean of CAG repeat size for these haplotypes. Haplotype CCG 7 was the most common haplotype in normal alleles, which represents 56.19% and 60.23% for the Venezuelan HD families and the Scottish population, respectively. CCG 7 was associated with longer CAG repeat lengths on normal chromosomes in Scottish and Venezuelan populations.

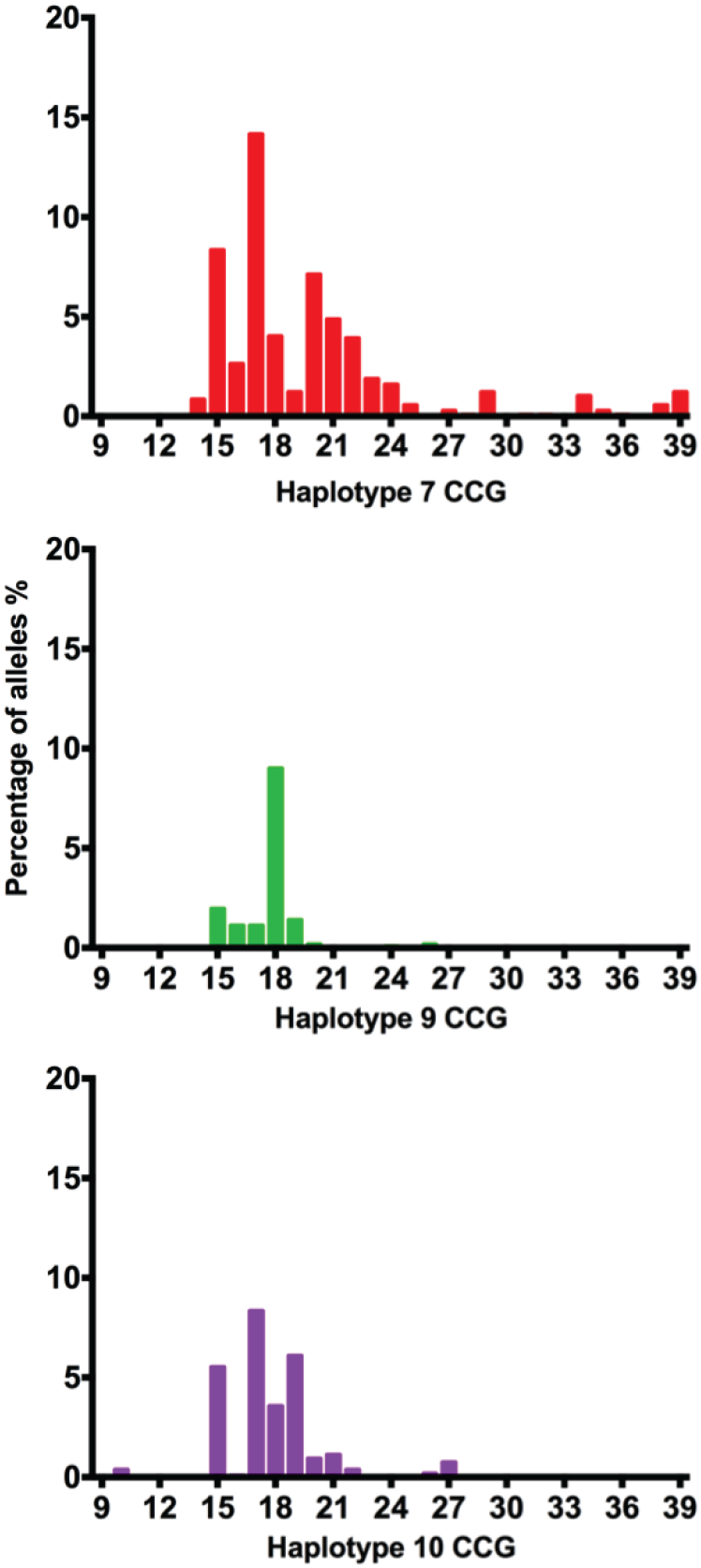

Figure 3-18 Haplotype analysis of *HTT* in unaffected and affected individuals of HD families from the Venezuelan population. The graphs show the distribution of CAG size on normal chromosomes in the 1,066 normal chromosomes for individuals carrying the 7 CCG, 9 CCG and 10 CCG.

3.2.5 Characterization of atypical alleles

The sequence of the 12 bp segment between the CAG and the adjacent CCG tracts in the HD gene may also significantly contribute to the instability of the CAG repeats. We were able to identify atypical HD alleles from sequencing *HTT* alleles in Scottish and Venezuelan populations. Atypical HD alleles were different than the typical structure: $(CAG)_nCAACAGCCGCCA(CCG)_n$.

3.2.5.1 Characterization of atypical alleles in the Scottish population

In our study, five different variants were identified from sequencing *HTT* normal chromosomes in unaffected individuals from the Scottish population. The first variant was observed in 13 chromosomes and was only seen in alleles having 15 CAG or 16 CAG with 7 CCG repeats (Figure 3-19). This variant was characterized by changes of the adjacent CCA CCG into a CCG CCG sequence. This variant could result from A→G substitution or a duplication of CCG repeats and deletion of the CCA codon. It is a silent mutation as both codons still code for proline. This variant has been reported in Margolis et al, 1999 as a rare variant that was found in one case among 1,236 cases tested for HD (0.04% of normal chromosomes). This variant was found in an unaffected spouse of a patient with HD. In contrast, this variant was found in ~ 4% of normal chromosomes in our study. That is higher compared to published data, which may be as our sample size is very small or because this variant is at a higher frequency in the Scottish DM1 population.

The second atypical allele involved the changes of the adjacent CAA CAG into CAG CAG sequence (Figure 3-19). This variant could result from an A→G substitution or a duplication of CAG repeats and deletion of the CAA codon. This variant was observed in two chromosomes and was only seen in alleles having 24 CAG and 29 CAG with 9 CCG repeats.

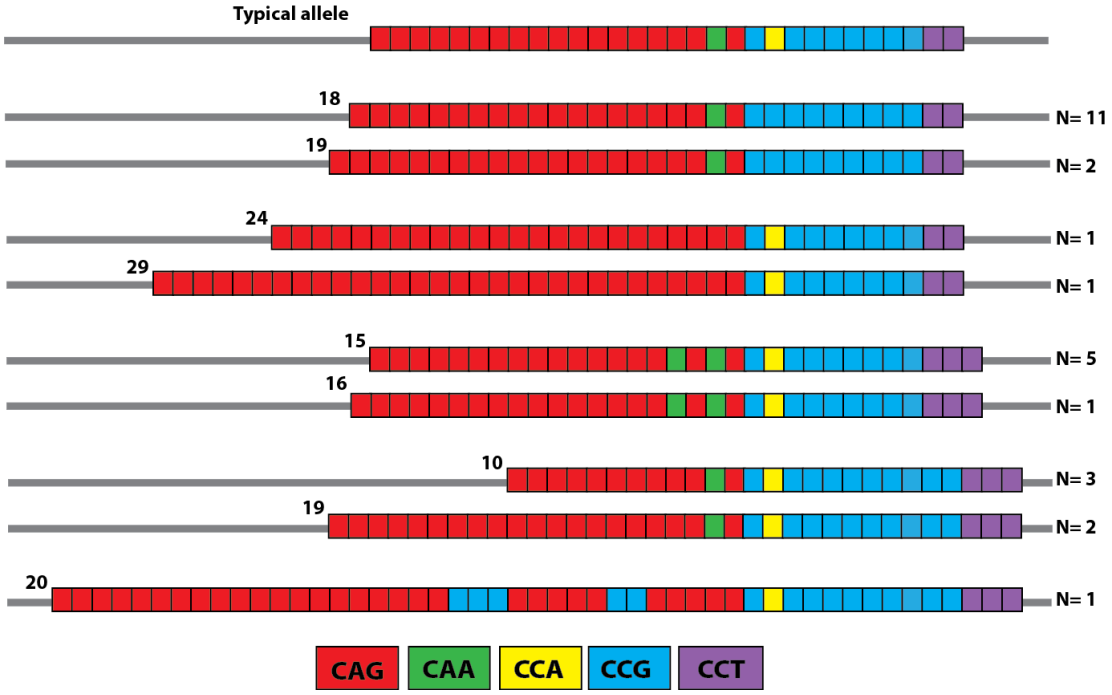

Figure 3-19 MiSeq sequencing of 27 atypical HD alleles. Five different structures were seen in unaffected individuals of the Scottish population. The majority of chromosomes (393) have the typical structure of the gene: $(CAG)_n$ CAACAG CCGCCA $(CCG)_n$, which is presented in the schematic representation on the top. The atypical alleles presented here are different from the typical structure in the intervening sequences and also within the repeats (CAG, CCG and CCT). The number on the right corresponds to the number of times each of these allele types was genotyped. The number on the left corresponds to the number of CAG repeats, which were associated with each allele.

The third variant that was analysed involves an additional CAA CAG at the 3'-end of the CAG repeat, along with the addition of a CCT sequence to the CCT duplication following the CCG repeat (Figure 3-19). This allele was observed in six normal chromosomes in our data from the Scottish population. This atypical allele was only seen in alleles that have 15 CAG or 16 CAG with 7 CCG repeats. This variant could result from a duplication of the CAA CAG sequence at the 3'-end of the CAG repeat, or a base substitution of G→A in the second last CAG of the uninterrupted CAG stretch. CAA still codes for glutamine. Also, the addition of a CCT sequence to the CCT that follows the CCG repeats could be a duplication of CCT normally present at the end of the CCG repeats or a base substitution of G→T for the last CCG repeat. This allele is more likely to have occurred because of length change mutations rather than a base substitution because otherwise it would have originated from an extremely rare 8 CCG allele. This atypical allele has been reported in previously published data (Pêcheux *et al.*, 1995).

The fourth atypical allele also involved an additional CCT sequence at the end of the CCG repeat (Figure 3-19). This atypical allele was observed in five chromosomes. This allele was found with 9 CCG repeats and CAG repeat size of 10 and 19. Therefore, the CCT triplet number may also account for the size variation of the CCG region in *HTT* chromosomes. CCT repeats can be either two or three (rare) triplets in length.

The last atypical allele has not been described before and is associated with interruption of CCG repeats into the CAG repeats tract, deletion of CAA in the intervening sequence along with an additional CCT sequence to the CCT duplication present at the end of the CCG repeat.

3.2.5.2 Characterization of atypical alleles in the Venezuelan population

We identified three atypical normal allele structures from sequencing HD 1,066

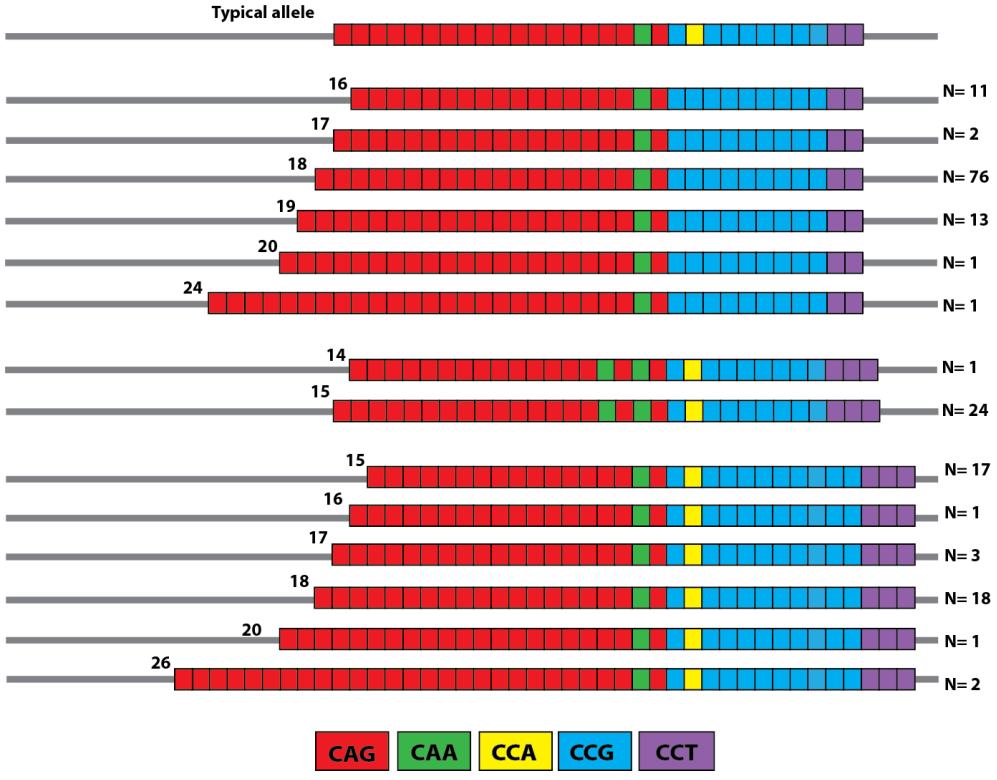

Figure 3-20 MiSeq sequencing of 107 atypical *HTT* alleles. Three different structures are seen in unaffected and affected individuals of HD families from the Venezuela population. The majority of chromosomes (959) have typical structures of the gene: $(CAG)_n$ CAACAG CCGCCA $(CCG)_n$ that is represented in the schematic representation on the top of the figure. The number on the right corresponds to the number of times each of these allele types was genotyped. The number on the left corresponds to the number of CAG repeats that were associated with each allele.

normal chromosomes from unaffected and affected individuals of the Venezuelan population.

The first atypical allele was observed 104 times (Figure 3-20). This allele is characterized by a change of the CCA CCG into CCG CCG sequence. This atypical allele was observed with alleles having 9 CCG repeats and a variable number of CAG repeat: 16, 17, 18, 19, 20 and 24 repeats. A high number of alleles (76) have 18 CAG repeats. This atypical allele was the most common allele among the three atypical alleles.

The second atypical allele involves an additional CAA CAG sequence at the 3'-end of the CAG repeat, along with an additional CCT sequence at the end of the CCG tract (Figure 3-20). This allele was observed in 25 normal chromosomes of unaffected and affected individuals. This atypical allele was seen mostly in alleles having 15 CAG with 7 CCG repeats, and in one allele with 14 CAG repeats.

The last atypical allele was observed 42 times, which involves an additional CCT sequence at the end of the CCG repeat. This allele was found with 9 CCG repeats, and a variable number of CAG repeats: 15, 16, 17, 18, 20 and 26 (Figure 3-20). The most common CAG repeat lengths associated with this atypical allele were 15 and 17 repeats.

All three atypical alleles were also found in Scottish samples sequenced in the previous section. There are no atypical alleles identified in this cohort from the sequencing of intermediate or reduced penetrance HD alleles. The actual alignment for three atypical alleles is illustrated in Figure 3-21.

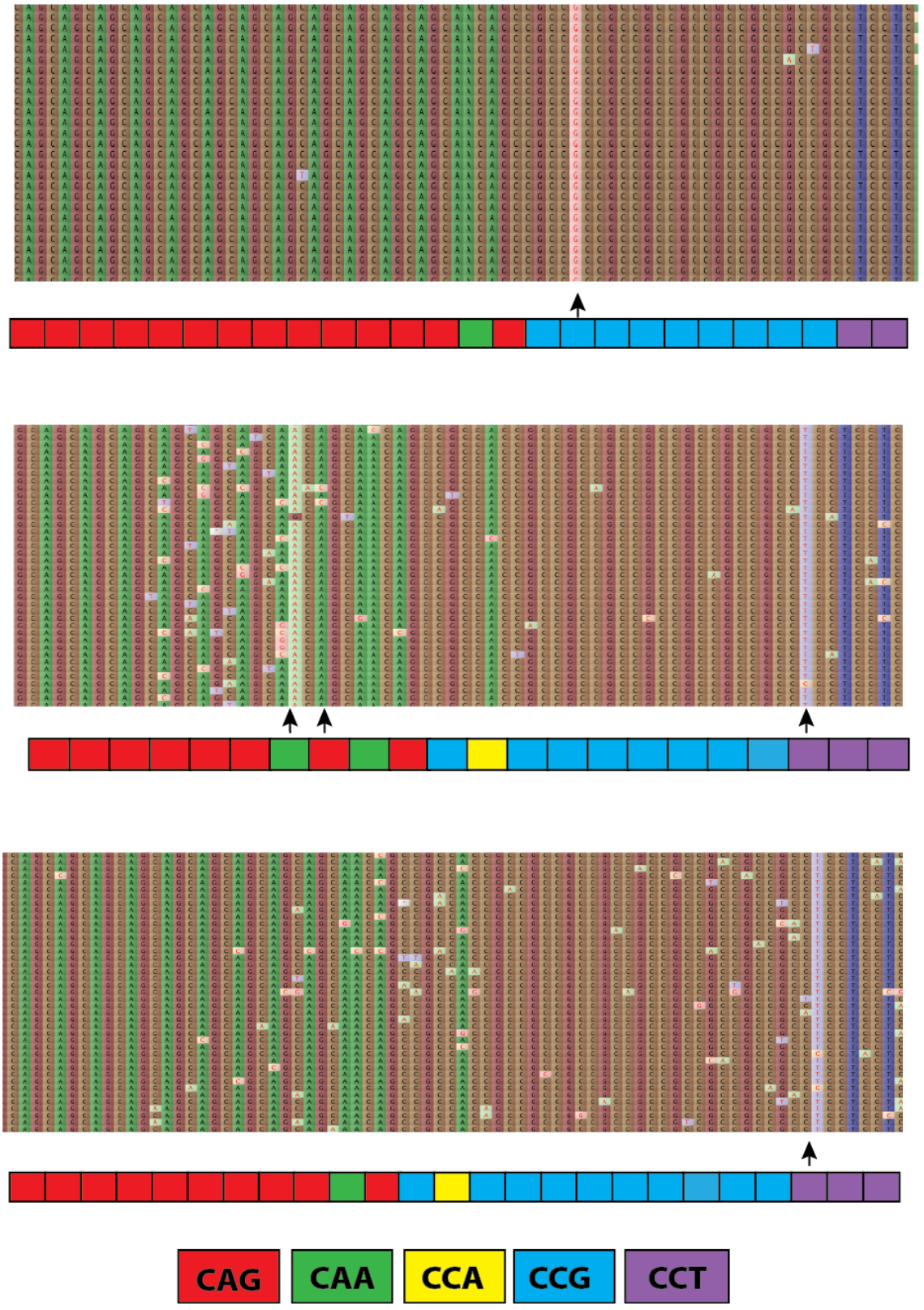

Figure 3-21 MiSeq sequencing reads of atypical HD alleles. Three alleles were shown in our data from Scottish and Venezuelan populations. Atypical HD alleles were different than the typical structure: $(CAG)_nCAACAGCCGCCA(CCG)_n$. These atypical alleles were described in detail in sections: 3.2.5.1 and 3.2.5.2. Black arrows indicate the change in each atypical allele from the typical allele structure. Tablet was used for visualisation the aligned reads.

3.3 Discussion

Since the discovery of the *HTT* gene, molecular genotyping in HD has been limited by examining the CAG repeat by automated fragment length analysis of PCR products. It is complicated by the presence of an adjacent polymorphic CCG repeat and provides no information on variant repeats or flanking sequence differences. In this chapter, we discussed the methodological and bioinformatic issues associated with next generation sequencing to develop new sequence-based approaches to high-throughput genotyping of the CAG repeat in HD. This new approach makes it feasible to consider NGS as a clinical tool in the near future. We have demonstrated the ability of NGS technology using the MiSeq platform to be able to genotype and characterize the size distribution of the CAG repeats of the *HTT* gene, including the polymorphic CCG repeats and the flanking sequences.

It is important to note the sensitivity of this sequencing method and the ability to identify sequence variants within the repeat region. The accuracy observed in sequencing the repeat and flanking CAG repeat region at great coverage produced sufficiently accurate sequence to genotype repeat number and detect variant repeats.

We have developed an optimised library preparation method using this approach. We can test 96 samples in each library. We established that it was possible to sequence and genotype normal HD CAG alleles and the adjacent CCG repeat using the MiSeq platform by amplifying the region using locus-specific primers combined with MiSeq sequencing adapters. The optimal primer pair (HD319 F/33935.5) was used for MiSeq sequencing to ensure sequencing of the region immediately flanking the repeat, which could be polymorphic, and also to minimize the length of the flanking region. Reduced flanking region would potentially allow the sequencing of longer CAG repeats. Although these primers (HD319 F/33935.5) produced a prominent primer dimer, it was possible to separate the desired products via an AMPure beads based approach because of the greater size difference between the full-length amplification products and primer dimers. This step is essential for product purification and to generate enough products to yield excellent HD sequence reads.

To investigate the utility of MiSeq sequence read data to genotype potentially longer CAG repeats (170 repeat allele), we evaluated the use of a MiSeq sequencing run up to 600 bp. Two sequencing libraries were prepared using locus-specific primers incorporating the sequencing adapters for the same DNA samples. After that, the two sequencing libraries were then sequenced using a 600 bp single direction run, and a paired-end read run (2x300 bp). Although we conducted a 600 bp unidirectional read, most of the fragments were less than 600 bp in length. However, most of the reads generated were 600 bp long. At least part of the reason for this is that we sequence through the fragment and into the adapter sequence. In this case, the adapter sequences would be present in the output file at the 3'-end of the reads. However, some of the sequences were even longer than that, i.e. they extended beyond the end of the fragment length. The reason for this is unclear but could be due to some sequencing signal from neighbouring sequence clusters on the MiSeq flow cell. A major drop in sequence quality was observed after about 400 bp. Thus simple quality trimming cannot be used to address this problem. Sequencing CAG repeats along with flanking region and CCG repeats from the same direction can be performed for most of the alleles before a significant drop off in sequencing quality. However, sequencing quality and mapping efficiency were lower when sequencing longer alleles because a drop in sequencing quality occurred within the CCG repeats and at the 3-end flanking regions. Thus, we have explored the utility of the reverse reads when conducting sequencing reaction in the 2x300 bp or 400x200 bp sequencing format. These data have revealed that good quality sequences may also be obtained on the reverse strand and that these can be used to improve the efficiency of genotyping the CCG repeat and the 3'-flanking DNA sequence. Thus, sequencing longer CAG alleles can be done from one direction, and the CCG repeats from the other direction because we cannot match or merge the reads in the middle of the CAG or CCG repeats. Therefore, using the 2x300 bp or 400x200 bp run to sequence HD alleles that performed better than at 600 bp read for MiSeq in terms of the number of reads and mapping coverage.

The obtained sequencing reads were then processed using CLC genomics workbench software as an NGS mapping tool for analysing the read alignment and visualizing the data. CLC can be used for alignment of both paired-end and single-end reads. The analysis of our data proceeded via quality assessment of

the raw data, trimming, demultiplexing, read alignment to the references, visualization of the data and genotype determination. The reads were aligned to custom references including the region flanking the repeats and various numbers of CAG and CCG repeats that yielded a greater alignment for each sample in comparison to a single reference. The optimal alignment parameters of the reads were determined based on the accuracy and the highest proportion of reads that were aligned to the reference. Adjusting these parameters improved the alignment quality to allow us to genotype normal *HTT* alleles. Examination of actual alignments revealed that highly accurate read alignments were generated, across the repeat region.

The aligned reads showed differences from the references that are shown as mismatches throughout the reads. These differences might occur as a result of PCR error during library preparation or bridge amplification during MiSeq sequencing, genuine somatic variation or error in base calling during sequencing.

Using this approach, we sequenced normal *HTT* alleles in 210 unaffected Scottish individuals and 733 unaffected and affected individuals from HD families from Venezuela. These data have revealed accurate genotypes, i.e. the exact length and sequence of both CAG and CCG repeat for most individuals. The length of the CAG repeats varied from 9 to 32 repeats in 210 unaffected Scottish individuals, with the 17 CAG allele occurring most frequently (30%). The size of CAG repeats in the Scottish population was similar to those reported among Caucasians (Kremer *et al.*, 1994; Squitieri *et al.*, 1994; Costa *et al.*, 2006; Warby *et al.*, 2011).

The Venezuelan kindreds are unique in that they comprise the largest genetically related HD community in the world. The normal allele distribution for Venezuelan population was 10 to 39 CAG repeats among 333 unaffected individuals and 400 HD patients. The most common allele was 17 repeats found in 21.1% of normal chromosomes of unaffected individuals and 28.8% in HD patients. The distribution of the CAG repeat size in the Venezuelan population was similar to that observed in HD families in the same population (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004).

Our findings suggest that intermediate alleles (27 to 35 CAG) for HD are relatively common among individuals in the Scottish DM1 population with no association with HD. The relative number of intermediate alleles found in our data was 5% (21/420) in the Scottish population. The allele frequency is higher than in a previous study that has examined the frequency of intermediate alleles in Western populations (3.13%)(458/14,630) (Kay *et al.*, 2016). However, there were marginally significant differences in the intermediate allele frequencies (chi-square= 5.25, P =0.021) between the normal allele in Scottish population and the published result in three Western countries (United States, Scotland and British Columbia, Canada).

The intermediate allele frequency was 4.65% of unaffected individuals and 2.5% of affected individuals from Venezuelan HD families. The intermediate allele frequency was 3.84% (41/1,066) in HD families from our data. Alleles in the intermediate range were estimated at 4.34% in the general population from Brazil (Raskin *et al.*, 2000). That is similar to our result of HD families in Venezuela. Intermediate alleles are linked with a higher risk of expansion in the next generation, giving rise to new cases of HD. Approximately 10% or greater of HD new mutations cases, who develop the clinical disease, have parents who were carriers of intermediate alleles (Falush *et al.*, 2000). Intermediate alleles which expand into the reduced penetrance or full mutation range of HD, are more likely seen with a paternal transmission (Goldberg *et al.*, 1993; Telenius *et al.*, 1993). The chance of transmitting an expanded allele to the offspring is higher with longer CAG repeats (Brocklebank *et al.*, 2009). A study showed when intermediate alleles have more than 30 CAG repeats, the higher will be the chance of transmission of an expanded allele, with complete penetrance, to the next generation (Semaka *et al.*, 2013).

There were no reduced penetrance alleles identified in the Scottish population. However, 0.1% (15/14,630) of alleles were identified within the reduced penetrance range among tested alleles in Western populations (Kay *et al.*, 2016). The reduced penetrance allele frequencies on normal chromosomes of the Scottish population were not significantly different (chi-square P=0.4) from Western populations in general. The undetected reduced penetrance alleles in Scottish population may be due to our study having a small sample size. Utilizing

much larger samples of different ethnic populations is required to estimate the frequency of reduced penetrance alleles in the general population.

There were nineteen reduced penetrance HD alleles (36 to 39 CAG) identified in the unaffected Venezuelan population and one allele identified in an HD patient. This presented an allele frequency of 2.85% in unaffected and 0.25% in the affected Venezuelan population. This gives a frequency of reduced penetrance alleles of 1.8% (20/1,066) within HD families in Venezuela. The proportion of reduced penetrance alleles varied between 1 and 2% in individuals from HD families previously analysed in the same population (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004; Paradisi, Hernández and Arias, 2008). We obtained a similar frequency compared to previous studies in the Venezuelan population.

For these experiments, although the Scottish DM1 patient's population was analysed as a control, they are not a random selection of the Scottish population with respect to DM1 patients. However, there is no known association between DM1 and HD. Both disorders are caused by independent expansions at different genetic loci. There is no reason to think that HD will be under or over represented in DM1 patients. Therefore, with respect to the HD locus, these DM1 patients should be representative of the general population in Scotland. However, it can be ruled out that there may be a bias in potential modifier loci variants in this population.

Among the individuals studied in the Venezuelan population, alleles in the intermediate and reduced penetrance ranges were observed more in unaffected individuals than in affected. The differences between them may be because we tested the non-mutant chromosome in affected individuals, and both chromosomes in the unaffected individuals. Some of the intermediate and reduced penetrance alleles in the unaffected individuals may arise from contractions or deletions transmitted from affected individuals. The second allele is prone to expand to a full mutation in unaffected individuals when transmitted to the offspring.

Intermediate and reduced penetrance allele frequencies are high in our samples, which could correspond to a high prevalence of HD in the Venezuelan

population. Also, the observed difference in intermediate and reduced penetrance alleles between the Scottish and Venezuelan populations may be due to a difference in the ethnic origin.

Individuals with reduced penetrance alleles may show the HD phenotype in 40% of cases (Agostinho *et al.*, 2013), such that they may have a very late onset of symptoms, which may be similar to the classic HD phenotype, milder symptoms, or other clinical presentation unrelated to classic symptoms. It is worth examining a large number of individuals with reduced penetrance alleles to determine if they have any clinical presentation over a long period of time. Those individuals with reduced penetrance alleles may also represent the source of new mutations for HD.

We emphasize the importance of determining the frequency of intermediate and reduced penetrance alleles in the general population and within HD families because of the risk of expansion and subsequent risk of new mutation. Also, individuals who have CAG sizes in the intermediate or reduced penetrance sizes range may benefit from better information on the risk estimate of developing the disease, and the risk of new mutations in their offspring.

CCG 7 alleles were the most common in both populations. In the Scottish population, the 7 CCG repeats were found at 60.23% and at a frequency of 56.19% in the Venezuelan population. The CCG 10 repeat was the second most common allele in both populations being present with a frequency of 30.95% and 27.39% in Scotland and Venezuela respectively. Together those two alleles (7 and 10 CCGs) account for the majority of normal chromosomes in both populations.

The distribution of CCG repeats in the Scottish population is similar to that in other western European populations (Andrew *et al.*, 1994; Squitieri *et al.*, 1994) in which CCG 7 and 10 were the most prominent alleles. However, we did not observe any 8, 11 and 12 CCG repeats which have been found in previous studies. The distribution of CCG repeat sizes in the Venezuelan cohort showed a difference to that in other Latin American populations (Agostinho *et al.*, 2012). We have observed the 9 and 11 CCG alleles, but the previous study showed

individuals of HD families having 6, 7, 8 and 10 CCG repeats. Also, we did not observe any 8 CCG repeat alleles in our data.

No difference was found between the numbers of CCG repeats in the unaffected and affected individuals of HD families. However, the allele frequencies of the CCG repeat in the HD families were different to those obtained in the Scottish population. There were three alleles identified in the Scottish population, while five alleles were identified in the Venezuelan cohort. The 6 and 11 CCG repeats were not found in Scottish individuals.

Haplotype analysis, comprising CCG and CAG polymorphic markers, revealed the existence of two haplotypes that were over-represented among our data. Haplotype CAG 17-CCG 7 has the highest frequency, and CAG 17-CCG 10 has the second highest frequency among all haplotypes in the Scottish population. Haplotype analysis in the Venezuelan population showed that the CAG 17-CCG 7 has the highest frequency and CAG 18-CCG 9 has the second highest frequency among all haplotypes. Therefore, there is a difference between the two populations in the frequency of the second haplotype.

The carriers of 7 CCG repeats were associated with a higher CAG mean and longer CAG repeat lengths on normal chromosomes in both populations. The difference was not significant and was not associated with differences in mean CAG size between the two populations. The mean CAG repeat size was higher in normal chromosomes containing 7 CCG allele than in those with 9 and 10 CCGs. The 9 and 10 CCG haplotypes were associated with a similar mean CAG repeat size. The distribution of CAG repeat sizes among the 7 CCG haplotype was similar to the total distributions for both. There was no large variation in the mean CAG size among the Scottish and Venezuelan populations.

Our data have revealed the existence of atypical alleles from sequencing normal *HTT* alleles. Most of the atypical alleles have been reported in published data, suggesting that we determined the accurate genotyping for HD alleles. The similarity of our data confirms the general validity of our interpretation. Thus, we are confident that these atypical alleles are genuine germline variants existing in normal alleles, and are not PCR and/ or sequencing errors.

Five atypical haplotypes were analysed in 27 normal alleles from sequencing *HTT* alleles in the Scottish population. One of these atypical alleles identified was with an intermediate repeat allele (29 CAG repeats). We observed three atypical normal allele structures from 107 normal alleles from the Venezuelan population. All the three atypical alleles were also found in the Scottish population from our data. There were no atypical alleles identified in this cohort from the sequencing of intermediate or reduced penetrance alleles in the Venezuelan population.

Among the Scottish population, only one of these atypical alleles was not a previously described HD allele. This last atypical allele is associated with interruption of CCG repeats into the CAG repeat tract, deletion of CAA in the intervening sequence, along with an additional CCT sequence to the CCT duplication present at the end of the CCG repeat.

The variant repeats may potentially expand less or more in both the soma and germline and can modify the disease severity. For instance, the atypical allele was seen for the first time in our data, may stabilise the repeat in both soma and germline and may be associated with reduced severity if it is found with HD expanded alleles. Some variant repeats may have an impact leading to increase severity and level of instability. For example, variants that lead to longer uninterrupted CAG repeats tracts are predicted to be more unstable and biased toward expansion. The presence of atypical allele structures may have an impact on the stability of the DNA and possibly the symptoms. Those atypical alleles in HD are present in the general population, and they expand to the disease size range.

Variant repeats in DM1, SCA1 and Fragile X syndrome have been shown to modify the mutational dynamics by reducing the amount of germline and somatic instability (Chung *et al.*, 1993; Zhong *et al.*, 1995; Musova *et al.*, 2009; Braida *et al.*, 2010). Thus, we expect the atypical alleles with more variant repeats to be more stable, and those with less variant repeats to be more unstable, predisposing alleles to expansion and eventually to disease status. The level of instability is greater for longer uninterrupted CAG repeat tracts, and these are biased toward expansion. Notably, the presence of CCG interruptions which break the CAG repeat into three smaller repeat tracts in our data may stabilise

the alleles with an overall repeat number greater than 35. We expect these alleles to be associated with less somatic instability and to be less prone to expansion when transmitted to the next generation. Therefore, for most of the trinucleotide repeats disorders, the interruptions provide genetic stability to the repeat tract and the interrupted tracts are less likely to expand upon transmission. The identification of those atypical alleles can be difficult in the diagnostic test using PCR analysis, as it does not take into account the allele structure. Therefore, sequencing those alleles is necessary as they have an implication for the individual's risk as well as alleles stability through life and transmission to offspring.

By sequencing these atypical allele structures, we have clarified some of the discrepancies between our results and previously published data. First, the CAG 15 alleles were relatively underrepresented in previous studies. This can be explained by the presence of atypical alleles that involved an additional CAA CAG sequence at the 3'-end of the CAG repeat, along with an additional CCT sequence at the end of the CCG tract. This atypical allele was seen mostly in alleles having 15 CAG. Therefore, using fragment length approaches for genotyping this allele would give the incorrect genotype of 17 CAG repeats because of the assumption that the length of the fragment is equal to the repeat number. Also, we didn't find any 8 CCG alleles in our data. Although, it was reported as a rare allele in a previously published paper (Squitieri *et al.*, 1994; Pramanik *et al.*, 2000; Costa *et al.*, 2006). This also might happen as the result of an atypical allele with an additional CCT sequence at the end of the CCG tract. That could lead to inaccurate determination of CCG repeat length using a standard method that does not take into account any structural variation of the region, and based on the assumption that the length of the fragment is equal to the repeat number. Therefore, the genotype obtained for this allele would be 8 CCG and 2 CCT of the 3'-end of the CAG repeats. These atypical allele structures would be almost impossible to characterise accurately without NGS.

Analysis of HD alleles sequences has shown that the presence of CAA, CCT, etc within the alleles could yield a distinct pattern of mutations termed as mutational signatures. These mutational signatures may relate to DNA repair processes. We have not searched for mutational signature specific to HD alleles,

as germline sequence changes are present at a low frequency in our data. However, a detailed analysis using a larger data set could be performed to detect mutational signatures in HD alleles. Understanding the link between the HD alleles and mutational signatures is relevant given the implications for the development of therapeutic approaches potentially directed towards a typical allele structure.

The limitation of fragment length approaches for genotyping is inaccurate genotyping of CAG and CCG repeats because of the assumption that the length of the fragment is equal to the repeat number. Therefore, it is important to shed light on the classification of *HTT* alleles by repeat length. This could facilitate data discussion and analysis among different studies, and also would lead to more reliable studies for determination of actual frequencies of different allele ranges in the general population and in HD families. We noticed the variation of *HTT* allele frequencies between our data and previously published data. That can be resolved by using the classification of *HTT* alleles by accurately determining the repeat length.

In summary, the classification of HD alleles by repeat length would lead to better data sharing and comparison as well as the actual determination of allele frequencies of different allele ranges. NGS approach represents an efficient method compared to other testing methods, which can be applied using new, rapid sequencing technology and offer diagnosis within a few days. Genetic testing using NGS to determine the length of the CAG repeat is essential as it provides a definite diagnosis of the disease. Therefore, NGS is more likely to have significant clinical utility. We illustrate the success and challenges of using NGS in trinucleotide disorders and suggest that our experience is likely to be applicable to many other expanded repeat alleles.

Chapter 4 Development of next generation sequencing based approaches to genotype the Huntington disease CAG repeat

4.1 Introduction

HD prevalence is estimated at approximately 4-8 per 100,000 in Caucasian populations (Harper, 1992), but occurrs at variable prevalence in different parts of the world. In Japan and China, there is a lower frequency (0.1-0.5 per 100,000 persons) (Leung *et al.*, 1992; Rawlins *et al.*, 2016). The lowest frequencies have been found in black South Africans, with 0.02 affected persons per 100,00 but this could be an underestimate of prevalence because there were only a few cases detected (Rawlins *et al.*, 2016). However, geographic isolates of Lake Maracaibo in Venezuela has a prevalence of 7 cases per 1,000 (Castilhos *et al.*, 2016). This is a very high frequency of the disease and is due to a founder effect

The CAG repeat length is the primary determinant of age at onset of clinical symptoms (Duyao *et al.*, 1993; Andrew *et al.*, 1994). The inherited repeat length is inversely correlated with age at onset, in which the contribution of each additional CAG repeat unit reduced the age at onset by approximately 2 years (Andrew *et al.*, 1993; Gusella and MacDonald, 2009). Therefore, accurate determination of CAG repeats size has an implication for determining the average age at onset, and also for genetic counselling for at risk and affected individuals. Symptoms onset usually begins between 35 and 55 years of age, and the median survival time is 15 to 18 years. About 10% of patients have juvenile HD onset (before age 20 years) that occurs in individuals with repeat lengths of more than 65 CAG repeats (Telenius *et al.*, 1993). Juvenile cases tend to be inherited through the male germline.

The length of the expanded HD CAG repeats is unstable when transmitted from parents to offspring. It has been shown that the CAG repeat length changes are biased toward increases in most cases, but male transmission can result in large increases in size (De Rooij *et al.*, 1993; Duyao *et al.*, 1993; Telenius *et al.*,

1993). Also, the HD anticipation phenomenon has been shown to be more intense in paternal transmission, that is offspring having symptoms at earlier ages of onset and longer repeat length than their affected parents (Telenius *et al.*, 1993).

Most HD is inherited, but there are some occurrences of new sporadic HD cases, in which an affected individual have the disease but has no previous family history (Myers *et al.*, 1993). The frequency of these chromosomes in the general population is quite low. Sporadic HD cases can occur from the instability of intermediate alleles that may lead to expansion into HD affected range upon transmission.

The *HTT* repeat region is complex because it comprises a polymorphic mixture of CAG and CAA glutamine codons followed by another polymorphic mixture of CCG, CCA, and CCT proline codons. The fully penetrant range of the HD mutation is 40 or greater CAG repeats. The CAG repeats at *HTT* gene vary between 40 to 120 CAG repeats in expanded alleles. The CAG repeat is immediately adjacent to a polymorphic CCG repeat that varies from 6 to 12 repeats in length (Andrew *et al.*, 1994; Squitieri *et al.*, 1994; Pêcheux *et al.*, 1995), although a CCG 4 allele was the smallest repeat identified in India (Pramanik *et al.*, 2000). Most normal chromosomes and the majority of disease chromosomes are strongly associated with the 7 CCG repeats allele in high prevalence areas of HD (Barron *et al.*, 1994; Squitieri *et al.*, 1994; Costa *et al.*, 2006). In Japan where HD has a low prevalence, HD chromosomes are strongly associated with an allele of CCG 10 (Morovvati *et al.*, 2008).

Molecular genotyping in HD is traditionally performed by fragment length analysis of PCR products amplified across the CAG repeat tract. This method fails to identify atypical sequence variants or flanking sequence differences but provides an estimation of the size of CAG repeat tract. In this project, we have addressed many of the methodological issues associated with next generation sequencing (NGS) to develop new sequence based approaches to high-throughput genotyping of CAG repeats.

Previously, we established that it was possible to sequence and genotype normal HD CAG alleles and the adjacent CCG repeats using the MiSeq platform by

amplifying the region using locus-specific primers combined with MiSeq sequencing adapters (see Chapter 3). NGS technology has been shown to be more informative and more accurate than methods that depend on the fragment length analysis.

To establish whether this approach (MiSeq sequencing) is also suitable for genotyping the HD locus in affected individuals, we have initiated the sequencing of the HD repeat in buccal cell DNA from the Venezuelan Lake Maracaibo population. The aim of this chapter was to sequence and genotype HD alleles in HD patients using MiSeq sequencing.

4.2 Results

4.2.1 Library preparation

As described in Chapter 3, we established that it was possible to sequence and genotype normal HD CAG alleles and the adjacent CCG repeats using the MiSeq platform by amplifying the region using locus-specific primers combined with MiSeq sequencing adapters. Although these primers produced a prominent primer dimer, we managed to separate the products by AMPure beads based approach and generate a product that yielded good quality sequencing data. The AMPure bead based approach was used because of the large size difference between the full-length amplification products and primer dimer.

We have noticed in sequencing the HD repeat that PCR amplification and sequencing bias yielded more reads for shorter alleles. This had been observed in heterozygotes where the smaller allele is clearly sequenced to a greater depth than the larger allele (see Chapter 3). It was clear that the difference in the number of reads between the two alleles was dependent on the relative difference in size between those alleles. The most common normal allele in HD is 17 CAG repeats; however, the vast majority of expanded alleles only carry 40 to 50 repeats. This is not a large length difference suggesting that it may not be necessary for gel purification for the expanded alleles to be carried out.

After library purification, the PCR products were pooled, and the quality of the sequencing libraries assessed by capillary electrophoresis on a Bioanalyzer, to

check the fragment had the expected size and that primer dimer was absent or present in relatively small quantity.

Although we had previously generated good quality data, we noticed a drop in sequence quality was observed after about 400 bp when using a 600 bp run. Therefore, we explored the utility of reverse reads when sequencing longer HD CAG repeats. Thus, it is possible to genotype longer CAG repeats by sequencing through the CAGs and into the intervening sequence using a long forward read, while the CCG repeats can be genotyped using the shorter reverse read. This is necessary because reads cannot be matched or merged in the middle of the repeats. Our data have revealed that good quality sequences may also be obtained on the reverse strand, and that these can be used to improve the efficiency of genotyping the CCG repeat and the 3'-flanking DNA sequence. Therefore, using the 2x300 bp or 400x200 bp run to sequence HD alleles performed better than a 600 bp run for MiSeq sequencing.

Also, we previously observed with the data from a 600 bp MiSeq run (see Chapter 3) that many of the reads were longer than the DNA molecules sequenced. An examination of the sequencing reads that were longer than the template DNA molecules revealed that most of them corresponded to a high sequencing error rate at the end of those reads, making it more difficult for initial adapter trimming to correctly recognise and remove the sequencing adapters at the end of those reads. This problem can be resolved by performing an additional adaptor trimming procedure before mapping the reads to the reference.

To investigate the utility of the MiSeq sequencing approach to genotype expanded HD alleles, HD alleles were amplified from 409 buccal swab DNAs from the US-Venezuelan Collaborative Research Project consisting of affected individuals from HD families. 91 samples were prepared by my colleague Dr. Marc Ciosi working on a CHDI funded project. I carried out the data analysis for this library.

Sequencing libraries were prepared using locus-specific primers incorporating the sequencing MiSeq adapters, using the primer pair HD319 F/33935.5. Those sequencing libraries were sent to the Glasgow Polyomics Facility for sequencing using the MiSeq platform. The libraries were sequenced using either the 400

bp/200 bp or the standard 2x300 bp run parameters in both forward and reverse directions.

Trueseq barcodes for MiSeq sequencing have 96 unique combinations of indices generated by using 8 forward and 12 reverse indices in all possible combinations. These 96 combinations of barcodes may be used to identify 96 different samples, each associated with a unique pair of barcodes.

The sequencing reads obtained were mapped with CLC Genomics Workbench to reference sequences including various numbers of CAG and CCG repeats (1 to 200 CAG repeats and 1 to 20 CCG repeats) and the region flanking the repeats. See Chapter 3 for more details about: quality assessment of the raw data, alignment parameter optimisation, read alignment to the reference and genotype identification and visualization of the data. The optimal alignment parameters of the reads were determined based on the improved alignment quality and the highest proportion of reads that were aligned to the reference.

4.2.2 Genotyping of the HD alleles in affected individuals

Read mapping of expanded alleles revealed that the MiSeq sequencing approach can be used to genotype the expanded HD alleles. The sequencing reads were obtained for both the normal and the expanded allele for each individual. The vast majority of reads (>90%) could be correctly mapped to the reference sequences for each patient. In most samples, the expanded alleles were detected. Sequencing data have been generated for 409 samples. From this data, we were able to determine the genotype.

The MiSeq sequencing result for an HD patient is presented in Figure 4-1. The CAG repeat size distribution from that patient shows the mode at 25 and 43 CAG with 7 CCG repeats, i.e. 25 and 43 CAG repeats occur most frequently. That patient is heterozygous for 25 and 43 CAG repeats, as well as homozygous for 7 CCG repeats. The normal HD allele has 25 CAG repeats with 7 CCG repeats and the expanded allele has 43 CAG repeats and 7 CCG repeats. CAG and CCG PCR slippage are also seen, in which multiple reads were mapped against references of different lengths as seen in the allele distributions (Figure 4-1).

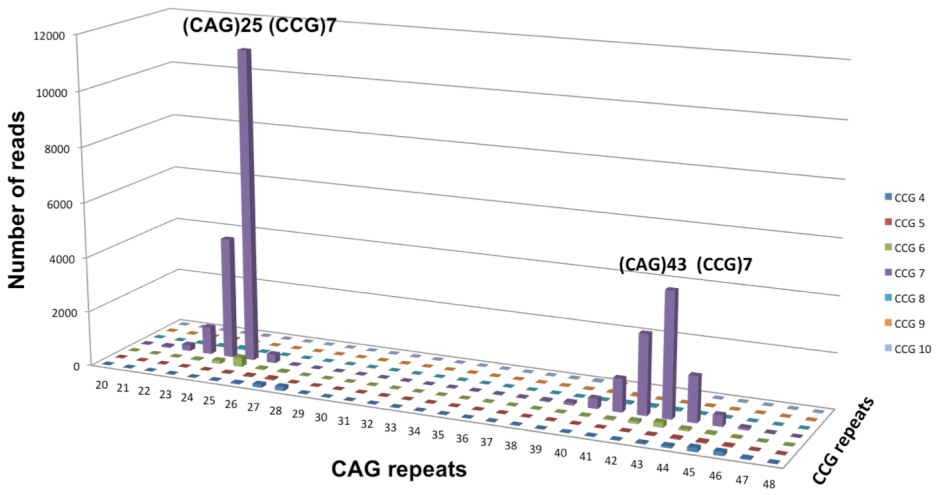

Figure 4-1 Genotyping of the CAG repeat in an HD patient using MiSeq sequencing. The number of sequencing reads mapped (y-axis) to the reference sequences associated with a different number of CAG repeats (x-axis). Allele length distribution is for one HD patient. The normal allele aligned against the 7 CCG repeats and 25 CAG repeats, represented as the highest peak. Also, expanded allele reads aligned against the 7 CCG repeats and 43 CAG repeats, represented as the second highest peak.

All expanded alleles detected had the typical allele structure: $(CAG)_nCAACAGCCGCCA(CCG)_n$ *i.e.* no atypical expanded alleles were identified in our data from the sequencing of Venezuelan HD patients. Atypical alleles structures have been previously detected from sequencing normal *HTT* alleles from unaffected individuals from the Scottish and Venezuelan populations (Chapter 3).

An example of mapped reads for three HD patients shows the CAG repeat at a mode of 40, 50 and 74 CAG repeats as shown inFigure 4-2. We used Tablet to visualize the read mappings results produced by the MiSeq sequencing (Milne *et al.*, 2013). Analysing the mapped reads confirms the presence of the expanded CAG alleles and 7 CCG repeats, and that both flanks to the repeats have the expected sequence, as no mismatches with the reference sequence were seen. These data confirm that it is possible to sequence HD expanded alleles at a mode of up to 74 CAG in the presence of normal allele. However, we are unable to reveal the 3'-flanking region for long CAG alleles (mode =≥74 repeats). Therefore, the reverse read can be used to genotype the CCG repeats and the 3'-flanking DNA sequence, and to detect any variants in that region and for haplotype analysis.

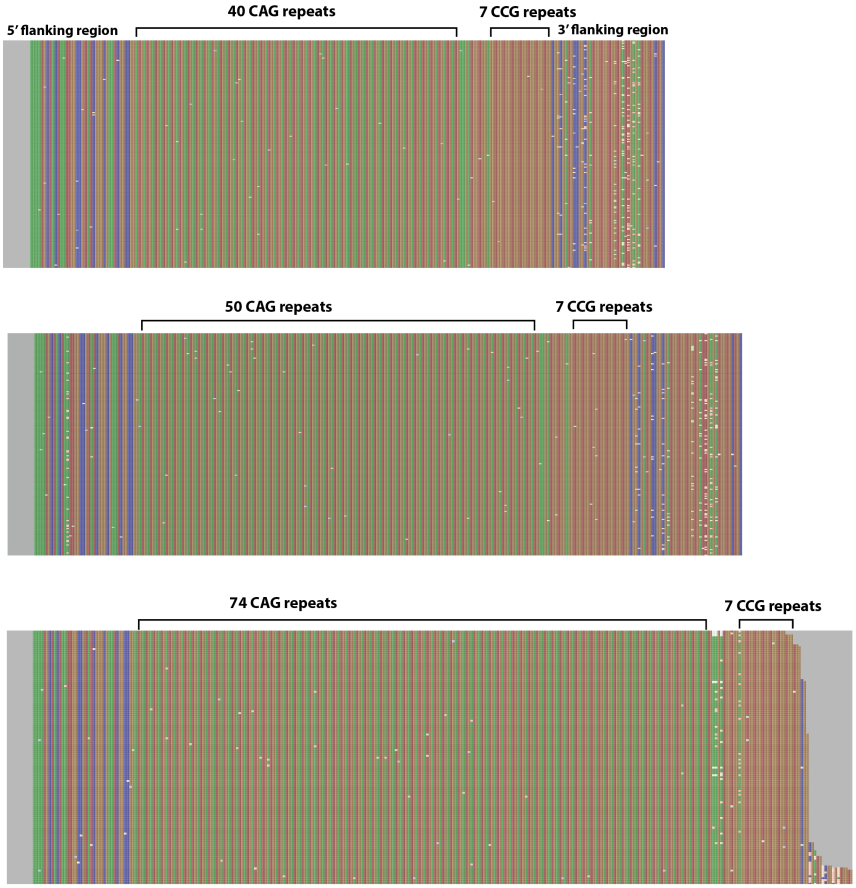

Figure 4-2 Mapped reads from the expanded alleles of three individuals with 40, 50 and 74 CAG repeats using MiSeq sequencing. Each sample reads were aligned against a set of reference sequences with (CAG) $_{1\text{-}200}$ and (CCG) $_{1\text{-}20}$ using CLC genomics workbench software. The CAG repeats, CCG repeats, the intervening sequences and the flanking regions are seen in the mapped reads. The intervening sequence is: CAACAGCCGCCA and is located between the CAG and CCG repeats. All three different alleles have 7 CCG repeats on their chromosomes. Tablet was used to visualise the aligned reads for all the three alleles obtained from CLC genomics workbench.

We analysed the CAG allele frequency distribution for all individuals. To capture a wide distribution of CAG changes across different repeat lengths, we illustrated the read length distributions of HD CAG repeat lengths for six individuals with different CAG repeats (Figure 4-3). The mapped reads of HD alleles shows multiple reads were mapped against references of different lengths. For each allele, the CAG/CCG genotype was defined as the genotype of the reference against which the highest number of reads had aligned, which is the mode of CAG distribution. Both CAG and CCG repeat sizes were determined for all individuals. The read count distribution shows expanded CAG alleles at modes of 40, 45, 50, 54, 60 and 72 CAG repeats respectively (Figure 4-3). All the expanded alleles aligned to the 7 CCG reference. The highest peak is defined as the primary allele, which is expected to be the length allele, transmitted from

the affected parent. The aligned reads, which are to the left of the main allele, represent repeat unit losses that most likely occur due to backward PCR slippage. PCR slippage is biased toward deletion mutation resulting in the generation of products smaller than the template molecule (Pearson *et al.*, 2002). The proportion of those reads becomes greater with increasing allele size, i.e. the proportion is higher for 50 CAG repeat than in the 40 CAG repeats as shown in Figure 4-3.

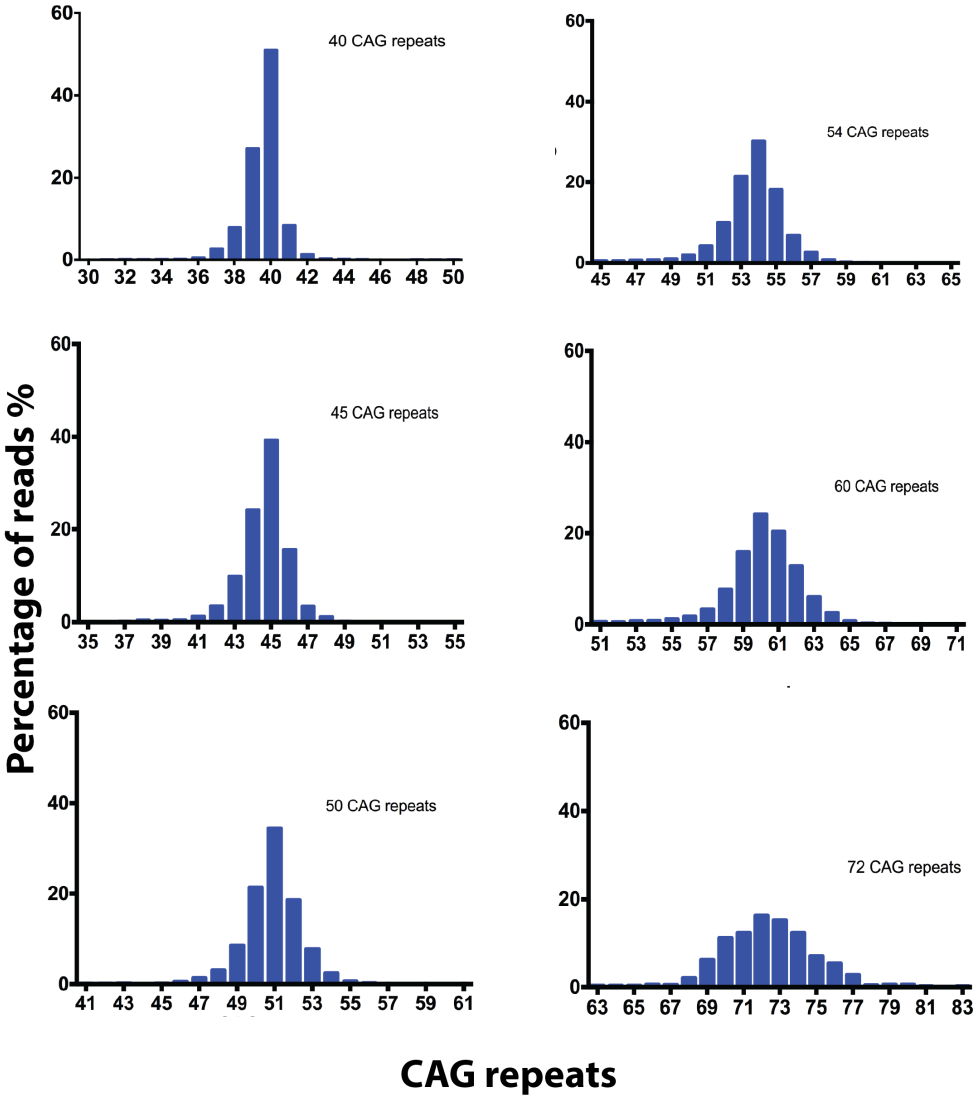

Figure 4-3 Allele length distributions for the CAG alleles in six different HD patients inheriting 40, 45, 50, 54, 60 and 72 CAG repeats obtained from MiSeq sequencing. Percentage of sequencing reads mapped (y-axis) to the reference sequences associated with different CAG repeats and 7 CCG reference (x-axis). All expanded alleles aligned to 7 CCG reference. The CAG genotype is defined as the genotype of the reference against which the highest number of reads had aligned. The main peaks for the allele read count distribution are 40, 45, 50, 54, 60 and 72 for each individual.

The peaks that are to the right of the main allele represent repeat unit gains captured by the MiSeq sequencing approach (Figure 4-3). Those reads are likely to derive from both forward PCR slippage and somatic instability. We think most of these reads are due to somatic instability because it is expansion-biased. As somatic mosaicism is allele length dependent and expansion-biased (Kennedy *et al.*, 2003; Veitch *et al.*, 2007), it is expected to result in a high proportion of reads larger than the progenitor allele in larger alleles. The proportion of those reads appears to be higher for larger alleles as shown in Figure 4-3. The repeat length distributions revealed the longer alleles had a higher proportion of PCR slippage and putative somatic mosaicism compared to the shorter alleles. The somatic mosaicism and PCR slippage will be discussed in detail in Chapter 5.

For the expanded CAG alleles (≤50 repeats), the main alleles are clearly recognized. We believe the mode of the CAG distribution is equal to the progenitor allele, which is a single length allele, inherited from the affected parent. For small expanded alleles (40 to 50 CAGs), the slope of the CAG distribution to the right of the main allele is steeper than the slope to the left. As the CAG alleles become >50 CAG, the slope to right of the main alleles becomes less steep, and for very large alleles (72 CAG), the distribution becomes symmetrical. Therefore, the progenitor alleles cannot be clearly identified in such alleles, due to PCR slippage and potentially somatic mosaicism (see Chapter 5 for more details about PCR slippage and somatic mosaicism). Therefore, we are not able to accurately genotype the expanded alleles with more than 50 CAG repeats because the reads associated with alleles longer than the inherited CAG repeats will affect the distribution of allele read length. Also because backwards slippage, which depends on allele length, may result in reads associated with alleles shorter than the inherited CAG repeats being more or less the same as the number of reads for the main allele.

In order to genotype expanded alleles (≤50 repeats), we examined the expected distribution of CAG repeats obtained from MiSeq sequencing to attempt to define the progenitor allele. The top graph represents the distribution of CAG repeats with peaks biased toward deletions (Figure 4-4). This generates a tail of products smaller than the starting molecules, and only a few reads from longer products as seen at the top distribution on Figure 4-4. This distribution is likely

generated by backward PCR slippage generating mostly repeats smaller than the starting molecules (green peaks to the left). Repeat slippage is common during PCR of repeats sequences and can result in backward slippage events of n-1, n-2, n-3, *etc*. The red peaks to the right of the main allele could be potentially somatic mosaicism captured by MiSeq sequencing, as the somatic instability of expanded alleles is expansion-biased. Therefore, the genotype of the expanded allele represented by this distribution would be the main peak, which represented as n (Figure 4-4).

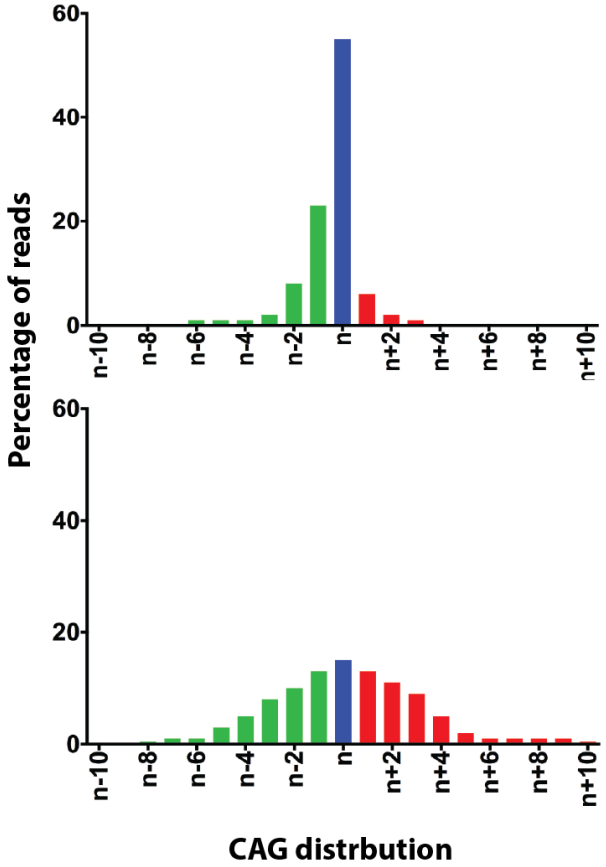

Figure 4-4 CAG repeat distribution in two different expected shapes obtained from MiSeq sequencing. The main allele is shown in blue (n), contractions (n-1 to n-10) in green, and expansions (n+1 to n+10) in red.

The bottom distribution on Figure 4-4 shows asymmetrical shape for the distribution of CAG repeats with many reads shorter and longer than the main allele. High levels of both somatic instability and PCR slippage result in this shape of the distribution, and no single clear peak is detected. This has been seen with very long alleles (>50 CAG) and that allele could be shifted to the right or left of the distribution. The peaks to the left of the main allele (n)

distribution are most likely generated by PCR slippage, generating products a multiple of repeats smaller than the starting molecule. This could shift the main allele to the left of the distribution because the longer the template, the more backward slippage. Also, peaks to the right of the main allele (n) are probably mostly generated due to somatic mosaicism. That could shift the main allele to the right of the distribution. Thus, we are not certain if the main allele is shifted to the left by slippage on the distribution or to the right by somatic instability, or whether the main allele remains as a major peak in the middle of the distribution, which is the mode. Although it remains unclear how to define the inherited progenitor allele, we decided to identify it for alleles with more than 50 CAG repeat also by the mode of the distribution.

4.2.3 CAG distribution in expanded HD alleles

We assessed CAG repeat length in the 409 affected individuals of HD families of the Venezuelan cohort (Figure 4-5). The CAG repeat size on expanded HD chromosomes ranged from 40 to 74 CAG repeats (mean = 45.56 ± 0.2 and median = 45), and the normal chromosomes from these patients ranged between 10 and 39 CAG repeats (mean = 21.2 and median = 18) (our results from Chapter 3). The most frequent alleles were with 45 (15.3%), 44 (11.5%) and 42 (11.5%) CAG repeats (Figure 4-5).

The distribution of the CAG repeats in HD patients revealed that the repeat range between 40 to 50 repeats was most common (382 alleles, 91.4%), with alleles greater than 50 CAG repeats being much less frequent (36 alleles, 8.6%). One 74 repeat allele was found which was the highest CAG repeat size in our study.

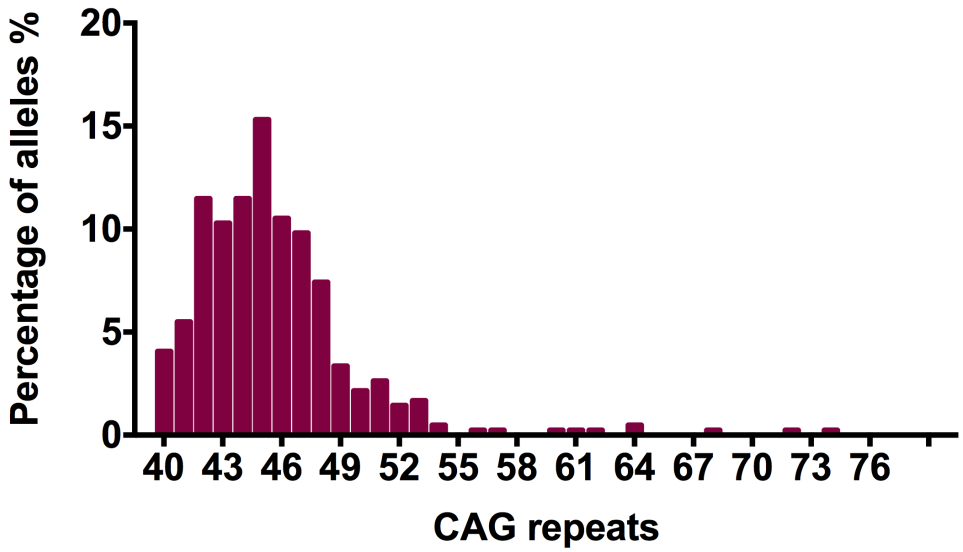

Figure 4-5 Distribution of expanded CAG alleles of the Huntington disease (HD) gene in 418 chromosomes from 409 HD patients of the Venezuelan population from MiSeq sequencing. The range of 40 and above CAG repeats is defined as a full penetrant range. The CAG repeat size on the expanded alleles ranged between 40 and 74 CAG repeats, with 45 CAG repeats being the most frequent.

Nine individuals (2.2%) were homozygous for the expanded alleles, having alleles containing 40 CAG repeats or more on both chromosomes. Previous studies on HD showed homozygosity rarely occurred and it ranged from 0.1 to 0.4%, which is lower than our obtained result (Kremer *et al.*, 1994; Alonso *et al.*, 2002). Of these, three of CAG homozygous individuals were homozygous for exactly the same number of CAG and CCG repeats. All three individuals were homozygous for 7 CCG repeats, and one of them has 44 CAG repeats, one has 45 CAG repeats and the last one has 48 CAG repeats. Therefore, those individuals may be truly homozygous because they had one of the most frequent alleles in the mutant range, although we cannot rule out the failure of amplifying the second allele. The remaining six individuals were homozygous for 7 CCG repeats and different CAG repeats in the mutant range, these individuals carrying CAG repeat numbers of 42/46, 44/47, 45/47, 47/49, 40/50 and 50/53 as their genotypes on both chromosomes. Those individuals are true homozygotes, as two expanded alleles were detected in our data. Because most of these CAG repeats are the most frequent, we are confident that the homozygous individuals are true homozygotes. Therefore, the CAG repeat distribution was obtained for 418 chromosomes from 409 HD patients.

4.2.4 Comparison between genotyping CAG repeats by MiSeq sequencing and fragment length analysis

To further investigate allelic variability in genotyping CAG repeats, we have now sequenced and genotyped the *HTT* repeat in 742 (333 unaffected (our result from Chapter 3) and 409 affected) individuals from the Venezuelan cohort. Among the 742 unaffected and affected individuals, 707 had also been genotyped for the number of CAG repeats by fragment length analysis from buccal samples. This was performed previously in the Lab, by Dr. Nicola Veitch (unpublished data). This allowed a direct comparison to be made between the numbers of CAG repeats estimated by MiSeq sequencing and by fragment length analysis for those individuals.

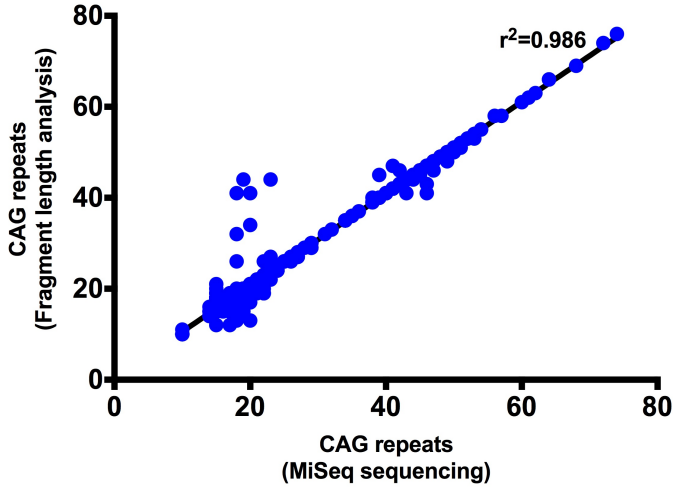

Figure 4-6 The correlation between CAG repeat number obtained by MiSeq sequencing and fragment length analysis for buccal DNA samples for 707 individuals (1,414 alleles).Blue circles represent the estimated CAG repeat by MiSeq sequencing compared to the number of repeats estimated by fragment length analysis that had been generated and measured previously Dr. Nicola Veitch. These alleles are from unaffected and affected individuals.

To compare the numbers of CAG repeats estimated by MiSeq sequencing and by fragment length analysis for buccal samples, we plotted the number of CAG repeats obtained by MiSeq sequencing and fragment length analysis for buccal samples for each individual (Figure 4-6).

Our results showed high similarity in determining the number of CAG repeats between MiSeq sequencing data and fragment length analysis for buccal samples (r^2=0.986) (Figure 4-6). The remaining variation (2%) is explained by differences

in estimation of CAG repeats by one or two repeats by fragment length analysis compared to MiSeq data. However, there are four alleles in the plot that are estimated to be expanded alleles from the fragment length analysis results and to be normal alleles from our MiSeq data (Figure 4-6). These outliers are overestimated of the number of CAG repeats by many repeats by fragment length analysis in comparison to MiSeq data. This could happen as result of human error in the laboratory such as labeling errors or swapped samples.

Also, the differences in our data compared to the fragment length method could be due to the presence of atypical allele structures within normal alleles that sometimes cannot be amplified using the PCR method. This can cause inaccurate determination of the number of CAG repeats. For the atypical alleles, it is likely that the intervening primer used for PCR genotyping method misprimes at more than one location, leading to an ambiguous length profile. Another explanation of the variation in estimation of expanded alleles in these two methods is that genotyping of large alleles is compromised by a high level of instability that might be hard to detect using fragment length analysis.

Fragment length analysis approaches may fail to identify atypical sequence variants or flanking sequence differences, and may result in inaccurate genotyping of expanded alleles due to the high level of instability. However, using NGS sequencing approach by MiSeq alleles has a high-throughput genotyping of CAG repeats, and has the potential to detect insertions, deletions or other sequence variants within or flanking the repeat.

4.2.5 Analysis of CCG repeats in expanded HD alleles.

The size of CCG repeats in the *HTT* gene and the frequency of each allele were analysed in HD chromosomes. One CCG allele was identified from sequencing and genotyping HD alleles of the HD patients. CCG 7 is the only detected CCG allele in 409 individuals. In contrast, we found different CCG alleles from genotyping the normal *HTT* chromosomes in the Scottish and Venezuelan populations (Chapter 3). The CCG alleles in the normal chromosomes from these patients were 6, 7, 9, 10 and 11 CCG repeats (Chapter 3). Analysis of CCG repeats haplotypes revealed that the expanded alleles presented only one haplotype with the 7 CCG allele.

4.2.6 The correlation between the CAG repeats on the normal and expanded alleles

We tested whether subjects with expanded CAG repeat lengths had a different range of normal CAG repeat length. We analysed the relationship of the number of CAG repeats between the normal and mutant alleles (40 to 53 CAG repeats) from heterozygous HD patients in the Venezuelan population (Table 4-1). We excluded the homozygous cases in this analysis as they have expanded alleles on both chromosomes. CAG repeats of greater than 53 CAG repeats were not included in the analysis, as they are less frequent alleles, therefore we can not detect the association between these alleles.

Expanded CAG allele number	Normal CAG alleles range	No. of subjects	Mean of normal CAG alleles	Median of normal CAG alleles
40	15-22	16	18.63	18.5
41	15-24	22	18.59	18
42	15-26	47	17.81	17
43	15-29	43	18.14	17
44	10-39	45	19.16	18
45	10-34	60	17.6	17
46	15-29	43	18.95	18
47	14-29	39	18.54	17
48	15-24	29	17.86	17
49	15-27	13	18.38	18
50	15-25	7	20.29	21
51	15-22	11	17.73	18
52	18-22	6	19.67	19
53	17-22	6	19	18.5

Table 4-1 The expanded allele associated with the normal allele for each individual from heterozygous HD patients from Venezuela.

There was no obvious correlation between the range of normal allele sizes and the expanded allele (Table 4-1). The mean and median were determined for each normal CAG repeat range for HD patients carrying each expanded CAG repeat. The mean and median of normal alleles were similar for all expanded alleles from 40 to 53 CAG repeats. Figure 4-7 shows the CAG repeat size in the

normal and expanded HD alleles for all heterozygous HD patients of the Venezuelan population. There were no clear differences for the normal allele ranges associated with each expanded CAG allele.

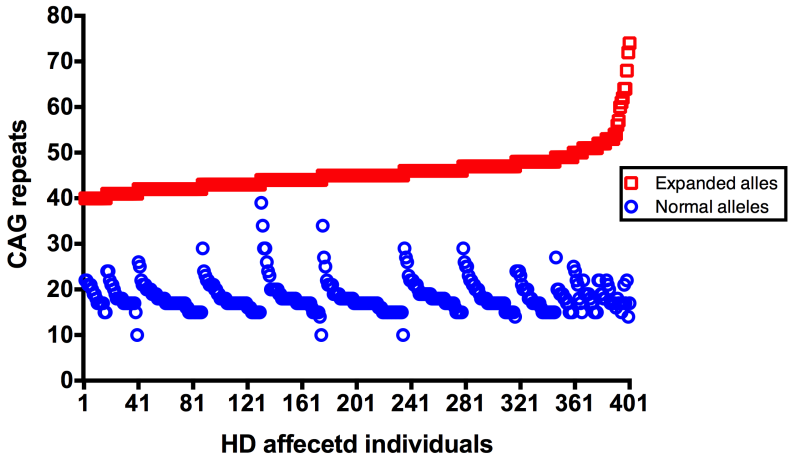

Figure 4-7 The normal and expanded alleles sizes in 400 heterozygous HD subjects from the Venezuelan population. The expanded alleles represented in red squares and the normal alleles represented in blue circles. Expanded allele ranged between 40 to 74 CAG repeats and the normal allele ranged between 10-39 CAG repeats.

4.2.7 CAG repeat length in the mutant HD allele and sex

We had assessed the effect of sex on the mutant CAG repeat sizes of affected individuals (Figure 4-8). Among 409 individuals, 173 were males, 225 were females and the remaining individuals (11) were of unknown sex from the Venezuelan population. The analysis includes 398 individuals of known sex. The CAG repeat size was from 40 to 68 in male patients (mean= 45.4, median =45) and 40 to 74 CAG repeats in female patients (mean = 45.7, median = 45). We found the CAG repeat size was not correlated with the sex of the patients since the mean and median of CAG repeats for both sexes were similar even though the expanded CAG repeat range was larger for female patients (Figure 4-8). Our results demonstrated no difference in the CAG repeat size between sexes.

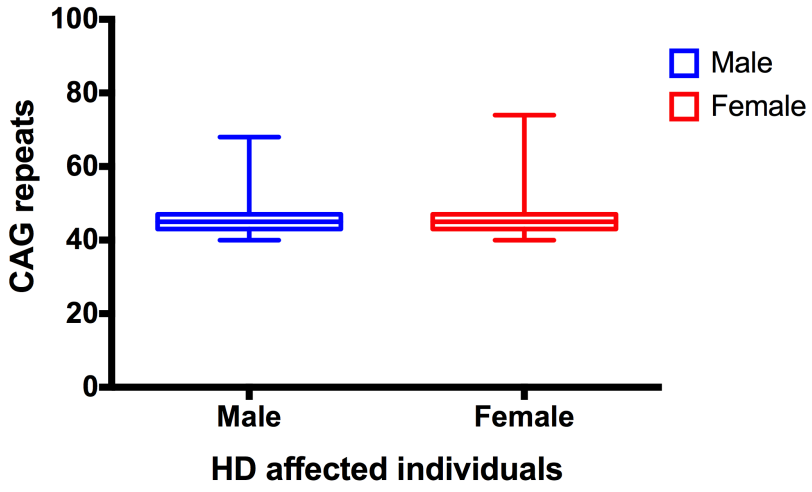

Figure 4-8 The CAG repeat size effect for the sex of 398 affected individuals from the Venezuelan population. The expanded CAG size ranged from 40 to 68 in male patients and from 40 to 74 in female patients. The CAG repeat size medians are 45 for both sexes and the means are 45.40 and 45.67 for male and females, respectively.

4.2.8 CAG repeat length and phenotype

The expanded CAG repeat is the major modifier of the age at onset in HD. Although the dominant inheritance of HD implies that one expanded allele is a trigger for the HD pathogenesis (Lee *et al.*, 2012), normal CAG alleles have been suggested in few studies, but not all, to modify the onset of the disease (Djoussé *et al.*, 2003; Li *et al.*, 2003). We therefore looked at the correlation between the phenotype and genotype of HD, particularly on the effect of the length of the expanded and normal CAG repeats.

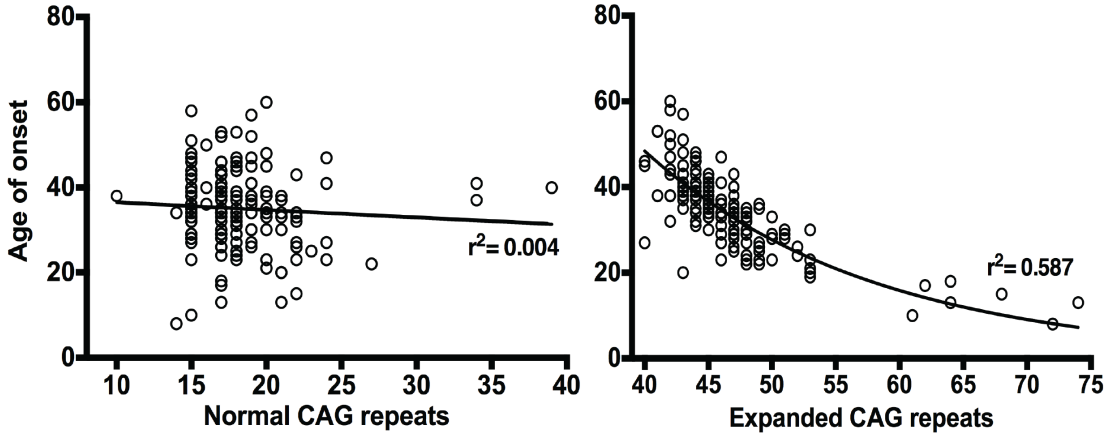

Figure 4-9 The relationship between the age at disease onset is shown for both the normal and expanded CAG alleles in 169 Venezuelan HD patients. The graph on the left shows no correlation between the normal CAG alleles and age at onset in HD patients (r^2=0.004, P=0.394). The graph on the right shows the inverse correlation between the expanded CAG alleles and age at onset (r^2=0.587, P < 0.001). This graph shows exponential regression lines fitted to the age at onset and the expanded CAG repeat length in HD.

We analysed 169 patients with HD with known age of disease onset. The relationship between age at onset and the CAG repeat size in HD alleles in Venezuelan affected subjects is shown in Figure 4-9. The age of onset of these HD subjects ranges from 8 to 60 years (median =34.5). This result showed that CAG repeat size is the main determinant of age at onset in HD patients, accounting for 58.7% of the variation in age at onset in the non-linear regression correlation. As expected, there was an inverse correlation between the age at onset of symptoms and the number of CAG repeats. An exponential regression model yielded adjusted R-squared values of 0.587 for the correlation between CAG repeats and age at onset, indicating that there are factors other than the expanded CAG allele length that modifies age at onset and disease pathogenesis in HD. The wide range of onset ages associated with any given repeat length points also to the possible influence of other genetic or environmental factors. Therefore, this variation in age at onset precludes an accurate estimate of when the disease will manifest for any single individual.

We also tested whether the normal CAG alleles had an effect on the age of disease onset in HD patients. Our results showed no significant effect of the normal CAG allele on age at onset (r^2=0.004, P=0.394) (Figure 4-9). Thus, our finding indicates the expanded CAG allele length is the most important factor in

modifying age at onset of HD and that normal CAG allele length does not play a significant role in the pathogenic process of the disease onset.

4.2.8.1 Phenotype-Genotype correlation of homozygous HD and also the effect of sex

We identified 9 homozygous individuals, indicating a CAG repeat expansion on both alleles, as mentioned earlier. Homozygosity has not been reported to influence the age at onset compared with heterozygosity, but there was an impact of HD homozygosity on disease progression and severity (Alonso *et al.*, 2002; Squitieri *et al.*, 2003).

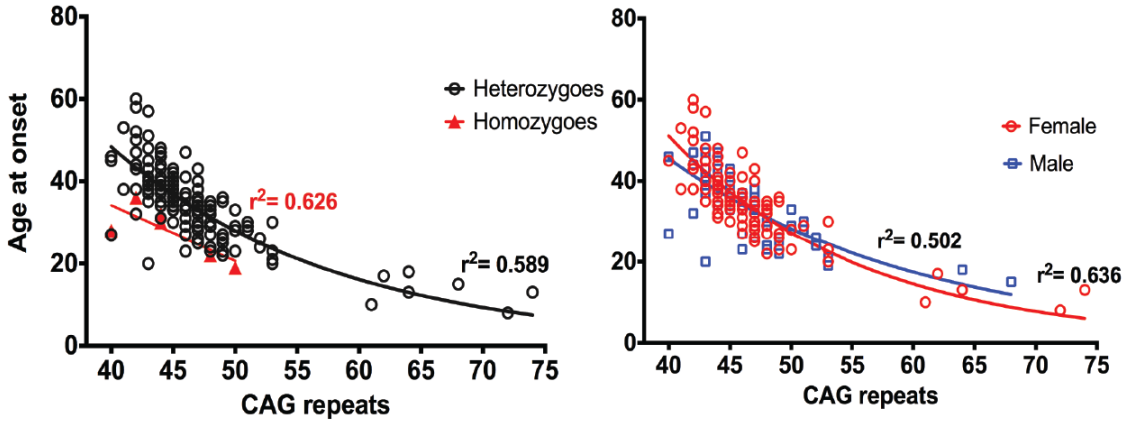

Figure 4-10 The relationship between age at onset in 169 individuals with HD and expanded CAG alleles in homozygous and heterozygous cases and also the effect of sex. The graph on the left shows the longer expanded alleles for HD cases with homozygous individuals (with 2 expanded alleles >40 CAG repeats) as red triangles and heterozygous individuals as black circles. The exponential regression lines fitted to the age at onset and CAG repeat length for the heterozygous individuals (r^2= 0.589, P<0.001) and for the homozygous individuals: the linear regression fitted to the age at onset and CAG repeat length (r^2= 0.626, P= 0.061). The graph on the right shows the exponential regression between the CAG repeat length and age at onset with adjusted R squared 0.502 and 0.636 for males and females respectively. The blue squares represent males and red circles represent female HD patients.

We compared the homozygous subjects' age at onset, as the only clinical feature available for our data, with 163 heterozygous cases. The age at onset was not determined for some of the homozygous cases. Therefore, 6 individuals were included in the study for whom we have the clinical age at onset. For the homozygous individuals, we plotted the longer CAG allele length against the age at onset. We found the correlation between the CAG repeat and age of disease onset in the heterozygotes cases (r^2=0.589, P<0.001) for an exponential

regression model. A linear regression fitted the correlation between the CAG repeat and age at onset for the homozygous individuals, giving an adjusted r^2 of 0.626 and P value of 0.061, which is a marginally significant relation. The mean distribution of age at onset was different for heterozygous and homozygous individuals (34.83 and 27.83 respectively). Student's t-test was used to compare mean age at onset between the homozygous and heterozygous individuals (using SPSS). The result shows a significant difference between the mean age at onset between homozygous and heterozygous individuals (t-test, t=1.908, P= 0.029).

Also, to evaluate the relationship between the age at onset, and the inherited allele length for the homozygous and heterozygous individuals, we used a linear regression model using SPSS statistics software (IBM). We found that the homo/heterozygous cases were significantly correlated with age at onset (P=0.039) (model 1, Table 4-2). Also, the t-test result was significant between the mean age at onset for homozygous and heterozygous cases. Thus, our result determined the homozygous individuals have an earlier age at onset and subsequently they are more severely affected than heterozygous individuals.

Model	Adjusted R square	P value	Parameters	Coefficients	Stanadrd error	t-statistics	P value
Model 1: Age at onset = Allele length + Homo/heterzygotes cases	0.561	<0.001	Intercept		4.821	20.844	<0.001
			Allele length	-0.739	0.09	-14.401	<0.001
			Homo/ Heterozygotes	-0.107	2.454	-2.085	0.039
Model 2: Age at onset = Allele length + Sex	0.555	<0.001	Intercept		4.515	21.004	<0.001
			Allele length	-0.747	0.091	-14.562	<0.001
			Sex	0.064	0.95	1.254	0.212

Table 4-2 Regression model of the relationship between the age at onset, and the inherited allele length for the homozygous and heterozygous cases in 169 individuals and also for the relationship between the age at onset and the inherited allele length for male and female cases. Allele length, homo/heterozygous cases and sex of the patients were used as independent variables in the regression models using SPSS statistics software (IBM). The table shows the adjusted squared coefficient of correlation (adjusted r^2) and statistical significance (*P*), the coefficient, standard error, t-statistic and statistical significance (*P*) associated with each parameter in each model.

The graph in Figure 4-10 illustrates the correlation between the expanded CAG repeats and age at onset between sexes for HD individuals; including 105 females and 64 males. There were many previous studies investigating the difference in age at onset of HD between sexes in different populations

(Telenius *et al.*, 1993; Kremer *et al.*, 1995). We sought to examine this correlation in our data for the Venezuelan population. The correlation between the age at onset and expanded CAG repeats in male was r^2=0.502 and 0.636 among female. There was a slight difference in the correlation between male and female. However, the mean distribution of age at onset was similar for male and female (34.23 and 35.07 respectively). The mean age at onset was not significantly different between male and female (t-test, t= 0.583, P=0.28). Therefore, our result revealed no sex effect for the age at onset in HD patients.

We also investigated the sex-dependent effect of the relationship between age at onset and inherited allele length using a linear regression model. These analyses revealed a non-significant difference (P =0.212) between males and females in a phenotype-genotype correlation (model 2, Table 4-2). From these two models, we can establish that the age at onset is dependent on the allele length and there were differences in the phenotype-genotype correlation between homozygotes and heterozygotes individuals. However, there were no differences in the phenotype-genotype correlation in patients of either sex.

4.3 Discussion

In this chapter, we described genotyping of the expanded CAG repeat in HD by next generation sequencing as a new sequence based approach. We have established that it was possible to sequence and genotype normal HD CAG alleles, including the polymorphic CCG repeats and the flanking sequences, using the MiSeq platform by amplifying the region using locus-specific primers combined with MiSeq sequencing adapters (Chapter 3). We have expanded this to show we are also able to genotype the expanded CAG allele including the polymorphic CCG repeats and the flanking sequences using MiSeq platform.

Using this approach we sequenced 409 buccal swab DNAs from the US-Venezuelan Collaborative Research Project consisting of affected individuals from HD families from Venezuela. The libraries were produced using the MiSeq compatible PCR primers with the locus-specific primers using protocols previously optimised (Chapter 3). Those libraries were sequenced using a standard 2x300 bp run or 400/200 bp parameters in both forward and reverse directions. During the library preparation, each individual was associated with a

unique barcode combination included in the sequencing adapters incorporated in the PCR primers. The sequencing reads obtained were then mapped with CLC Genomics Workbench to reference sequences including various numbers of CAG and CCG repeats (1 to 200 CAG repeats and 1 to 20 CCG repeats) and the region flanking the repeats.

These data have revealed accurate genotypes for the expanded CAG repeat and flanking regions for most individuals. Good quality sequencing reads were obtained for both the normal and the expanded allele for each individual. *HTT* CAG/CCG genotypes were obtained for 409 affected individuals from the Venezuelan population. From this data, we were able to sequence CAG repeats up to 74 repeats with the presence of the normal alleles for most of the samples. However, we noticed that we were unable to reveal the 3'-flanking region from the mapped reads for patients with 74 CAG repeats. Thus, the reverse read can be used to genotype the CCG repeat and the 3'-flanking DNA sequence to detect any variants in that region and for the haplotype analysis. Sequencing longer alleles can be done by genotyping the CAG repeats from the forward strand and the CCG repeats from the reverse strand because we cannot match or merge the reads in the middle of the repeats. We also investigated the utility of reverse reads and that these can be used to improve the efficiency of genotyping the CCG repeat and the 3'-flanking DNA sequence.

We analysed the distribution of the expanded CAG allele across different CAG repeats. The read length distribution for MiSeq sequencing was used in the analysis to examine the wide distribution of CAG changes across different repeat lengths. We have shown in this chapter the CAG read length distributions for 6 individuals, with expanded CAG alleles at modes of 40, 45, 50, 54, 62 and 74 repeat. The data revealed the repeat length distribution is more variable for the longer alleles compared than the shorter alleles. Overall, longer alleles had a higher proportion of (n-1 and n+1) reads that are shorter and larger than the main peaks when compared to shorter repeat length alleles. It was noticeable that the progenitor alleles of CAG repeats (>50 CAG) have progenitor alleles that cannot be clearly identified due to PCR slippage and potentially somatic mosaicism (see Chapter 5 for more details about PCR slippage and somatic mosaicism). Thus, we are not confident about the genotypes of the expanded

alleles with more than 50 CAG repeats. The read length distributions of those alleles were associated with alleles longer than the inherited CAG repeats which affects the distribution of allele reads. Although it remains unclear how to define the inherited progenitor allele for larger alleles with more than 50 CAG repeats, we decided to identify the progenitor allele as the mode of the CAG repeat distribution.

Accurate determination of CAG repeat length is critical in genetic counselling for affected individuals, as it makes a difference in predicting the average age at onset. Also, determining the CAG repeat size accurately can allow us to differentiate whether that individual is in the affected range, low penetrance, or not at risk for HD.

The CAG size distribution of these HD patients' chromosomes ranges from 40 to 74 CAG repeats. The most frequent alleles had 45 repeats. The CAG repeat range between 40 to 50 repeats was most common, comprising 91.4% of the total alleles, with alleles greater than 50 CAG repeats being less frequent (8.6%). The distribution of CAG repeats among HD cases did not differ significantly from previously published data with the same population (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004). The number of CAG repeats in HD samples ranged from 40 to 86 repeats in 956 affected individuals from the Venezuelan population, with 44 CAG repeats being the most frequent (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004). Expansions between 40 to 50 repeats were most common, representing 90% of the expanded alleles. This is similar to our findings.

Comparing our obtained results from genotyping CAG repeats from MiSeq sequencing and fragment length analysis genotypes for the same samples showed there were similarities between both methods (r^2=0.986). However, the slight differences could be due to a failure of fragment length analysis in identifying atypical sequence variants, flanking sequence differences, inaccurate genotyping of expanded alleles due to a high level of instability or human error. Therefore, an NGS sequencing approach using MiSeq allows a high-throughput genotyping of CAG repeats and has the potential to detect insertions, deletions or other sequence variants within or flanking the repeat.

All expanded alleles were associated with 7 CCG repeats in the Venezuelan population, which reveals the CCG repeat is stably transmitted. The lack of variation in CCG repeats supports a possible common source of mutation in Venezuela due to a founder effect inheriting 7 CCG repeats on the expanded alleles. In contrast, the study of the CCG repeats in other ethnic populations, revealed each ethnic population shows a predominance of certain CCG alleles that can vary between 6 to 12 CCG repeats, with CCG 7 and 10 being the most frequent. For example, in Japanese and Chinese populations, the expanded alleles were associated with 10 CCG repeats (Morovvati *et al.*, 2008). The CAG repeat number in Chinese normal allele was higher in 10 CCG than in 7 CCG alleles. In contrast, the expanded alleles in the Western population were associated with 7 CCG, although the 10 CCG was present in some pathogenic chromosomes (Barron *et al.*, 1994; Squitieri *et al.*, 1994; Costa *et al.*, 2006). This suggested that their mutations originated from a different origin. The differences in CCG alleles in different populations could be related to the differences between populations, in which other modifiers and the haplotype may be involved in the expanded chromosomes.

All detected expanded HD alleles retained the expected structure of CAG and CCG repeat: $(CAG)_n$ CAACAGCCGCCA$(CCG)_n$. Our analysis revealed no atypical alleles identified from sequencing of Venezuelan HD patients. In contrast, analysis of normal *HTT* alleles from unaffected individuals from the Scottish and Venezuelan populations revealed atypical allele haplotypes in the normal alleles (Chapter 3). Also, some atypical expanded alleles have been detected previously. These atypical alleles differ structurally mainly within the intervening sequence, but also can differ in the number of CCT repeats present downstream of the CCG repeats. The expanded alleles sequenced here were associated with typical haplotype could be due to a founder effect in the Venezuelan population of Lake Maracaibo, so the mutation has a common ethnic descent with HD families.

In HD, we have seen atypical normal alleles (Chapter 3) and no atypical expanded alleles. Dr. Marc Ciosi has seen mutant atypical alleles, but they are already described in normal alleles (unpublished data). This suggests the atypical expanded alleles in HD arise from the general population. However,

there were no atypical alleles in the general population for DM1 and the atypical alleles, interrupted with variant repeats, seen in DM1 affected families (Braida *et al.*, 2010). Thus, the atypical expanded alleles in DM1 may arise from *de novo* mutations in DM1 patients.

We analysed the relationship of the number of CAG repeats between the normal and mutant alleles. There was no correlation between the normal allele length and the expanded CAG allele. The mean and median of normal alleles were slightly similar for all expanded alleles from 40 to 53 CAG repeats. Thus, our findings showed the length of the CAG repeat on the normal allele does not influence the expanded CAG alleles.

Also, we investigated the effect of sex on mutant CAG repeat sizes of affected individuals from the Venezuelan population. There is a slight difference in the correlation test between male and female. However, the mean and median of CAG repeat were similar for male and female. We did not observe any differences in correlation between the CAG repeat size and the sex of the patients, even though the expanded CAG repeat range was larger for female patients. Our result obtained similar mean and median of CAG repeats for both sexes, even though the expanded CAG repeat range was larger for female patients. Our results demonstrated no difference in the CAG repeat size between sexes.

We analysed the relationship between the number of CAG repeats within the normal and mutant alleles and the age at onset for 169 patients with clinically known age at onset. Within these patients, the number of CAG repeats of the normal allele was found uncorrelated to the age of disease onset. However, we demonstrated an inverse correlation between the expanded CAG allele length and age at onset of HD. The regression model showed that 69% of the variation of the age at onset was accounted for by the number of CAG repeats of the expanded alleles. Our analysis confirmed there was no evidence that normal allele length modifies the age at HD onset, and that rather length of the expanded allele is the major factor in modifying the age at onset in HD. The remaining variation in age at onset could be generated by other genetic or environmental factors.

Furthermore, we analysed the association of CAG repeat length of homozygous and heterozygous patients as well as the different sexes effect on age at onset. We found there were no differences in the genotype-phenotype relationship in patients of either sex. Interestingly, our results demonstrated the homozygous cases have an influence on age of disease onset. These homozygous individuals have an earlier age at onset than heterozygous individuals. Therefore, the homozygosity for HD is expected to result in a more severe phenotype and rapid disease progression.

This finding is different to previously published studies on the effect of homozygosity on age at disease onset. An earlier study reported a similar age at onset for homozygous and heterozygous individuals, but homozygotes had a more severe phenotype of HD (Squitieri *et al.*, 2003). These different findings could be because this study was conducted on a small number of HD patients (eight homozygotes and 75 heterozygotes). In addition, it may be difficult to determine the accurate age at onset in HD due to an overlap of cognitive, motor and psychiatric symptoms in the early stages of HD. Our data suggest that homozygous individuals have an earlier age at onset of disease than heterozygotes.

We noticed that the frequency of homozygotes (2.2%) for HD is higher in the Venezuelan family compared to previous studies in Caucasian populations. Previous studies on HD showed homozygosity rarely occurred, and it ranged from 0.1 to 0.4%, which is lower than our obtained result (Kremer *et al.*, 1994; Alonso *et al.*, 2002). This is most likely because of the high frequency of the HD mutation in the Venezuelan cohort. In addition, studies in other polyglutamine expansion disorders, such as the spinocerebellar ataxias (SCA3 and SCA6) and dentaterubro-pallidoluysian atrophy (DRPLA) have reported a more severe phenotype in homozygotes than in heterozygotes, with an unknown reason for these differences (Gusella and MacDonald, 2000).

Patients with homozygous HD mutations are an important aspect for genetic counselling, given the fact that all their offspring will inherit the HD mutation and will be affected by the disease, unless a contraction of the CAG repeat tract occurs, which is highly unlikely. Therefore, additional counselling is very

important, as it is essential for the patients to know about the increased risk of transmitting the mutation to their offspring.

Our results indicate the expanded CAG allele length is the primary determinant of age at onset and the repeat length is inversely correlated with age at onset. We also observed that one unit of CAG repeat increase causes a decrease in age at onset of approximately 2 years. This is similar to the previously published data (Goldberg *et al.*, 1993; Gusella and MacDonald, 2009). The broad range of age at onset associated with each expanded allele size could be due to other genetic or environmental modifying factors. The importance of defining these factors is that they may be targets for delaying the onset of disease and slowing the progression in HD patients. Moreover, the identification of CAG repeat length is not a certain predictor of onset age, due to variation in age at onset for each CAG repeat size.

In summary, in addition to genotyping *HTT* alleles by NGS technology, our studies in the phenotype-genotype emphasize the importance of the expanded allele for the phenotype in HD patients.

Chapter 5 Somatic mosaicism of expanded CAG repeats in HD patients of the Venezuelan cohort: a modifier of disease severity

5.1 Introduction

The length of the expanded HD alleles in the germ-line is unstable through intergenerational transmission, with a major bias toward an increase in size in paternal transmission (Duyao *et al.*, 1993; Snell *et al.*, 1993; Goldberg *et al.*, 1995; Falush *et al.*, 2000). This causes the anticipation phenomenon, where the age of onset decreases and the severity of the disease increases in successive generations (Nahhas *et al.*, 2005). Most juvenile HD cases are reported in patients with paternal inheritance, which is characterized by inheritance of very large CAG repeats and early age at onset.

A number of studies have shown that there is a variation in repeat size within and between tissues of affected individuals with a higher level of somatic instability observed in affected brain regions relative to cerebellum and blood in HD patients (Telenius et al. 1994; De Rooij et al. 1995). The sensitive small pool PCR (SP-PCR) technique has been used to analyse the mutation length profile from different brain regions in HD cases (Kennedy *et al.*, 2003). Brain samples from two individuals with Vonsattel grade 0, which means no neuropathological changes were identified in the striatum (Vonsattel *et al.*, 1985), were examined. According to the Vonsattel grading system HD autopsy brains were classified into five grades (0 to 4) according to neuropathological severity. Each grade reflects the extent of neuronal loss by macroscopic and microscopic examination of the striatum. These two individuals are likely to be pre-symptomatic, as they both died (~ 6 and 13 years) earlier than the expected age at onset of neurologic symptoms. The mutation length profiles showed differences in mutational instability in each specific brain region and mutation length changes were expansion-biased. Interestingly, the greatest instability was observed in these two individuals brains in the region most affected in HD patients, the striatum. The same pattern was also seen in a knock-in mouse model (where expanded CAG repeats were inserted into the mouse Htt locus) in which a higher mutation

level was observed in striatum compared to cortex and cerebellum (Kennedy and Shelbourne, 2000). These somatic expansions are also quite remarkable in their size in HD patients. Striatal cells may acquire expansions that exceed several hundred or even a thousand repeats in length, which reflects up to a 20 times length increase in comparison to the allele size inherited from the affected parent (Kennedy *et al.*, 2003). These data have shown that allele length expansion in the striatum, the most affected brain region; occurs prior to the HD pathogenesis and before the onset of symptoms.

Interestingly, although such very large expansions were observed in these individuals with no detectable pathology, there is significant neuropathology seen in individuals with end-stage HD disease (Kennedy *et al.*, 2003). Brain samples were examined by SP-PCR for an end-stage HD patient who died 10 years after disease onset. Although there was no formal nueropathological grading of this sample according to the Vonsattel's HD grading system, there was notable striatal atrophy and cell death in the autopsy brain specimen. A high level of mutation instability was present in the cortex, but not in the striatum, which is expected to exhibit the largest level of instability. This observation suggests that the cells with the largest repeat expansions in the striatum might be lost during the disease process. In this study, the evaluation of somatic mosaicism from a small number of samples makes it difficult to draw clear conclusions about the differences in mutation length changes in each HD disease stage, given that the greatest expansions are associated with specific target tissues, and this somatic instability occurs prior to the HD pathogenesis. These findings have led to the hypothesis that the somatic instability of the CAG repeat may contribute to the tissue-specific pathology and the progressive nature of the disease.

The report by Kenndey et al., 2003 was very interesting showing there were long expansions in striatal tissue; however those analyses were based on bulk DNA, and it was not clear whether those large expansions were observed in the neurons or in the supporting glial cells. In order to address that question, Shelbourne et al. 2007 used laser capture microdissection to selectively choose either neurons or glial cells based on cell surface markers and assess the repeat lengths in those specific cell types (Shelbourne *et al.*, 2007). They were able to

show that indeed repeat length expansions were greater in neurons than they were in glial cells from two affected individuals with advanced stage disease. These findings are very important and show that repeat length expansions are able to accumulate in non-dividing cells and the repeat lengths are greater in neurons. This is consistent with the hypothesis that neurons die because they have bigger repeats. Laser capture microdissection studies on mutation length also show that striatal neurons tend to have the longest expansion compared to those from cortex in two individuals with early-stage disease. This finding is consistent with the idea that CAG repeat lengths in striatal neurons increase early during the disease process. This evidence supports the hypothesis that somatic increases of repeat length might contribute to cell type differences in vulnerability in HD. Their findings also demonstrated that somatic instability might influence disease progression.

To determine the association between somatic instability in HD patients and variation in the age of disease onset in HD, SP-PCR was used to quantify the CAG repeats at the single molecule level in cortex samples from postmortem brain (Swami *et al.*, 2009). Selecting the cortex over the striatum in this study to measure somatic instability in HD was because the striatum is more vulnerable to damage and subject to neuronal loss in HD, thus cannot be easily examined in autopsy brain (Kennedy *et al.*, 2003). The study involved an analysis of 48 patient cortex brain regions, with 24 cases of extreme early onset and 24 with extreme late-onset (Swami *et al.*, 2009). Those patients developed symptoms at either an earlier or later age than expected age at onset relative to the number of repeats inherited. They observed the level of somatic instability of CAG repeat varies between all individual's cortex samples, with a significant bias towards expansions. As expected, a greater degree of somatic mosaicism in HD brain was found in individuals with longer CAG repeat length.

To test whether there was an association between variation in age at onset and somatic instability, the magnitude and frequency of CAG repeat expansions for the cortex samples were determined for the extreme early and extreme late onset groups (Swami *et al.*, 2009). The results showed there were no differences in the mean CAG expansion, the total expansion frequency, or the frequency of expansion of at least 10 CAGs between the groups. However, they noted that the

largest expansions were found in individuals with extreme early onset. To evaluate if the greater somatic expansions were associated with earlier disease onset, they measured skewness from the distribution of mutant CAG alleles in the cortex. Skewness is a measurement of the symmetry of the distribution of expanded CAG repeat lengths. The overall distributions of repeat length changes were skewed to the right, as most repeat length changes were expansions. Thus, samples that exhibit greater somatic expansion would show greater positive skewness than those with smaller somatic expansions. The result showed a negative correlation between residual variation in age at onset and skewness. This evidence showed that somatic mosaicism of the mutant allele in the brain is significantly associated with age of disease onset after accounting for the effect of CAG repeat length. These data thus support the hypothesis that somatic instability contributes to the pathogenic process, with higher somatic instability directly linked to disease progression and earlier age at onset.

Although detailed studies have been carried out on somatic mosaicism in the affected brain regions, it is difficult to further determine the possible contribution of somatic instability to disease age at onset and the nature of HD pathogenesis in human autopsy brain material. Brain tissues cannot be evaluated until death, and it is difficult to analyse the most affected tissues because many cells have already been lost during the disease process. Obviously, brain samples cannot be obtained from living HD individuals at the time of disease onset. Therefore, other tissues that allow us to test a larger number of samples in living individuals are preferable to use for measuring somatic instability.

Few studies on human peripheral tissues have been carried out to study HD instability. It has been previously reported the blood DNA showed a relatively low level of somatic mosaicism (Telenius *et al.*, 1994; Leeflang *et al.*, 1995). Although buccal cell DNA showed a low level of repeat length variation and small allele length changes in affected individuals, somatic mosaicism is measurable in buccal cell DNA (Veitch *et al.*, 2007). A study of the somatic instability of HD was carried out in buccal cell swabs from HD patients from Venezuela collected by The US-Venezuela Collaborative Research Project team. First, they determined the level of somatic mosaicism using SP-PCR within 20 individuals, who had inherited CAG repeat lengths between 40 and 66 repeats. These data

showed the level of somatic mosaicism is low in most cases in buccal cells, except in the very longest HD alleles (~75 CAG repeats), which have a higher level of somatic instability. Given that most affected individuals inherit the classic range of HD (40 to 50 CAG repeats) and a relatively small number of HD patients (16%) inherit long alleles (Gusella and MacDonald, 2000), by using SP-PCR and agarose gel-based methods we cannot quantify somatic instability in the majority of cases with HD. Therefore, a sensitive single molecule PCR assay was established to investigate mutation length changes in buccal swab cells, which allows for accurate sizing of a single repeat difference in a polyacrylamide gel (Veitch *et al.*, 2007). Samples from 12 patients who were all 39 years old at the time of sampling, but with variable CAG repeats (39 to 48 repeats), were examined. Individuals were matched for age at sampling to ensure the effects of age did not confound the association between somatic instability and allele length. For each individual, a DNA sample was diluted 50 times to produce an average of 0.5 molecules of the template in 10 µl reaction and the sample was amplified by PCR. After that, PCR products were resolved by agarose gel electrophoresis to detect the presence of expanded alleles. This step was necessary as agarose gel electrophoresis is more cost-effective than polyacrylamide gel electrophoresis. The analyses of the polyacrylamide gel electrophoresis use fluorescently labelled primers to visualise the DNA in the gel and automated DNA sequencing machine that makes the procedure more expensive, time-consuming and requires special facilities and makes this protocol impractical for most laboratories. Then, the size of the expanded alleles was precisely determined by using either the ABI polyacrylamide gel or capillary electrophoresis automated sequencing system. The somatic mutation frequencies detected were generally low in buccal cells, but with a varying degree of mutation between individuals. Although the level of somatic mosaicism is low in buccal cells and patients may acquire very large expansions in brain tissues, this data confirmed that the inherited allele length is the major modifier of somatic instability and accounts for ~70% of the variation observed in individuals with identical age at sampling. Using single-molecule PCR is sensitive, but is impractical and labour intensive for high throughput analysis in a large cohort of patients. To overcome the limitation of previous methods, a novel method should be developed to facilitate high throughput analysis of somatic instability.

Studies on HD strongly implicate somatic instability in disease pathogenesis and the tissue specificity of the disease. Thus, understanding the mechanisms and role of genetic instability remains an important subject in HD research and may provide a potential route to therapy. Previously (in Chapters three and four), it was demonstrated that using high throughput next generation sequencing (NGS), we were able to derive accurate somatic genotypes of normal and expanded alleles in addition to the repeat structures from buccal DNA samples. The aim of the current chapter therefore, was to quantify somatic mosaicism of the CAG repeats in HD directly from NGS read length distributions. We hypothesise that the main drivers for somatic mosaicism are repeat length and age and that individual-specific residual variation for both effects will also be associated with disease severity. In this chapter, we will address the issues associated with NGS technology as an approach to quantify somatic mosaicism of the CAG repeat in HD. Thus, we sought to measure the degree of somatic variation in buccal cell DNA of large numbers of HD patients by quantifying the effects of allele length and age at sampling in generating somatic mosaicism, and to test whether somatic mosaicism contributes to disease severity.

5.2 Results

5.2.1 Suitable statistical measures for quantifying somatic mosaicism and the role of allele length and age in defining instability.

In order to determine the utility of an NGS based approach in being able to define somatic mosaicism, we used MiSeq sequencing data from all individuals as described in the previous chapters. *HTT* alleles were amplified from each individual from buccal swab DNAs from the US-Venezuelan Collaborative Research Project. DNA samples were successfully genotyped using PCR amplification with MiSeq adapter primers that flank the CAG·CTG repeat tract of the *HTT* locus, followed by sequencing the products using the Illumina MiSeq platform. Those reads were aligned to custom references to reveal read length distributions from each individual from which we were able to determine the genotype in terms of numbers of CAG repeats for both normal and expanded alleles (Chapter 3 and 4). We also determined the presence or absence of atypical structures in *HTT* alleles. Most samples were successfully genotyped by

MiSeq sequencing. All affected individuals with expanded alleles had pure CAG repeats and were present on the typical allele structure with CCG 7 repeats.

To evaluate the instability of the *HTT* CAG repeat, we first analysed our very large dataset from the Venezuelan cohort comprising 740 individuals (439 unaffected and 301 affected) to examine both normal and expanded CAG repeats (Figure 5-1). The large numbers of samples available for analysis allowed us to capture the effect of different CAG repeat lengths on repeat instability. This analysis included all samples with CAG repeats in the normal range, excluding any samples for individuals who have atypical *HTT* allele structures because they don't match exactly the reference which can modify the allele length distribution. We also excluded any individual with two *HTT* alleles that were less than 6 repeats apart, as read length distributions from the two alleles overlap and therefore we cannot unambiguously assign reads to either allele.

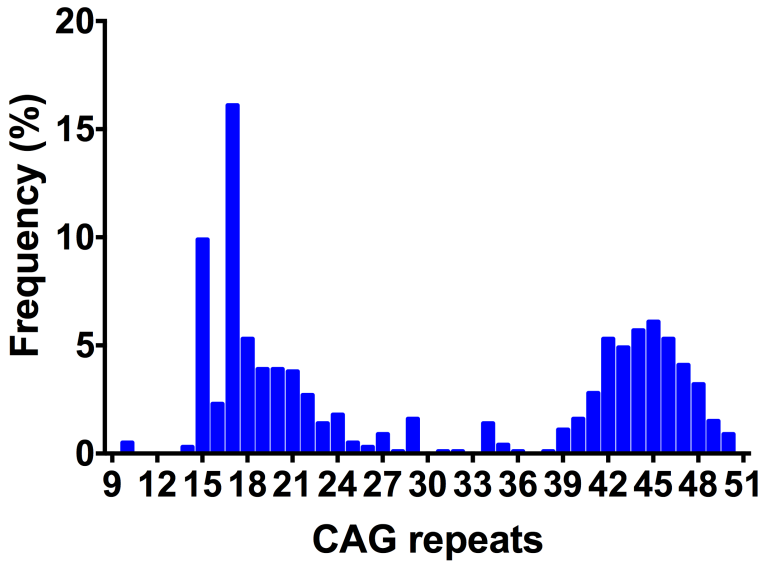

Figure 5-1 Distribution of CAG repeats of 740 *HTT* alleles.The distribution is shown for the number of CAG repeats observed on *HTT* normal and expanded alleles ranging from 10 to 50 repeats in unaffected and affected individuals.

Also, individuals for whom we did not have details about age at sampling were removed from the subsequent analysis, as well as any samples that have a low number of reads. We excluded samples with the low number of reads because of problems, with random sampling errors incorrectly addressing the amount of somatic mosaicism and slippage, and because of potential confounding effects of

background that either result from contamination or primer barcode swapping effects. Alleles with more than 50 CAG repeats were also removed from further analysis as the progenitor allele cannot be clearly identified in such alleles due to PCR slippage and potentially somatic mosaicism.

The data used for the expanded allele measurement (40 to 50 CAG repeats) corresponds to most (90%) of the HD MiSeq data collected for expanded alleles. In order to analyse the data and evaluate the read length distribution for each sample, Alastair Maxwell, who is a bioinformatician working on a Cure Huntington's Disease Initiative (CHDI) project, designed a Python script that arranges all sample output files in one file that centralises the reads to the progenitor allele peak. This makes measuring somatic mosaicism much easier, allowing us to examine the reads around the progenitor alleles in one file.

5.2.1.1 CAG repeat length effects on the magnitude of repeat changes

We analysed the wide distribution of CAG length changes on the magnitude of expansion and contraction across different CAG repeats. Read length distributions of *HTT* CAG repeat lengths are shown in Figure 5-2 for six individuals, five with the normal CAG range (10, 15, 20, 25 and 31 CAG repeats) and one with a CAG repeat expansion (40 CAG repeats). The mapped reads of *HTT* alleles showed multiple reads were mapped against references of different lengths. The main peak is the most common allele, represented as (n) in blue, which is the number of reads corresponding to the highest peak of an allele read count distribution: 10, 15, 20, 25, 31 and 40 CAG for the *HTT* allele (Figure 5-2). This peak is defined as the progenitor allele length, i.e. the single allele transmitted from the affected parent.

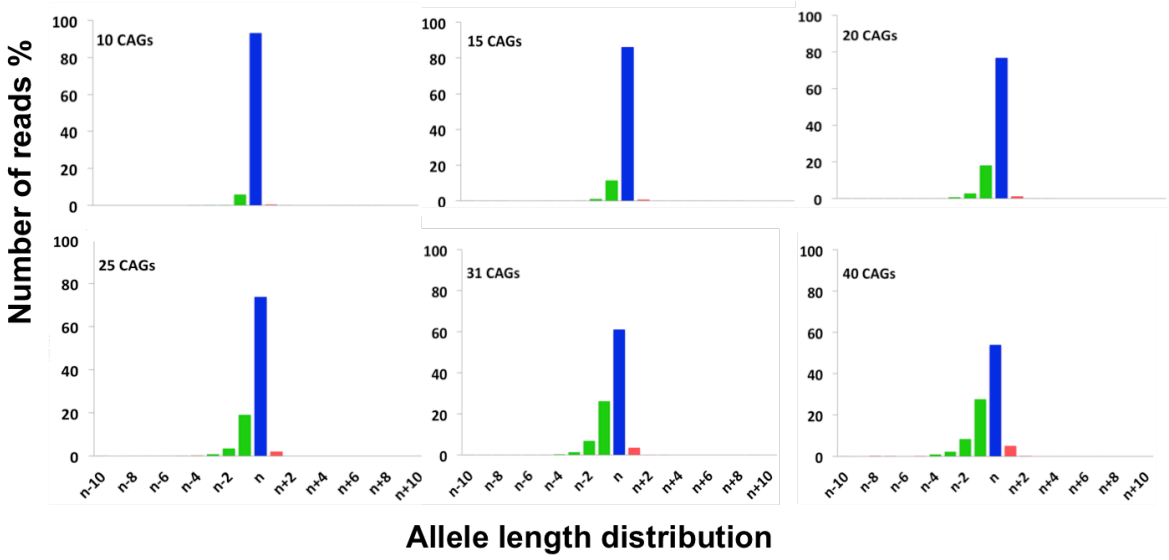

Figure 5-2 Qualitative assessment of CAG allele length distributions in 6 different individuals who have different CAG repeats (10,15, 20, 25, 31 and 40 CAG repeats).The proportion of allele length was quantified from the MiSeq sequencing read distribution, the inherited allele length (blue), contractions (green), and expansions (red).

Green peaks (n-1 to n-10 peaks), which are to the left of the main allele, represent repeat unit losses that most likely occur due to backward PCR slippage. The PCR slippage is biased toward deletion and results in generating products smaller than the starting molecules (Pearson *et al.*, 2002).

The n-1 reads are relatively low for small alleles within the normal size. For larger alleles (40 CAG repeats), the proportions of n-1 reads are greater than in normal alleles as seen in Figure 5-2. The degree of n-1 reads appears to correlate strongly with the CAG repeats length. The red peaks to the right of the main allele (n+1 to n+10) are the number of reads corresponding to additional CAG repeat gains, and peaks larger than (n) are likely to be somatic variants captured by MiSeq sequencing approach (Figure 5-2). As somatic mosaicism is allele length dependent and expansion-biased, it is expected to result in a higher proportion of reads larger than the progenitor allele in larger alleles. Indeed, the proportion of reads (n+1) appears to be higher for larger alleles as shown in Figure 5-2. The individual with 40 CAG repeats has a higher proportion of reads in n+1 peak than the individuals with the normal size of CAG repeats. Overall, the repeat length distributions revealed that longer progenitor repeat had a higher proportion of n-1 and n+1 reads compared to shorter repeat lengths.

5.2.1.2 Repeat length dependent PCR slippage.

In order to identify the determinants of the proportion of reads shorter than n, we calculated the proportion of (n-1) reads relative to the number of reads for the main allele. There was a very strong correlation (r^2= 0.978, P < 0.001) between (n-1) reads and inherited allele length (Figure 5-3). Therefore, the proportion of (n-1) reads is dependent on allele length, and ~ 98% of the variation is accounted by allele length. This is consistent with the number of (n-1) reads being driven largely by PCR slippage. We would also expect that somatic instability would be highly age-dependent. Although it is possible that age-dependent somatic mosaicism contributes toward the remaining 2% of the unexplained variability, it is clear that most of the n-1 read variation is highly repeat length-dependent, which is consistent with it being primarily driven by PCR slippage.

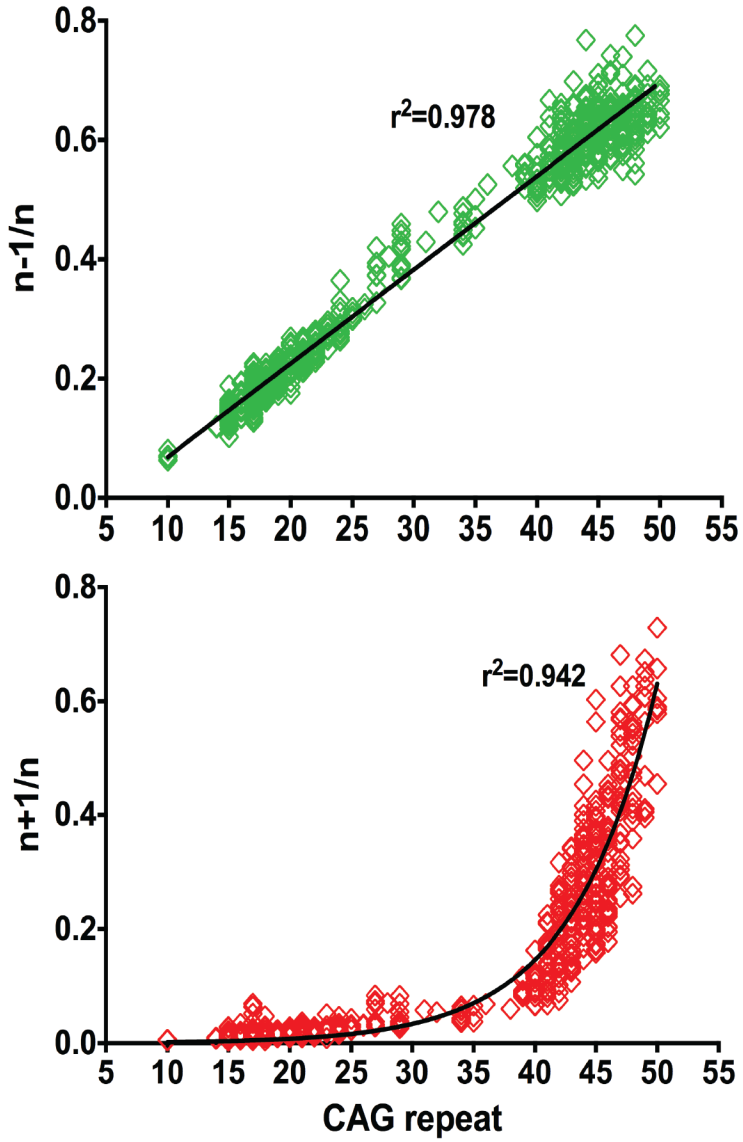

Figure 5-3 The relationship between the degree of repeat length variation and inherited repeat length in 740 alleles including normal and expanded alleles ranging from 10 to 50 CAG repeats. The scatter plot on the top display the proportion of n-1 reads relative to the main allele in green dots. The lines fitted to the scatter plot reveals a strong correlation (r^2=0.978, *P* <0.001) between the n-1 reads and inherited allele length. The scatter plot on the bottom displays the n+1 reads relative to the main allele in red dots. The lines fitted to the scatter plot reveal an exponential association (r^2=0.942, *P* < 0.001) between the n+1 reads and inherited allele lengths. The degree of repeat length variation was measured from MiSeq read count distribution (see Chapter 3 and 4).

Then, the frequency of (n+1) reads compared to the inherited allele reads was tested in the same way as the (n-1) reads (Figure 5-3). The read length distributions were examined to test whether there is any evidence of somatic mosaicism. As expected, there was a significant correlation between the (n+1) reads and inherited allele length. However, the shape of the distribution was very different from the distribution of n-1 reads. Although allele length is a

major determinant, the shape is clearly different in these two distributions. This shape is obviously not linear, and if fitted with a simple linear regression model, the correlation coefficient was relatively lower (r^2 = 0.52). Analysis of a range of curve estimation regression models revealed an exponential association (r^2 =0. 942, P < 0.001) was found to best describe the relationship between allele length and n+1 reads for the overall distribution than a single line across the entire CAG repeat sizes. We noticed the slope of the curve correlating CAG repeats length over 40 repeats with n+1 reads is steeper in this range than in the normal range. Therefore, we considered whether a two-segment regression response could improve the correlation and produce a better fit for the data.

To fit the two response models at each CAG repeat length range, we split the data into two groups at the point of 40 CAG repeats as shown in Figure 5-4. This allows more meaningful comparisons to be made between two regression models within the different CAG repeat ranges. We plotted the CAG repeat length for the normal and expanded alleles against the n+1 reads for each individual.

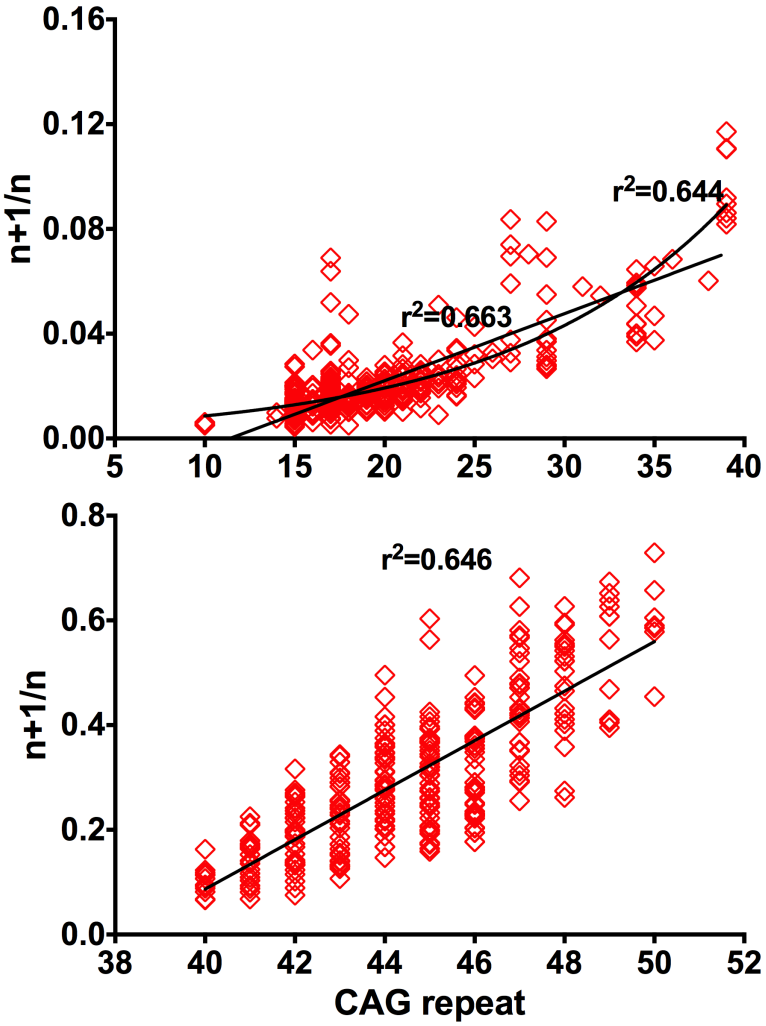

Figure 5-4 The relationship between n+1 reads and the inherited CAG repeat in the normal allele range and mutant range. The scatter plot on the top displays the proportion of n+1 reads relative to the main allele in the normal allele length range. The two lines fitted to the scatter plot reveal a linear regression (r^2=0.663, P <0.001) and an exponential regression (r^2=0.644, P <0.001) between the n+1 reads and inherited allele lengths. The scatter plot on the bottom displays the proportion of n+1 reads to the main allele in mutant CAG range. The lines fitted to the scatter plot to reveal a linear association (r^2=0.646, P < 0.001) between the n+1 reads and expanded CAG repeats.

For the data in the range of normal alleles, the distribution of n+1 reads fits linear regression (r^2= 0.664, P <0.001) with inherited allele length (Figure 5-4). Exponential association analysis of the data showed a slightly lower correlation coefficient (r^2= 0. 644, P <0.001), therefore the linear correlation fits the data better. The distribution of n+1 reads for expanded alleles displayed a good fit to a linear association (r^2= 0.646, P <0.001) (Figure 5-4). The remaining variation in n+1 reads of affected HD allele range that is not explained by the repeat length

is possibly age-dependent somatic instability. The level of variation in the n+1/n reads was obviously greater in the expanded alleles than in the normal alleles.

Clearly, there are two responses between n+1 reads and CAG repeat size. Also, there is a clear division in the data from the linear curve in the normal range to linear regression fit to the mutant range. These data allowed us to assess the allele length distribution and relate these to the inherited allele length using NGS read distribution.

5.2.1.3 Age-dependent somatic instability in HD individuals

Having established the repeat length is the major determinant of read length variability and showing that people with the same CAG repeat appear to have a different degree in variation in n-1 and n+1 reads, factors other than HD CAG repeat size may influence instability. Thus we tested whether the residual variation of CAG repeat is influenced by age at the time of sampling. Therefore, we sought to determine the effect of age at sampling on the relative frequency of allele length changes that have a specific CAG repeat length of both normal and mutant alleles.

The wide range of CAG repeats sizes in our dataset allowed us to perform an analysis of the effect of age at sampling on the variation of repeat instability in specific CAG repeats. This dataset is composed of n-1/n, and n+1/n reads for all individuals inheriting one of four different repeat sizes of CAG repeats in the normal range (15,17, 18 and 20 CAG repeats) and one of four different repeat sizes in the mutant range (41, 43, 46 and 48 CAG repeats). These CAG repeats were selected because they are associated with the most frequent alleles.

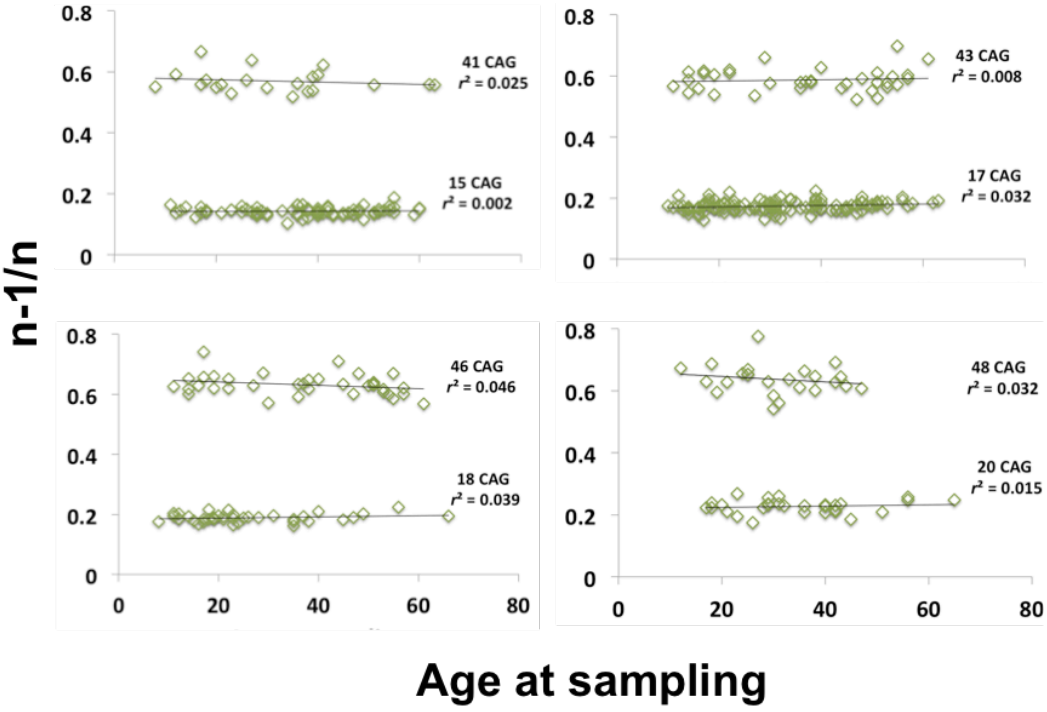

Figure 5-5 The correlation between n-1/n reads and age at sampling for four normal and four expanded alleles. A. The correlation between the proportion of n-1 reads relative to the main allele for CAG repeat numbers in the normal range: 15, 17, 18 and 20; 15 CAG repeats (r^2 = 0.013, *P* = 0.743), 17 CAG repeats (r^2 = 0.024, *P* = 0.05), 18 CAG repeats (r^2 = 0.013, *P* = 0.231) and 20 CAG repeats (r^2 = 0.022, *P* = 0.532). B. The correlation between the proportion of n-1 reads to the main allele for four CAG repeat numbers in the affected range: 41, 43, 46 and 48 CAG repeats; 41 CAGs (r^2 = 0.026, *P* = 0.491), 43 CAG repeat (r^2 = 0.021, *P* = 0.607) 46 CAG repeat (r^2 = 0.039, *P* = 0.191), and 48 CAG repeat (r^2 = 0.012, *P* = 0.406).

In this dataset, there was no relationship between age at sampling and n-1/n reads for normal repeats 15 CAG repeats (r^2 = 0.013, *P* = 0.743), 17 CAG repeats (r^2 = 0.022, *P* = 0.05), 18 CAG repeats (r^2 = 0.013, *P* = 0.231) and 20 CAG repeats (r^2 = 0.022, *P* = 0.532)(Figure 5-5). P-values were non significant for all CAG repeats expect for 17 CAG repeat which was marginally significant. Probably this would be non significant after performing a correction for multiple testing. There is no evidence of any age effect; therefore there is no effect of somatic instability.

Also, there was no correlation between n-1/n reads and age at sampling in the expanded allele ranges, 41 CAGs (r^2 = 0.026, *P* = 0.491), 43 CAG repeats (r^2 = 0.021, *P* = 0.607), 46 CAG repeats (r^2 = 0.020, *P* = 0.191), and 48 CAG repeats (r^2 = 0.012, *P* = 0.406)(Figure 5-5). However, P-values were not significant for any of the expanded alleles. Age at sampling did not significantly alter the frequency of the n-1 reads. These data showed a higher level of n-1 reads with larger CAG

repeats, that is most likely explained by PCR slippage and not by age-dependent somatic mosaicism. These data further suggest n-1 reads are most likely driven by backward PCR slippage, as it is allele length dependent and not age-dependent.

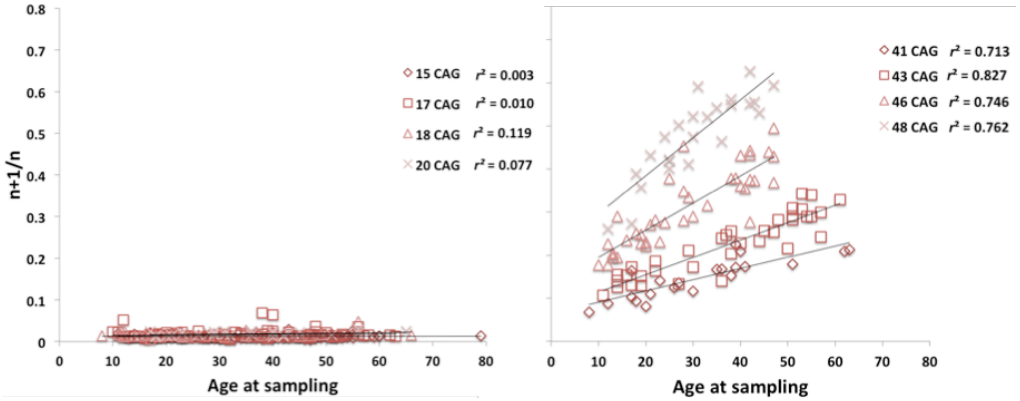

Figure 5-6 The correlation between n+1/n reads and age at sampling for the normal and expanded CAG repeats. A. The correlation between the proportion of n+1 reads relative to the main allele for four CAG repeat numbers in the normal range: 15, 17, 18 and 20: 15 CAG repeats (r^2= 0.011, *P*= 0.667), 17 CAG repeats (r^2= 0.001, *P*= 0.291), 18 CAG repeats (r^2= 0.095, *P* = 0.032) and 20 CAG repeats (r^2= 0.043, *P* = 0.145). B. The correlation between the proportion of n+1 reads relative to the main allele for four CAG repeat numbers in the affected range: 41, 43, 46 and 48 CAG repeats: 41 CAG (r^2= 0.698, *P* <0.001), 43 CAG repeats (r^2= 0.822, *P* <0.001) 46 CAG repeats (r^2= 0.739, *P* <0.001), and 48 CAG repeats (r^2= 0.751, *P* <0.001).

There was no correlation between the proportion of n+1/n reads and age at sampling for alleles in the range 15 to 20 CAG repeats (Figure 5-6). Analysing the normal alleles showed no significant effect of age at sampling on the repeat length changes for 15, 17 and 20 CAG repeats: 15 CAG repeats (r^2 = 0.011, *P* = 0.667), 17 CAG repeats (r^2 = 0.001, *P* = 0.291) and 20 CAG repeats (r^2= 0.043, *P* = 0.145). For 18 CAG repeats (r^2= 0.095, *P* = 0.032) there was a significant correlation with age that might not show any significant P-value if we do multiple testing corrections. Overall, the CAG repeat in the normal range showed no consistent changes with age.

As expected, there was a significant correlation between the (n+1)/n reads and age at sampling with mutant alleles (Figure 5-6): 41 CAGs (r^2 = 0.698, *P* <0.001), 43 CAG repeats (r^2 = 0.822, *P* <0.001), 46 CAG repeats (r^2 = 0.739, *P* <0.001), and 48 CAG repeats (r^2 = 0.751, *P* <0.001). Older patients had more expansions for any particular repeat number, and the slope was much steeper with larger

alleles. These data are not consistent with PCR slippage and suggest a high proportion of the n+1 reads are mediated by allele length and age-dependent somatic mosaicism. These findings support the view that age plays a major role in determining *HD* CAG instability that is quantified by n+1 reads.

5.2.1.4 A model to quantify somatic instability in expanded CAG repeats

We have generated clear evidence that n+ measurement for alleles over 40 repeats likely contains a measure of somatic instability. We are aiming to develop a somatic instability quantification method that is applicable to high throughput assays and also, to gain insight into factors that influence the level of somatic mosaicism in buccal DNA of HD patients and relate these to disease severity. The total data set comprised 301 individuals, of which two are homozygous, and 299 are heterozygous, who have a CAG allele more than 40 repeats.

In the previous section (5.2.1.2), we quantified putative somatic mosaicism from MiSeq read count distribution using the ratio n+1/n, where n is the number of reads corresponding to the highest peak of an allele read count distribution and n+1 is the number of reads corresponding to one additional CAG repeat gain larger than n. This proportion of reads may oversimplify the somatic instability measure for these sequenced reads larger than n, as reads of n+2, n+3, n+4, *etc.*, are also present in some distributions.

	Measure of somatic mosaicism	**Allele length interval considered**
1	$\frac{n+1}{n}$	The number of reads for one read larger than inherited allele length
2	$\sum_{i=1}^{5}\frac{n+i}{n}$	The number of reads for n+1 to n+5 peaks
3	$\sum_{i=1}^{10}\frac{n+i}{n}$	The number of reads for n+1 to n+10 peaks
4	$\sum_{i=1}^{10}\frac{n+i}{n}\times i$	The number of reads for n+1 to n+10 peaks were multiplied by the changes from the progenitor allele

Table 5-1 Measurement of somatic mosaicism in MiSeq data from expanded alleles. (n) is the number of reads corresponding to the highest peak of an allele read count distribution obtained from MiSeq sequencing.

To define appropriate measures for quantifying somatic instability in HD MiSeq data, we have investigated four measures (Table 5-1). We have correlated the proportion of reads larger than n against allele length. By quantifying the proportion of these reads, we can test whether there is any evidence that it may be reflecting somatic mosaicism. Table 5-1 illustrates the instability indices used for quantifying somatic mosaicism. First, the instability index was calculated by dividing the number of reads for one read larger than the progenitor allele by the number of reads of the progenitor allele, which is presented as n+1 divided by n. This has been selected because most expansions are seen in the n+1 peak. Then, the second instability index was derived by dividing the sum of the reads from n+1 to n+5 by the number of reads for the progenitor allele that is presented as n+5/n. These reads were included in the instability index as some samples have reads observed with those peaks that may be reflecting higher levels of somatic mosaicism. Likewise, n+10/n was chosen as an instability index because some samples have reads in this range. Reads for n+10 was chosen as the higher boundary of the interval because it includes 99.9% of the sequenced reads longer than inherited allele. The last index selected is n+10 where the peak heights were multiplied by the changes from the progenitor allele. This instability index yields an average gain per read that gives more weight to large expansions as they likely result from multiple events rather than a single event. We then attempted to determine whether the different possible somatic instability measures might be considered as an indication of somatic mosaicism.

Somatic instability model	SM Measure 1		SM measure 2		SM measure 3		SM measure 4	
Model 1: SM = Allele length	r^2= 0.64	P < 0.001	r^2= 0.60	P < 0.001	r^2= 0.59	P < 0.001	r^2= 0.52	P < 0.001
Model 2: SM= Age S	r^2= 0.16	P < 0.001	r^2= 0.14	P < 0.001	r^2= 0.15	P < 0.001	r^2= 0.15	P < 0.001
Model 3: SM= Allele length + Age S	r^2= 0.86	P < 0.001	r^2= 0.80	P < 0.001	r^2= 0.80	P < 0.001	r^2= 0.72	P < 0.001
Model 4: SM = Allele length + Age S + (allele length x Age S)	r^2= 0.90	P < 0.001	r^2= 0.88	P < 0.001	r^2= 0.86	P < 0.001	r^2=0.80	P < 0.001
Model 5: SM= Allele length + Age S + (allele length x Age S)+ Primers	r^2= 0.90	P < 0.001	r^2= 0.87	P < 0.001	r^2= 0.87	P < 0.001	r^2=0.81	P < 0.001
Model 6: SM= Allele length + Age S + (allele length x age S)+ Run	r^2= 0.90	P < 0.001	r^2= 0.86	P < 0.001	r^2= 0.86	P < 0.001	r^2=0.80	P < 0.001
Model 7: SM= Allele length + Age S + (allele length x age S)+Primers +Run	r^2= 0.90	P < 0.001	r^2= 0.87	P < 0.001	r^2= 0.87	P < 0.001	r^2=0.82	P < 0.001

Table 5-2 Regression model of the relationship between somatic mosaicism (SM), and the inherited allele length and age at sampling (A_S) using SPSS statistics software (IBM). The table shows the adjusted squared coefficient of correlation (adjusted r^2) and statistical significance (*P*) associated with each model by testing four different somatic mosaicism measure (SM measure) as described in Table 5-1. n=number of reads aligned for the main HD allele, primers are HD primers used in amplifying the HD allele and run is the MiSeq run in which samples were sequenced using the MiSeq platform.

To assess the utility of these measures, we performed a series of statistical analyses (Table 5-2). We first evaluated the relationship of allele length to somatic mosaicism measures using linear regression models. Using a simple linear regression model, we found that CAG repeat length is correlated with somatic mosaicism and accounts for ~50 to 60% of the variation in somatic instability among somatic instability measures (Table 5-2, model 1). From this model, we can establish that CAG repeat length is the major determinant of somatic instability in HD (r^2=0.52 to 0.64). Previously a similar analysis using single molecule PCR on a small subset of Venezuelan samples revealed that the inherited allele length explained approximately 70% of the variation in somatic mosaicism (Veitch *et al.*, 2007). This is higher than observed in our study. However, all individuals in that study were selected with the same age, which is different from our study that includes individuals with variable age at sampling. Nonetheless, the degree of somatic variation was different for individuals with the same inherited allele length, suggesting that factors other than HD repeat size contribute to the somatic mosaicism. Given that the measure of somatic instability is likely to change through the individual's lifetime (Kennedy *et al.*,

2003), we then tested this hypothesis using regression analysis. As expected, age at sampling was correlated with somatic instability (r^2= ~ 0.15) (Table 5-2, model 2). Therefore, a more complicated model may be required to elucidate the factors involved in somatic variation that includes the allele length and age at sampling.

We also evaluated additional models that include both variables. Thus, we tested a model with both allele length and age at sampling effects on somatic instability measures (Table 5-2, model 3), as well as another model for these two factors and their interaction (model 4). Both models showed an improvement in explaining the variation in somatic instability (Table 5-2, model 3 and 4, r^2= ~ 0.8, $P < 0.001$) over model 1 and 2. Lastly, we assessed the contribution of other factors in addition to allele size, age at sampling and their interaction. These factors are the different HD flanking primers used in amplifying the HD allele, and the MiSeq run in which the samples were sequenced in MiSeq platform. My colleague Marc Ciosi who is working on the CHDI project, prepared one of the libraries that includes 92 samples. He used the inner primers (31329F/33934) to amplify HD alleles in those samples. The rest of the samples were amplified using the intermediate primers (HS 319F-3395.5), resulting in sequence reads with different length flanking sequences. Also, each library has been sequenced on different MiSeq runs, which might introduce a variation between libraries, as there are differences in quality and the total number of reads generated by different MiSeq runs. Therefore, we tested another model with all variables considered above and a primer parameter (Table 5-2, model 5) and another model with a possible run effect (model 6). Lastly, we assessed a model with the same factors as model 4 and both run and primers effects (Table 5-2, model 8). With the addition of the primers and run variables into the model, the coefficients and the model did not show any improvement over model 4. The primers and runs show minor differences, and none of the individual parameters were significant on their own. Thus, the primers and run variables did not explain a significant proportion of the variation in the somatic mosaicism models.

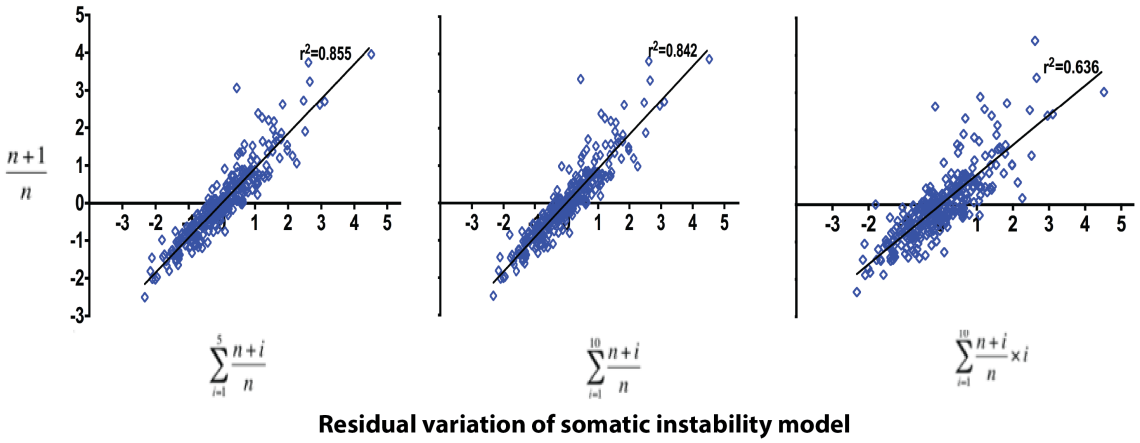

Figure 5-7 Comparison of standardized residual variation of the selected Somatic mosaicism model. Scatterplots show the relationship between standardised residuals of the selected model for each measure. The somatic mosaicism measures were explained in Table 5-1. The selected model is model 4 that is described in Table 5-2.

To confirm all the parameter effects, stepwise linear regression was performed in SPSS software for model selection. This step is essential in order to remove the weakest correlated variable and to identify the most relevant descriptive variables for somatic instability among all the models considered. The selected model was model 4 that will be used in further analysis (Table 5-2, model 4). In each case, the run and primers effects were excluded from the models, as they were not significant.

To gain some idea whether these different somatic measures were assessing the same information, we quantified the residual variation for all individuals for each of the four measures of somatic mosaicism under model 4. The standardized residuals were used for pairwise comparisons between the measures (Figure 5-7). The four measures contain very similar information about quantifying somatic instability as shown by the significant correlations between the standardised residuals of the somatic instability model for these measures (Figure 5-7). Those residuals were similar, but not exactly the same with r^2 of 0.85 between measures 1 and 2 and r^2 of 0.84 between measure 1 and 3 (Figure 5-7). In contrast, the correlation between measure 1 and 4 was lower (r^2= 0.63) than the correlation between measure 1 and measures 2 and 3. The low correlation between measures 1 and 4 could be results of increased noise in one

of the measures. Alternatively, measure 4 might be revealing some other aspect of somatic mosaicism that is more related to the frequency of larger expansions.

We cannot distinguish between these two possibilities in measure 4. However, we expect that allele length and age at sampling to be the primary driver of somatic mosaicism. Therefore, it is likely that any measure of somatic mosaicism should be well explained by those two variables. By considering at the overall adjusted r^2 for the same model that incorporate allele length, age and the interaction, we realised that the measure 1 has the most of it's variation explained using that model r^2 of 0.9 whereas measure 4 has r^2 of 0.8. Although we cannot completely exclude the idea that a different biological feature is being revealed in measure 4, it seems more likely that measure 4 is incorporating a greater degree of noise than present in measure 1. Measures 2 and 3 were related to slightly less variation in somatic instability that is explained by age, and allele length with all models than measure 1 and also their residual variation was highly significant with the n+1 measure. In addition, measure 4 was associated with a low variation in somatic instability for all models, and it's residual variation compared with measure 1 was low. This suggests all four measures are measuring something very similar, and the n+1/n had the highest r^2 value, so was chosen as the most useful model. Among the four measures, measure (1), which is the proportion of n+1 divided by n, was strongly correlated with the other three measures and associated with the highest adjusted r^2. The measure 1 will be used for subsequent phenotypic analysis. Table 5-3 shows the regression analysis model summary for the selected model and measure of the relationship between somatic instability, inherited allele size and age at sampling (r^2= 0.903, P < 0.001). The model accounted for 90% of the variation in somatic mosaicism and was a highly significant model. This model was fitted to measure somatic mosaicism from HD affected individuals who inherited CAG repeats between 40 and 50. The model showed the changes in somatic instability increased as a function of repeat length and time of sampling and their interaction.

Model	Adjusted R sqaure	*P* -value	Parameters	Coefficient	Standard error	t-statistics	*P* - value
Model 4: SM = Allele length + Age S + (allele length x Age S)	0.903	< 0.001	Intercept		0.136	-5.3	< 0.001
			Allele length	0.318	0.003	6.077	< 0.001
			Age S	-3.404	0.004	-9.315	< 0.001
			Allele length x Age S	3.876	0.001	10.63	< 0.001

Table 5-3 The selected model for somatic mosaicism measure showing the relationship between somatic mosaicism (SM), and the inherited allele length and age at sampling (Age S) using SPSS statistics software (IBM).The table shows the adjusted squared coefficient of correlation and statistical significance (*P*), the coefficient, standard error, t-statistic and statistical significance (*P*) associated with each parameter in the model.

5.2.2 Genotype and phenotype correlation

CAG repeats number is the primary determinant of age at onset of HD. However, wide ranges of ages of onset are observed for individuals with identical inherited CAG repeat numbers. A strong inverse correlation exists between the CAG repeat size and age at onset for motor signs in HD. The CAG repeat length contributes to approximately 70% of the variability in age at onset, indicating that modifying factors other than CAG repeat are likely to be involved in determining the age of onset (Wexler et al, 2004). Analysis of the Venezuelan kindred estimates ~40% of this variability in age at onset is attributable to genetic factors other than the HD gene, and ~60% is environmental factors (Wexler et al, 2004). We hypothesised that part of that variation in age at onset is mediated by somatic instability.

5.2.2.1 Variation in age at onset

In order to examine the correlation between age at onset of symptoms and the inherited CAG repeat length in the HD patients of the Venezuelan cohort, the CAG repeat length in the HD gene was plotted against age at onset for 137 HD patients (Figure 5-8). This sample includes all individuals that were heterozygous for an HD mutation and for whom we have a clinical age at onset and repeat length data. There was a statistically significant negative correlation between the CAG repeat size and age of disease onset in HD patients (r^2 =0.46, P < 0.001).

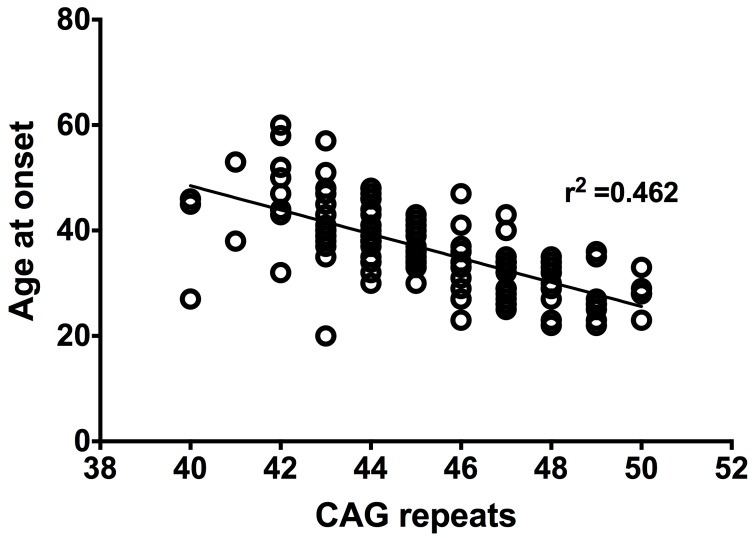

Figure 5-8 The relationship between CAG repeats and Age at onset in 137 HD patients who have CAG repeats between 40 and 50.The straight line shows the regression line for all HD patients using a linear model (r^2= 0.462, P- value < 0.001). This data revealed that the inherited CAG repeats are a major modifier of the disease onset and severity in HD patients.

Three different methodological approaches have been used previously to investigate the correlation between CAG repeat size and age of disease onset. Some studies used linear regression to assess the correlation between the two variables. The second type was to estimate the correlation between CAG and the natural logarithm of age at onset, as an exponential relationship was noticed in the data. In a dataset with a wide range of alleles from 40 to 110 CAG repeats, an exponential model was shown to fit the data better than a linear model (Andresen *et al.*, 2007). The third approach was to assess the correlation between the natural logarithm of both age at onset and the CAG repeat (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004). We therefore tested three models for each different method used in previous studies for the age at onset and allele length (Table 5-4). Both models (2 and 3) showed no improvement in explaining the variation in age at onset, despite the logarithmic transformation of the variables. We found that the correlation coefficient between the two variables was slightly higher using linear regression (r^2 =0.46, P < 0.001) (model 1,Table 5-4) compared to the logarithm of the age at onset and logarithms of both CAG repeat and age of onset. Examination of the relationship between CAG repeats and age at onset showed the data might have non-linear components. Therefore, we also tested both quadratic and exponential models of the CAG repeats. There was no additional variation explained with these

models (quadratic, adjusted r^2=0.46, P < 0.001, exponential, adjusted r^2=0.45, P < 0.001), we therefore assumed the linear regression provides the best fit for the data among the three models.

Model	Adjusted R sqaure	P -value	Parameters	Coefficient	Standard error	t- statistics	P - value
Model 1: Ao = Allele length	0.458	< 0.001	Intercept		9.636	14.539	< 0.001
			Allele length	−0.680	0.213	−10.772	< 0.001
Model 2: log (Ao) = Allele length	0.449	< 0.001	Intercept		0.116	23.919	< 0.001
			Allele length	−0.672	0.003	−10.583	< 0.001
Model 3: log(Ao) = log(Allele length)	0.447	< 0.001	Intercept		0.443	14.033	< 0.001
			Allele length	−0.672	0.268	−10.532	< 0.001
Model 4: Ao = Allele length (excluding the outliers)	0.552	< 0.001	Intercept		8.846	17.01	< 0.001
			Allele length	−0.745	0.195	−12.886	< 0.001
Model 5: Ao = Allele length (excluding the outliers and (40-41) CAG repeats)	0.558	< 0.001	Intercept		9.49	16.563	< 0.001
			Allele length	−0.749	0.208	−12.756	< 0.001

Table 5-4 Regression model between age at onset (Ao) and allele length.The table shows the adjusted squared coefficient of correlation and statistical significance (*P*), the coefficient, standard error, t-statistic and statistical significance (*P*) associated with allele length in each model.

We also evaluated whether analysis of subjects with an exclusion of the data outliers might provide a better correlation to age at onset variation. We also considered excluding individuals who inherited alleles of 40 and 41 repeat because very few individuals have those CAG repeat and they appeared to be outliers. Thus, we tested a model with exclusion of outliers as well as removing these outliers and CAG repeats from 41 to 42 CAG repeats and separate analysis of those individuals excluded from the model (Table 5-4, model 4 and 5). We obtained a statistically significant result with an increase of the correlation coefficient r^2=0.552, r^2=0.558 for model 4 and 5 respectively. Both models showed an improvement in explaining the variation in age at onset over model 1, but they are not widely different from model 1. Therefore, the last model in which we exclude the outliers and individuals who inherit 40 and 41 CAG repeat, didn't explain additional variation compared to model 4. Therefore, the selected model for the linear regression analyses between CAG repeat and age at onset was model 4, which will be used in the further analysis for 135 individuals.

Thus, our results revealed that the inherited allele length is an important determinant of disease severity in HD, accounting for 55% of the variation in age at onset in the linear regression correlation. The remaining variation (45%) of the age of onset must be explained by factors other than the inherited CAG repeat number. This is especially true for individuals with identical repeat lengths showing wide variability in age at onset. For example, individuals with 44 CAG repeats exhibit age at onset between 30 to 50 years of age. The wide range span in age at onset suggests the extensive variation in onset age is not explained by CAG repeat size alone. Also, it is clear the correlation coefficient for the association between age at onset and allele length in our model is lower than in a previous study. Andresen et al. have analysed the inherited allele length in the range from 40 to 110 CAG repeats in 443 Venezuelan patients, which accounts for approximately 70% of the variation in age at onset (Andresen *et al.*, 2007). The narrow repeat length range in our data and relatively low number of samples could explain the obtained different correlation coefficient from previous studies since the coefficient tends to be higher for tests measuring a wide range of data.

5.2.3 Does the level of somatic instability contribute to the age of HD disease onset?

In HD patients, inherited CAG repeat length is the major determinant of disease severity. Clearly, longer CAG repeats cause more severe HD in humans. Longer alleles lead to a greater level of somatic mosaicism that is biased toward expansions. Therefore, it is logical to assume that somatic instability contributes to the pathogenic process of the disease. Individuals in whom the repeats expand more rapidly than average for a particular CAG repeat number would have an earlier age of onset than expected and vice versa. We thus expect there is an inverse correlation in residual variation in age at onset that is independent of the effect of the repeat size and residual variation somatic instability, which is independent of the effect of age at sampling and the repeat size. In order to investigate the age-dependence of somatic instability, we used simple linear regression of residual variation in age of onset (model 1, Table 5-4) and residual variation of somatic instability (model 4, Table 5-2) in the 135 individuals.

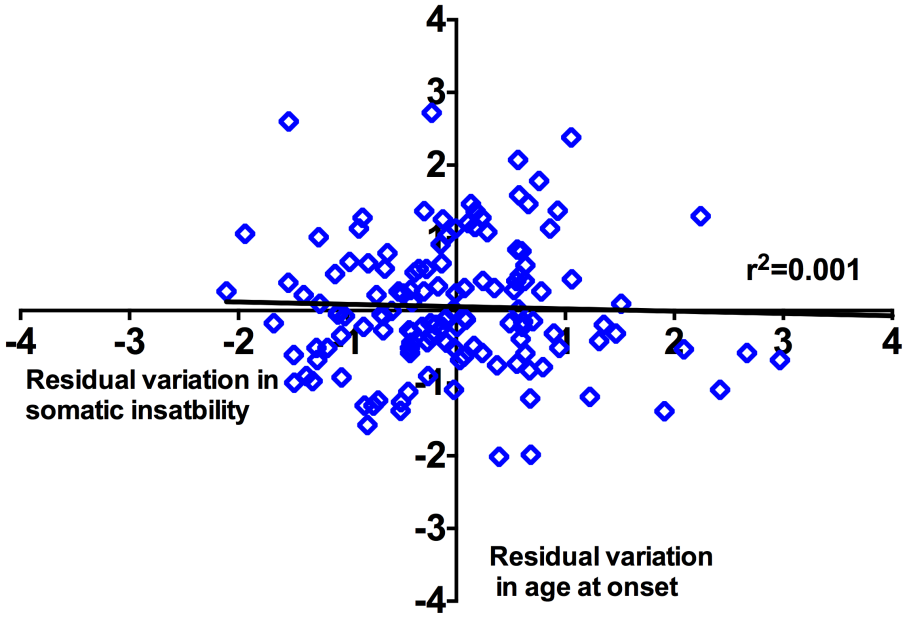

Figure 5-9 The relationship between the residual variation in age at onset and residual variation in somatic instability. The graph shows the relationship between standardized residual variation in age at onset not accounted for CAG repeats (model 1,Table 5-4), and standardized residual variation in somatic instability not accounted forage at sampling and CAG repeats (model 4, Table 5-2) in 137 HD patients. Linear regression shows a negative correlation (r^2= 0.04) but it is not significant (P = 0.47).

Our analysis did not provide direct evidence or support the correlation between the residual variation of both age at onset of HD and somatic instability (r^2= 0.001, P = 0.47) (Figure 5-9). This non-significant association between age at onset and somatic instability was not expected, as we assumed that the somatic instability accounted for some of the variation in HD age at onset and the repeat expands more rapidly for individuals who have an earlier age at onset.

5.3 Discussion

The expanded CAG repeat at the HD locus is unstable, which means it can mutate through germline transmission and also within individual tissues. The repeat is somatically unstable in a process that is tissue-specific, age-dependent and expansion-biased (Kennedy *et al.*, 2003). It has been shown that in HD, there is a noticeable correlation between the tissues vulnerable to neuropathology and those exhibiting a great level of somatic instability of the repeats. This observation has led to the hypothesis that somatic instability contributes to HD pathogenesis and progression.

The somatic instability has been investigated by bulk PCR amplification and subsequent Genescan analysis of PCR products. This does not provide a full presentation of the repeat sizes, or reflect the dynamic nature of CAG repeat instability in the HD patients. Although earlier studies using single molecule PCR analysis showed success in precisely estimating the repeat size and somatic instability, this approach is impractical and labour intensive for high throughput analysis in a large cohort of patients. To overcome the limitations of using previous approaches in evaluating somatic instability, we thus used high throughput NGS to quantify somatic mosaicism of the CAG repeats in HD directly from NGS read length distributions. Our hypothesis was that the main drivers for somatic mosaicism are repeat length and age, and that individual-specific residual variation for both effects is also associated with disease severity.

In this study, we have analysed the pattern of somatic instability in buccal cell DNAs from unaffected and affected individuals from the Venezuelan kindred. This was done in order to investigate the role of various factors in somatic instability and to test whether somatic mosaicism contributed to disease severity. Our data highlighted the importance of factors such as CAG repeat length and age at sampling. We described the detailed pattern of read length distributions in the normal and expanded alleles from unaffected and affected individuals.

We also analysed the wide distribution of CAG length changes across different CAG repeat numbers. The read distribution of n-1 and n+1 reads were used in the analysis, to examine the role of CAG repeat number in the repeat changes. The data revealed that the repeat length distribution was slightly changed in the majority of normal alleles, while the longer alleles showed considerable levels of changes in the read distribution. Overall, the repeat length distributions revealed that longer repeat had a significantly higher proportion of n-1 and n+1 reads when compared to shorter repeat length. For further analysis of the n-1 and n+1 reads, we calculated the proportion of n-1 and n+1 reads compared to the number of reads for the main allele. These data allowed us to evaluate the allele length distributions and relate these to the inherited allele length using NGS read distributions.

We investigated the effect of age at sampling on the relative frequency of repeat length changes on both normal and mutant alleles. To do this, we correlated the n-1/n and n+1/n reads against age at sampling for four normal and four expanded alleles. These data showed age at sampling has an impact on the mutant repeat instability. Overall, there was no significant level of instability in the normal alleles, but with mutant alleles, instability was observed in all expanded alleles, and an even greater level was observed with increasing age. This is consistent with the greater level of somatic mosaicism observed in the expanded alleles (Veitch *et al.*, 2007). These data confirmed that for expanded CAG alleles somatic instability is age-dependent and expansion-biased and the normal alleles are highly stable. These analyses revealed clear differences in the patterns of instability between the normal and mutant alleles. This analysis suggests the allele length and age are major modifiers of somatic mosaicism as expected from previous studies, and that NGS approach can be used to quantify somatic mosaicism in HD patients.

There is strong evidence from our data that n+ measurement for alleles between 40-50 CAG repeats likely to contain a measure of somatic instability in samples. The aim was to develop a somatic instability quantification method that is applicable to high throughput assays. Here we show that the somatic expansions increase linearly with increasing allele length in human HD patients. These analyses revealed that CAG repeat length is a major modifier of somatic instability, accounting for 64% of the variation in somatic instability. The data show the level of somatic expansion is highly dependent on CAG repeat size. Notably, age at sampling alone has a low correlation coefficient with the level of somatic instability. The age effect has not explained the variation in somatic instability measure; this most likely occurs from age at sampling bias in HD patients in which individuals with larger alleles tend to be sampled at a much younger age, and individuals with shorter alleles tend to be sampled at an older age. Nonetheless, age was confirmed as a major factor in determining the level of somatic instability by using multivariate analyses that confirm the interaction between age at sampling and allele length. The variation in somatic instability was explained by CAG repeat length and age at sampling, explaining 90% of the variation.

For somatic mosaicism, each expanded allele contains molecules of different lengths and each may have a different value of the estimated inherited allele (n). Amplified PCR slippage molecules will generate molecule of n-1, n-2, n-3, etc, for each template molecule. The overall profile from a person will consist of all profiles added together. For a patient, n-1, n-2, etc. reads for estimated (n) will also include n-1, n-2, n-3 reads from other longer expansions generated by somatic instability. Therefore, the number of n+1, n+2, n+3, etc. reads for the distribution will be underestimated due to backward slippage events. Also, the n reads is also underestimated. To date, we have not attempted to correct for that effect.

We also studied the association between age at disease onset and the CAG repeat length in the HD patients. Our results indicate there is a significant role for expanded CAG repeat length in modifying age at disease onset. This association appears to fit with a linear regression model, as there was no evidence for improved fit using a non-linear model such as quadratic and exponential. These data revealed that allele length explains 55% of the variation in age at onset. This suggests that other genetic and environmental factors may determine the remaining variation in age of disease onset, and may modify disease outcome. Our result has emphasised a significant role of the expanded CAG repeat in modifying age at onset, indicating that expanded allele length is the initial trigger of HD pathogenesis and the predominant factor determining the rate of the processes that lead to disease onset.

Here, we have estimated the allele length and evaluated repeat dynamics in buccal cell DNA using NGS technology. We can confirm buccal DNA can be used in a detailed analysis of somatic instability and genotype-phenotype correlation, as it is a good source for sizing the CAG repeat and estimating somatic instability. Although we are evaluating the somatic instability in peripheral cells, it is important to consider the repeat length will be much larger in the affected tissues than the periphery. An analysis of DNA from buccal cells is a good source to test large numbers of samples from living individuals for measuring somatic instability. Nonetheless, somatic instability has been reported to be low in buccal cell samples for HD patients using the single molecule PCR approach (Veitch *et al.*, 2007). Although somatic mosaicism in buccal cells was very low in

that study, they confirmed the allele length is a major determinant of variation in somatic instability. We therefore rationalise the established utility of using high throughput approaches to quantify somatic instability and display a great understanding of the genotype-phenotype relationship in HD.

There has been much debate in the field whether age at onset and disease severity are influenced by somatic instability. This is a critical question, as the answer determines whether the pathogenic effect of expanded alleles can be modulated by CAG repeat instability. Therefore, we evaluated the residual variation of somatic instability not explained by allele length and age at sampling that is likely reflecting the specific mutation rate variation between individuals. We expect those individual specific differences will be associated with disease severity and the somatic instability will have an influence on age at onset. In disagreement with previous observations of repeat instability and age at onset of Myotonic dystrophy type 1 (DM1) patients, we did not find any evidence that the residual variation in somatic instability is an explanation for any portion of the variance in age at disease onset that is not explained by the allele length (Morales *et al.*, 2012). The evidence from the DM1 study suggests that the residual variation of somatic instability is individual specific and heritable as a quantitative trait and that residual variation is associated with residual variation in age at onset in DM1 patients. These data revealed the somatic instability accounted for some of the variation in age at onset and the repeat expands more rapidly for individuals who have an earlier age at onset in DM1 patients. However, it is important to note that correlation between residual variations of somatic instability and age at onset is relatively low (r^2=0.068). That implicates the role of individual specific environmental or genetic factors as a modifier of the individual-specific differences in the level of somatic instability. Furthermore, it has been reported that CAG repeat expansion in postmortem DNA is associated with an earlier age of disease onset, suggesting that somatic instability is a significant predictor of the age of onset and disease progression (Swami *et al.*, 2009). Although the published data implicating somatic mosaicism as an important component of the disease pathway are compelling and that somatic instability plays a role in age at onset, we could not replicate these findings in our study.

This non-significant correlation is supported by the absence of positive correlation of age at sampling on somatic instability, or inherited repeat length variation on age at onset in this Venezuelan cohort. This could also result from a relatively low number of samples in this analysis and the difficulties of precisely assigning age at onset to a patient. It is important to consider the measurement of somatic instability from the non-target tissue is likely underestimating the real repeat dynamics in affected tissues.

Nonetheless, the great part of the variation in the association between these two residual variables is likely explained by unknown environmental and genetic factors. The remaining variance has been reported to be heritable indicating the presence of genetic modifiers, which may have an effect on the disease onset and progression of HD. These genetic factors may play a role in repeat instability and could act either in *cis* or *trans*. Although *cis*-acting modifiers can influence somatic instability, sequences close the CAG repeat are unlikely to be responsible for differences in somatic instability in the Venezuelan kindred, as the HD chromosomes derive from the same founder chromosome (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004). Therefore, it is more likely these genetic modifiers act in *trans* influencing somatic instability in these patients. This has been suggested in a previous study with the limited number of patients with HD from the Venezuelan cohort (Leeflang *et al.*, 1995). Additional studies confirmed the role of *trans*-acting modifiers in HD CAG repeat instability and HD CAG dependent phenotypes in mice (Wheeler *et al.*, 2003; Pinto *et al.*, 2013).

The most obvious candidates for *trans*-acting modifiers of somatic instability are the DNA mismatch repair genes. These modifier genes have been shown to be critical in generating somatic repeat expansion in mice. Deficiency of *Msh2*, *Mlh1* and *Mlh3* mismatch repair genes delay the disease pathology and completely suppress the somatic instability (Wheeler *et al.*, 2003; Pinto *et al.*, 2013). In humans, several analyses have identified DNA repair genes as a likely explanation of variable instability and disease onset and progression, such as *FAN1, MTMR10, RRM2B* and *PMS2* (GeM-HD Consortium, 2015; Bettencourt *et al.*, 2016). These modifier genes may modify the level of repeat instability and account for variable levels of and disease severity and differences in age at

onset between individuals. Such modifier genes would provide potential new targets for therapeutic interventions in HD that alter the HD disease in human patients delaying disease onset and progression. If somatic expansion contributes to disease, then a therapeutic approach could be possible by inhibiting the somatic expansion that occurs in the brain during the lifetime. Nevertheless, the absolute importance of somatic expansion in disease progression remains obscure. Understanding the repeat instability mechanisms in HD patients could ultimately provide the possible therapy aimed at preventing CAG repeat expansion.

Chapter 6 Testing candidate DNA repair genes as potential *trans*-acting modifiers of genetic instability in HD

6.1 Introduction

It is clear that most of the variation in age at onset is explained by the expanded allele size. Even though all HD patients have the same mutation, two individuals with identical CAG repeat are unlikely to present with symptoms at exactly the same age. The variability in age at onset and disease manifestation has suggested the existence of modifier factors, and the search for these factors has captured significant attention. These could theoretically be environmental or genetic factors.

The identification of disease-modifying genetic factors from human HD subjects can yield clues for therapeutic development aimed at delaying or preventing the disease onset. Specific genes, pathways, and processes offer a promise of being identified using modern genetic techniques and could permit the development of new therapeutic interventions that target these pathways. Nonetheless, investigating modifier genes may be limited by naturally occurring genetic variation in human populations. Other reasons for seeking to identify these modifiers is to permit more informative clinical trials by recruiting a more homogeneous patient population and reducing genetic variation between individuals.

The HD-MAPS (Modifiers of Age at onset in Pairs of Sibs) study recruited affected sibling pairs and revealed evidence for the heritability of residual variation in age at onset (Li *et al.*, 2003). They suggested that 56% of the variation in age at onset may be attributable to other genes capable of modifying HD pathogenesis. Subsequently, a similar study confirmed in the large HD Venezuelan Kindred that approximately 40% of the variability in age at onset after accounting for the effect of CAG repeat length was attributable to genetic factors other than the *HTT* gene (The U.S.–Venezuela Collaborative Research Project and Wexler, 2004).

Earlier studies on human genetic modifiers in HD have relied on a genome-wide genetic linkage of both sib-pairs and extended families to search for chromosomal locations, which are potentially harbouring modifiers (Li *et al.*, 2003, 2006; Gayán *et al.*, 2008). The genomic regions implicated in these studies are large and have not led to identification of specific genes or sequence variants responsible for the variability in age at onset. Notably, genome wide genetic linkage studies detected multiple loci in distinct population samples. Also, some studies attempting to identify these genetic modifiers were small candidate gene studies. These studies were all superseded by the one large genome-wide association studies (GWAS) of HD modifiers that is an unbiased approach to identify naturally occurring genetic variations that modify the disease process. This study performed by the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, discovered significant modifier associated with modification of age at onset of the motor sign (The GeM-HD Consortium, 2015). GWAS was carried out on 2,131 individuals of European ancestry who carried between 40 and 55 CAG repeats. HD subjects were genotyped with the Illumina Omni2.5 array. Analysis of the GWAS data revealed a genome wide significant level in one gene on chromosome 8 and one on chromosome 15, with two independent signals at the same locus. This study identified new genetic modifiers that are located in the *FAN1* and *MTMR10* regions on chromosome 15 as well as the *RRM2B* gene on chromosome 8. In addition, a further interesting locus on chromosome 3 is the *MLH1* gene. Although this locus did not achieve genome wide significance, *MLH1* was revealed as a potential modifier. This locus is particularly attractive given that *MLH1* is known to regulate somatic instability of CAG repeats and is a modifier of the phenotype in mouse HD knock-in models (Pinto *et al.*, 2013). This study supports *MLH1* as a candidate gene that modifies HD pathogenesis and implicates DNA mismatch repair as a process of HD modification.

Interestingly, none of the most significant modifier genes in chromosome 15, 8 or 3 from the GWAS analysis corresponds to any suggested modifier genes or linkage regions identified in previous studies in HD subjects (Li *et al.*, 2003, 2006; Gayán *et al.*, 2008). The discrepancy between GWAS and the Venezuela linkage study may be a consequence of genetic heterogeneity, (in which mutations at different loci are responsible for the disease phenotype), and

therefore different modifier alleles, reflecting the different ethnic background in these populations. In addition, the increased sample size in recent GWAS delivered a more accurate assessment of the relationship between allele length and age at onset, yielding residuals that more accurately account for the effect of repeat length on age at onset.

Furthermore, pathway analyses from GWAS that examined whether variations associated with the age at onset cluster in genes with common biological function showed that variants in genes involved in DNA repair pathways associated with residual variation in age at onset (The GeM-HD Consortium, 2015). Those genes in the DNA repair pathway were replicated and validated in a separate study of HD and polyglutamine spinocerebellar ataxia (SCA) patients (Bettencourt *et al.*, 2016). They tested the association of DNA repair genes with residual variation in age at onset. The study involved 1,462 patients of HD and SCAs type 1, 2, 3, 6, 7 and 17. SNPs were chosen from the most significant SNPs ($p < 0.1$) selected from the GeM-HD study. All these SNPs cluster in DNA repair pathways. In addition, SNPs from *RRM2B* and *UBR5* genes were added to the study, as they are members of DNA repair pathway that were marginally significant in the GeM-HD analysis, but those genes have significant genome wide P-value (Table S5 of the GeM-HD article). The most significant SNPs were chosen for each gene as well as a small number of proxy SNPs in close LD (r^2 >0.8) with these most significant SNPs in the GeM-HD study. SNP analysis, which tested the effect of 22 SNPs on age at onset, was examined. The analysis of the combined samples for all polyglutamine diseases (HD and SCAs) yielded a significant association at *FAN1* and *PMS2* and for all the SCAs patients yielded association with *PMS2*. These genes' involvement is likely explanation of modifying disease onset and progression in HD. This study suggested that modulation of somatic instability may also be important in genetic variation that alters age at onset of polyglutamine diseases.

The data from the GWAS showed that genes involved in DNA mismatch repair are modifying age of onset in HD patients. Studies in HD mouse models have revealed that DNA mismatch repair genes are required to generate somatic expansions (Wheeler *et al.*, 2003; Pinto *et al.*, 2013). These have been carried out by an elimination of HD CAG repeat instability in Hdh knock-in mice. Hdh

knock-in mice were crossed with mismatch repair genes knockout mice to generate mouse models that have a targeted mismatch repair inactivating mutation and test the relative contribution of these genes in the mouse models. The effect of the mismatch repair genes knockout is to eliminate somatic expansion. The data have shown the genetic variations on *MSH2*, *MLH1* and *MLH3* mismatch repair genes in HD mice models underlie differences in CAG repeat somatic expansion. The deficiency of these mismatch repair genes suppresses the somatic instability and delay the phenotype (Wheeler *et al.*, 2003; Pinto *et al.*, 2013). These data support the role of the mismatch repair pathway in somatic instability.

In addition, several analyses have identified DNA repair genes as a likely explanation of variable instability and disease progression in human HD (GeM-HD Consortium, 2015; Bettencourt *et al.*, 2016). There is also evidence from DM1 patients for an association between variants in the *MSH3* gene and variation in the degree of somatic instability (Morales *et al.*, 2016).

Therefore, a plausible mechanism to explain the possible mechanism by which DNA mismatch repair genes modify HD is that they act through modifying the rate of somatic expansions in HD patients. Thus, reduced activity of the DNA mismatch repair pathway would be expected to reduce the severity of HD symptoms. Therefore, we sought to investigate whether the most significant SNPs reported in GWAS also have a modifier effect on somatic instability in HD. In this regard, we analyzed SNPs from both GeM-HD and the follow-up study (Bettencourt *et al.*, 2016) to evaluate whether these variants modify the somatic instability.

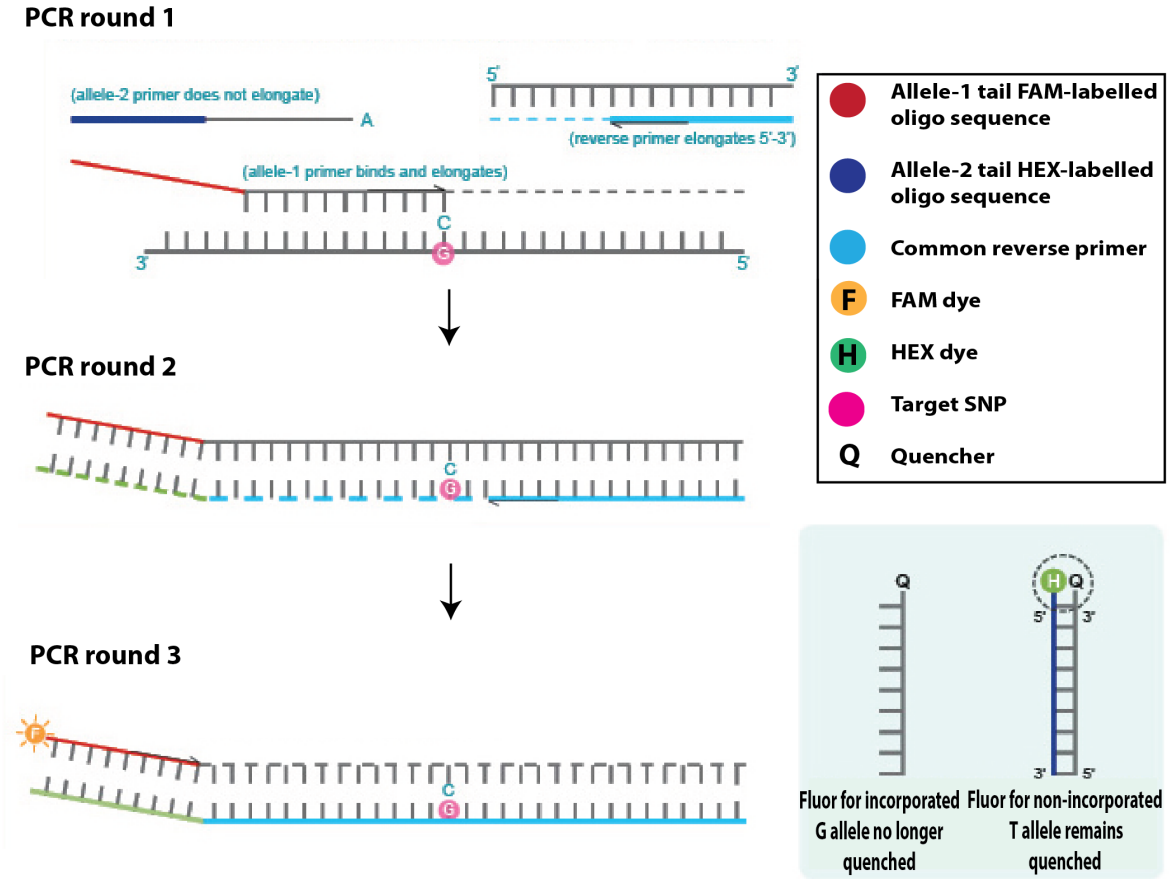

Figure 6-1 KASP assay principle. In the first round of PCR, one of the allele-specific forward primers matches the target SNP with common reverse primers to amplify the target region. The 3 prime parts of both forward primers are complementary to the target DNA and differ in the 3' end of each SNP. Each forward primer also includes unique additional nucleotides that are not complementary to the target DNA. These nucleotides are referred as the tail sequence at the 5' end that an important role for signal generation by hybridizing to either the HEX or FAM fluorophore. At the first round of PCR, the tail sequences will have been incorporated into the PCR product. During PCR round 2, the complementary sequences to the allele-specific tail sequences are first generated. In the third round of PCR, the presence of tail sequences allows the FAM labelled oligonucleotides to bind. The FAM labelled oligo binds to new complementary tail sequences and is no longer quenched and emits fluorescence. After a further round of PCR, the level of allele-specific tail increases. The fluor labelled part of the FRET cassette is complementary to new tail sequences and binds, releasing the fluor from the quencher to generate a fluorescent signal. (Modified from https://www.lgcgroup.com/LGCGroup/media/website-content/Products/Genotyping/KASP/how-does-kasp-work.jpg).

Given the progress in high throughput SNP genotyping available, several variables need to be considered including throughput, cost, turnaround time, sensitivity, reliability, flexibility and number of SNPs genotyped per run. The candidate genes approach is more suitable for our experiment than GWAS, as we aimed to genotype a relatively small cohort and thus less powered by GWAS. LGC genomics has developed high-throughput technology named Kompetitive Allele Specific PCR (KASP) (www.lgcgenomics.com). KASP was used to validate some of the candidate genes for HD and SCA modifier genes that were identified by

GWAS (Bettencourt *et al.*, 2016). The advantage of using KASP over other methods is the ability to design an assay to genotype virtually any SNP or insertion and deletion that makes KASP a very flexible method.

The KASP assay utilizes a homogeneous fluorescence-based genotyping system (Semagn *et al.*, 2014). The KASP assay uses allele-specific oligo extension and fluorescence resonance energy transfer (FRET) for signal generation. KASP can distinguish between two alleles for each SNP. For each SNP, one common reverse primer paired with one of the two specific forward primers (one for each SNP allele) designed for KASP assay are used (Figure 6-1). Each forward primer has a nucleotide sequence that is specific to one of the labelled fluorophores. As PCR proceeds, one of the fluorophores, corresponding to the amplified allele, is incorporated into the template DNA, and it is no longer bound to its quencher complement. As the fluorophore is no longer quenched, the fluorescent signal is generated. This fluorescence is detected at the end of the assay using a qPCR machine or a FRET capable plate reader. Genotyping of the samples is obtained from the proportion of fluorescence from HEX, FAM, or both. If the genotype at given SNP is homozygous, one of the fluorescent signals will be generated. If the individual is heterozygous, the result will be a mixed fluorescent signal. The KASP chemistry can be performed in 96, 384 and 1,536 well plate formats. KASP genotyping is available as reagent kit to use in any labs and as a genotyping service through LGC labs in North America and Europe. This technology provides cost-effective and greater flexibility for researchers in generating data for 1 to thousands of SNPs for over thousands of samples able to be analyzed (Semagn *et al.*, 2014).

The aim of this study was to develop KASP-based assays to genotype candidate modifier genes of genetic instability in the Venezuelan cohort of HD patients. Our hypothesis was that the variants in DNA repair genes might modify CAG instability of HD and subsequent severity.

6.2 Materials and Methods

6.2.1 HD allele genotyping and measurement of somatic mosaicism

The CAG repeats were amplified from each individual from buccal swab DNAs from the US-Venezuelan Collaborative Research Project. DNA samples were genotyped using PCR amplification with MiSeq adapter primers that flank the CAG·CTG repeat tract of the *HTT* locus, followed by sequencing the products using the Illumina MiSeq platform as described in Chapter 3. We also used MiSeq read length distributions to measure the degree of somatic instability of the expanded CAG repeats for each individual using the previously described method as in Chapter 5. For each individual, the residual variation in somatic instability that was not accounted for by CAG repeat length, age at sampling and their interaction, was determined. This individual-specific difference in somatic instability might be attributable to genetic modifier genes. We included 412 HD patients, 59 unaffected individuals and 20 control samples with known high-quality blood DNA samples for genotyping.

6.2.2 SNP selection criteria

SNPs were selected from the genes most significant by associated ($P < 1 \times 10^{-6}$) with residual variation in age at onset from the GeM-HD study (see Table 1 of the GeM-HD article). Additional SNPs in the DNA repair pathways were selected from a follow-up study of GeM-HD analysis (Table S4 of the GeM-HD article) (Bettencourt *et al.*, 2016). In this regard, we selected 31 SNPs from this study to evaluate whether these loci are associated with HD somatic instability. Also, we chose 9 SNPs out of 11 from the GeM-HD study that are most significant by associated residual variation of age at onset (see Table 1 of the GeM-HD article). Two SNPs were in common between these two studies and were included once. SNPs in the *MSH3*, *MSH6*, *PMS1*, *PMS2*, *MLH1*, *MLH3*, *FAN1*, *RRM2B*, *UBR5* and *LIG1* candidate modifier genes were tested in our analysis.

6.2.3 KASP genotyping

SNP genotyping was performed using custom KASP assays through LGC Genomics labs in Hoddesdon, UK. This service includes assay validation with their samples

and also our submitted or supplied DNAs, and optimization of assay conditions and primer sequences. LGC service lab runs customer samples and provides genotyping data. DNA samples were assembled in a 96 well plate. We included two no-template controls on each genotyping plate to improve confidence in the validity of the genotyping results. Also, positive controls were added to each plate for assay validation, but these controls were of unknown genotypes, but with good DNA quality and high concentration compared to the buccal swab DNAs.

The assay for several SNPs was designed in reverse orientation to the chromosome (rs1037700, rs1037699, rs3512, and rs20579). For this reason, for all SNPs in reverse orientation, genotypes resulting from these KASP assays will be complementary to those using HGVS nomenclature. This is reflected where the minor allele for these SNPs differs from GeM-HD but corresponds to the same allele (Table 6-1).

Primers were designed by LGC Genomics based on the SNP locus sequence, and KASP assay carried out according to the company's protocol (http://lgcgenomics.com). Following KASP PCR, plates were read, and the genotyping data analysed using cluster analysis viewing software. For our data, cluster analysis was performed using SNP viewer software that enables genotyping and cluster plot data to be viewed by the plate (https://www.lgcgroup.com/products/genotyping-software/snpviewer/snpviewer-help/#.WfxP7hS3mOp). The detected signal from each sample is represented as an independent data point on the cluster plot. The fluorescent signals from samples with the same genotype cluster together on the cluster plot (Figure 6-2). One axis is used to plot the FAM fluorescence value and the second axis is used to plot the HEX fluorescence value for each sample. Homozygous samples for the allele reported by FAM will produce FAM fluorescence only during the reaction, and that data point will be plotted close to one of the axes. In the same way, a sample that is homozygous for the allele reported by HEX will produce HEX fluorescence only during the KASP reaction. Heterozygous samples contain both alleles reported by FAM and HEX fluorescence and generate half as much FAM and HEX fluorescence as the samples that are homozygous for these alleles. This data point will be plotted in

the centre of the plot, representing half FAM and half HEX signal. No fluorescence will be generated for non-template controls, and the data point will be plotted in the origin. We can determine the genotypes of all samples based on the relative position of each cluster.

6.2.4 Statistical analysis

The somatic instability was corrected for repeat length, age at sampling and their interaction for each individual, which was performed using linear regression as described in Chapter 5. The residual variation in age of onset not accounted for by allele length was also estimated for each individual (as described in Chapter 5). This residual variation of somatic instability and age at onset provides a phenotype for the SNP analysis. Those residuals were estimated using SPSS statistics software (IBM).

Allele frequencies, Hardy-Weinberg equilibrium tests, linear regression analyses and family-based association study were performed using gPLINK (version2.050) (http://zzz.bwh.harvard.edu/plink/gplink.shtml). The gPLINK is a Java-based program that provides graphical user interphase for the commonly used PLINK software (command line options) (Purcell *et al.*, 2007). The gPLINK programme offers a simple, user-friendly tool for performing a wide range of basic and large-scale genetic analyses. The software includes only common PLINK commands, however it is possible to create any PLINK command that has not been incorporated into gPLINK. A regression analysis was performed to test the association of the residual variation of somatic instability with each SNP genotype. Also, we did a family-based quantitative trait association analysis (QFAM) that combines a linear regression of phenotype on genotype with a permutation test, which accounts for relatedness among family members. P-values <0.05 were considered statistically significant. The analysis tests whether there was an association of residual variation in somatic instability across all SNPs. For multiple corrections, we used the Bonferroni multiple significance tests. We compared the significance to that obtained from previous analysis in order to assess the association of these SNPs with the disease phenotype.

6.3 Results

6.3.1 Pilot study

The DNA concentration was quantified for a few samples from the Venezuelan cohort. We found their concentrations varied from 0.5 to 20 ng/μl. Some of the buccal swab DNA samples are at a lower concentration than the recommended DNA concentration (5 ng/μl) that is required for KASP genotyping. Therefore, a pilot study was necessary to test the probability of the buccal swab DNA samples from HD patients of Venezuelan cohort being genotyped when their concentrations are low.

LGC recommended conducting this pilot study that includes a minimum of 22 buccal swab DNA samples and two positive controls of known high-quality DNA. These samples were run by LGC lab to enable efficient cluster analysis for this study. These 22 samples were selected randomly to be representative of all samples in the cohort. Genotyping 31 SNPs in 24 samples was conducted by LGC lab by sending 95 μl DNA for each sample. Of the 31 SNPs used for genotyping, 19 SNPs (61%) were validated in our buccal DNA samples (Table 6-1). Although with LGC control samples, 26 SNPs were validated, 7 SNPs could not be validated in our samples due to these SNPs being sensitive to low DNA concentrations. The 19 validated SNPs were used to genotype all the DNA samples in the Venezuelan cohort.

	SNP ID	Minor allele	Major allele	SNP to chromsome	Sequence of KASP assay design
1	rs1037699	G	A	Reverse	TGTCCGCCCGCCCTC[A/G]CCGCAGCCTGGCTT
2	rs1037700	G	C	Reverse	CGGGGTGAGACTTAC[C/G]CCTGCGTTTATCCG
3	rs144287831	C	T	Forward	CATGGTGAAACCCCA[T/C]CTCTACTAAAAATA
4	rs146353869	C	A	Forward	ATTAAAATGTGAATC[A/C]CAAGAGTGATGTGT
5	rs148491145	-	GACTCTA	Forward	TGGTAGCTGAATCCT[GACTCTA/]GAATATTACCACAT
6	rs150393409	G	A	Forward	AATTGGCCAAACAGC[A/G]TTCAGTCTGCACTT
7	rs16869352	T	C	Forward	TAGCAATGCTTGGAA[C/T]ACACGCTTGCATTT
8	rs175080	G	A	Forward	CTTTCTCTCAAACTA[A/G]GCATCTGTTGTTCT
9	rs1799977	G	A	Forward	GACAATATTYGCTCC[A/G]TCTTTGGAAATGCT
10	rs1800937	T	C	Forward	GGTAGGCACAACTTA[C/T]GTAACAGATAAGAG
11	rs1805323	C	A	Forward	GACCCAGTGACCCTA[A/C]GGACAGAGCGGAGG
12	rs20579	T	C	Reverse	AAGGGAGAATTCTGA[C/T]GCCAACATGCAGCG
13	rs3512	G	C	Reverse	TTAAAAGTAAAGGCA[C/G]TTCCAAGAGTAACA
14	rs3735721	G	A	Forward	GCTTAGTTGTAAGAA[A/G]AACTATTATTGTAT
15	rs5742933	G	C	Forward	GCCTCGCGCTAGCAG[C/G]AAGGTAGTGTGGTG
16	rs5893603	G	-	Forward	CCGGGGCAGAGCAGC[/G]GAGCGGGACGCAAA
17	rs6151792	T	C	Forward	ATGTAAAATTAGGCC[C/T]GCAGACAATTCGAA
18	rs61752302	T	C	Forward	TATAAAATGAGCATT[C/T]GCCTTTCGATCCTT
19	rs72734283	G	A	Forward	GATTTGACCTTGACA[A/G]CCCATCTAGCCAAC

Table 6-1 SNP sequences for SNP KASP assay design.The genotypes for SNPs in reverse orientation to chromosome given by KASP assays are complementary (reverse) to HGVS nomenclature.

6.3.2 Genotyping candidate genes

We tested 19 markers for the KASP assay that have been validated through LGC lab. The analyses comprised 412 patients' samples (229 females, 168 males and 15 unspecified) that were chosen for this study. An additional 59 unaffected individuals and 20 controls from Scottish population were also genotyped. These HD subjects were members of 130 families used in this study. The family size ranged from 1 to 20 individuals with an average of 3.6, consisting of 44 (33.8%) families of 1 individual, 34 (26.2%) families of 2 individuals, and 17 (13.1%) families of 3 individuals. Families with 1 individual represent one affected individual data point with missing parental genotypes. The remaining 35 (26.9%) of families have 4 or more individuals, in which 7 (5.34%)of these families have 10 to 20 individuals. Those affected individuals selected for the study have one HD allele of 40-50 CAG repeats and known residual variation of somatic instability measure.

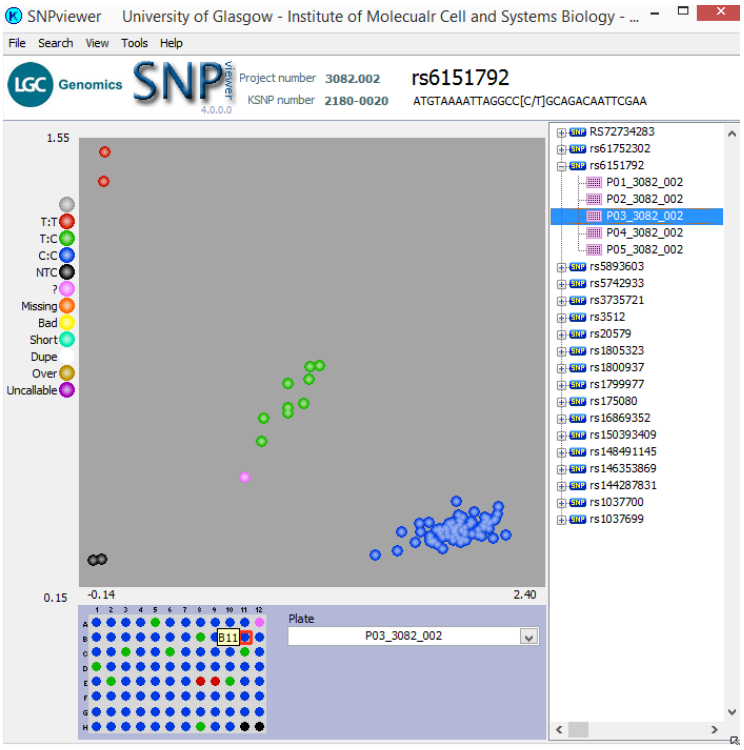

Figure 6-2 SNP viewer of KASP assay results for rs6151792 on one of the plates that contains 94 samples. Genotypes with a "C" are represented by blue dots across the bottom right cluster and those with a "T" by red dots in the upper left cluster. The data points in the centre of the plot, representing C:T heterozygous genotypes are represented as green dots. The two black dots near the bottom left are negative controls that have no fluorescence generated. The pink dot is a sample with unknown genotype.

For each SNP, three primers were designed, and two alleles were labelled with either FAM or HEX. The SNP sequences used for designing primers are provided in Table 6-1. An example genotyping result of a KASP assay using SNP viewer software is illustrated in Figure 6-2. The KASP assay was performed on 94 genotypes from our data illustrating the T/C SNP in three distinct clusters. As a result, the blue or red or green coloured dots represent the genotypes of each sample in the plot. For the rs6151792 marker, samples homozygous for the major allele (C) are shown in blue dots, while the red dots are samples homozygous for the minor allele (T) and C:T heterozygotes are shown in green dots. The no template controls were included across several plates of samples submitted for genotyping. These controls are expected to have no signal or very weak signal. Our results showed that the no template DNA controls clustered together and rarely produce signals.

The fluorescent signal from each individual DNA sample is represented as an independent data point on the cluster plot. Each data point represents one

sample genotyped for one SNP. Therefore, genotyping of 491 samples with 19 SNPs results in 9,329 data points. However, a proportion of the data points reveal unknown or unassigned genotypes. The frequency of these unknown genotypes is 206 data points (2.2%), 9,123 data points (97.8%) were informative for further analyses. Some samples were not scored for different SNPs due to either a weak amplification or uncertainty in differentiating the homozygote and heterozygote genotypes.

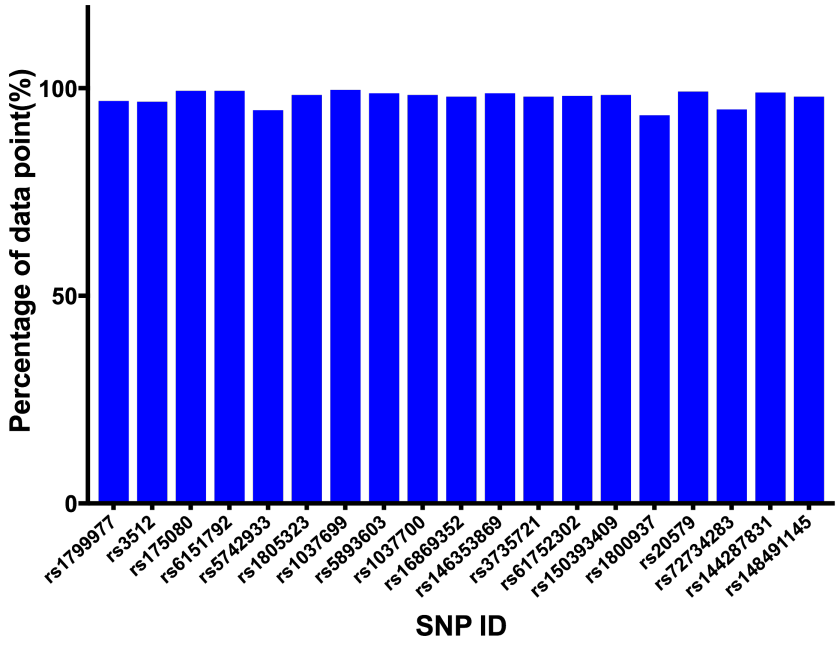

Figure 6-3 KASP genotyping data for 491 samples for 19 SNPs.The analysis of data points obtained for each SNP showed that between 93% to 99.5% of samples were genotyped for each SNP.

We also checked the genotyping success of samples and SNPs. In general, genotyping was of good quality and consistent within the families. Eighty-four % (16/19) of SNPs had data from 97% of samples (Figure 6-3). The number of genotypes obtained for all SNPs ranged from 93.5% to 99.5%. The overall genotyping efficiency of these patients was 97.8%.

For five SNPs (rs1805323, rs61752302, rs148491145, rs150393409, rs146353869) no individual tested was homozygous for the minor allele. Of these SNP targets, one SNP (rs144287831) was homozygous for all the genotyped HD patients, i.e. monomorphic; the rest were polymorphic between individuals. The

monomorphic SNP was removed from further analysis because there is no power to detect association with such variants.

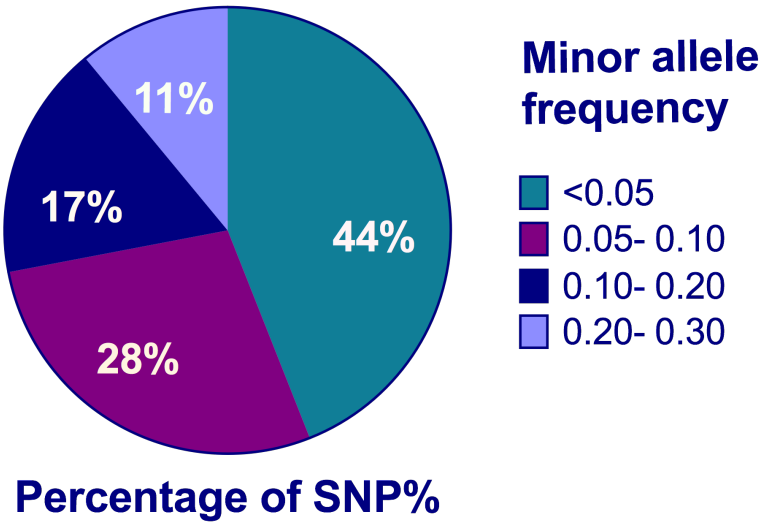

Figure 6-4 The frequency of minor alleles of 18 SNPs tested in HD patients. The number of SNPs for each MAF range is plotted, and the MAF ranged from <0.05 to 0.30. A large proportion of SNPs have a MAF of <0.10. 11% of SNPs were very common (MAF > 0.20). Many (44%) rare SNPs were also detected (MAF<0.05).

The frequency of each SNP is given in terms of the minor allele frequency (MAF). For each SNP, the MAF was determined from all samples. The characteristics of SNPs used in our study are shown in Figure 6-4. MAF ranged from 0.004 to 0.292 in HD patients and the majority of SNPs have MAF of <0.10. We found many rare SNPs (44% of the 18 SNPs) having a MAF of < 0.05 and also 28% of SNPs with a MAF between 0.05 and 0.10 (Figure 6-4). We also detected two of 18 SNPs (11%) which were common with a MAF > 0.20. The remaining 17% of SNPs have a MAF in the range of 0.10-0.20. These common SNPs comprise the majority of heterozygous individuals in the population. The presence of common SNPs within the study is important to find the alleles specifically associated with the disease phenotype.

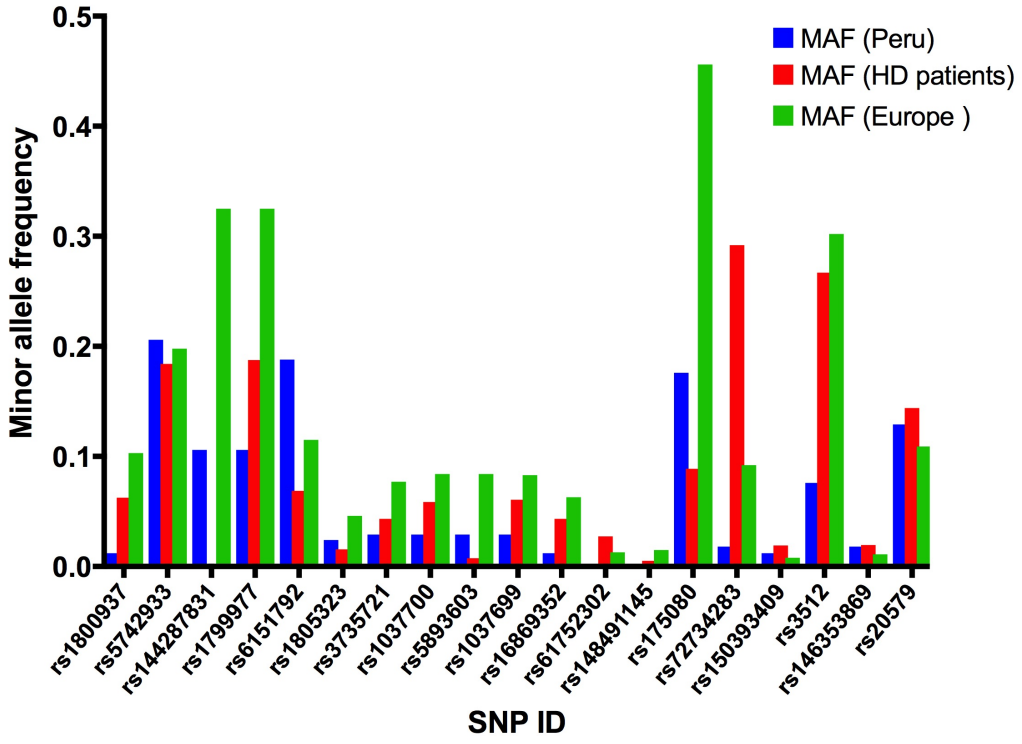

Figure 6-5 Minor allele frequency (MAF) comparison of 19 SNPs between HD patients from Venezuelan, European and Peruvian populations.Minor allele frequency for HD patients in our study was compared to MAF in European and Peruvian populations from 1000 genome project data.

The obtained minor allele frequency from our samples was compared with the 1000 genome project data for Peru, which we expect to be similar genetically to Venezuela, and also allele frequency in Europe. The distribution of minor alleles frequencies of each SNP in the Venezuelan, Peruvian and European population is shown in Figure 6-5. Minor allele frequencies of all SNPs were broadly similar across the Venezuelan and Peruvian population. However, it is notable the Venezuelan population had a significantly higher MAF in rs72734283 and rs3512 compared to Peruvian. Significant differences in allele frequency of these two SNPs could be due to ethnic differences or the differences in sample size in determining MAF in these populations (MAF of Peruvian was estimated from 1000 genome project study from 85 individuals). There were substantial ethnic differences in MAF across all three populations; European population had a higher MAF (0.456) in rs175080 than the Venezuelan (MAF=0.0887) and Peruvian populations (MAF=0.176). The minor allele was not detected for rs144287831 from HD patients of the Venezuelan population. When compared this SNP's MAF in the Peruvian (0.106) and European population (0.325), MAF of rs144287831 is

high in other populations and is expected to be detected in the Venezuelan population. These findings suggest the lack of variation might be attributed to an error in genotyping. Therefore, this SNP was eliminated from the further studies.

SNP ID	Chr:position (bp)	Gene	Functional annotation	MAF (HD patients)	P value HDW (HD patients)	MAF (unaffected individuals)	P value HDW (unaffected individuals)	MAF (controls)	P value HDW (controls)
rs5742933	2:189784590	*PMS1*	5′ UTR variant	0.184	0.366	0.213	1	0.3	1
rs144287831	3:37026588	*MLH1*	Intron variant	0	1	0	1	0	1
rs1799977	3:37012077	*MLH1*	Missense variant	0.1875	0.242	0.0877	1	0.194	1
rs6151792	5:80761142	*MSH3*	Intron variant	0.0687	0.106	0.042	1	0.15	1
rs1805323	7:5987311	*PMS2*	Missense variant	0.0156	1	0	1	0.475	0.0001
rs3735721	8:102205467	*RRM2B*	3′ UTR variant	0.0433	0.202	0.069	1	0.025	1
rs1037700	8:102238547	*RRM2B*	Intron variant	0.0586	0.352	0.144	1	0.05	1
rs5893603	8:102238611	*RRM2B*	Frameshift variant	0.0076	0.004	0	1	0.05	1
rs1037699	8:102238702	*RRM2B*	Missense variant	0.0606	0.383	0.068	1	0.05	1
rs16869352	8:102293805	*UBR5*	Synonymous variant	0.0433	0.202	0.068	1	0.05	1
rs61752302	8:102298925	*UBR5*	Synonymous variant	0.0273	1	0.026	1	0	1
rs148491145	14:71893459-71893465	LOC105370558	Intron variant	0.004	1	0	1	0.05	1
rs175080	14:75047125	*MLH3*	Missense variant	0.0887	1	0.314	0.544	0.475	1
rs72734283	14:75028356	*MLH3*	Intron variant	0.292	0.402	0.064	1	0.125	1
rs150393409	15:30910758	*FAN1*	Missense variant	0.0191	1	0	1	0	1
rs3512	15:30942802	*FAN1*	3′ UTR variant	0.267	0.371	0.232	1	0.425	0.354
rs146353869	15:30834198	*FAN1*	Intron variant	0.0195	1	0	1	0	1
rs20579	19:48165573	*LIG1*	5′ UTR variant	0.144	0.731	0.161	0.33	0.075	1

Table 6-2 Characteristics of single nucleotide polymorphisms (SNPs) used in our study. Genes annotated by the SNPs are indicated. Minor allele frequency and P value for Hardy–Weinberg equilibrium were estimated for HD patient, unaffected individuals and control samples in our study. Chr = chromosome; MAF = minor allele frequency; HWE = Hardy–Weinberg equilibrium.

We tested all SNPs to see whether they fit the expected genotype frequency distribution under the assumption of Hardy-Weinberg equilibrium (HWE). HWE is a mathematical equation that can be used to determine if allele and genotype frequencies in a population will not change from generation to generation. Since gPLINK software considers only founders for the analysis and therefore no HWE result would be given for only sibling dataset (where parental genotypes are unknown), a single individual was selected from each family pedigree of the 131 families to test for HWE accurately. These individuals were selected from each family for either parental genotypes or one of the siblings if the parental genotypes are missing.

The observed genotype frequencies of most of SNPs were in Hardy–Weinberg equilibrium except, rs5893603 (P=0.004) (we considered a p-value >0.01 significant for our data set) (Table 6-2). In our data, rs5893603 has a very low minor allele frequency (MAF=0.007) in HD patients compared to the Peruvian population (MAF=0.029) and also this minor allele was not detected in unaffected individuals from the Venezuelan population (MAF=0). It is unlikely to detect any association with a very rare variant; therefore this SNP was eliminated from any further analysis. The genotype frequencies of all SNPs were in Hardy–Weinberg equilibrium for unaffected individuals. However, rs1805323 was not in HWE for control samples, this could result from the DNA for the controls being of good quality and at a higher concentration than the buccal DNA for unaffected and affected individuals from the Venezuelan population and subsequently DNA will amplify at a slightly different rate and generate a different level of a fluorescent signal. Therefore, inconsistent DNA quantity and quality across the plate might cause the heterozygous group (green dots) to migrate toward the homozygous group (red dots) and the apparent heterozygous in the genotyping plot (Figure 6-6). Also, the position of the homozygous control sample (blue dot) is skewed, making genotype scoring difficult. This SNP (rs1805323) was excluded from the analysis because the genotypes were ambiguous.

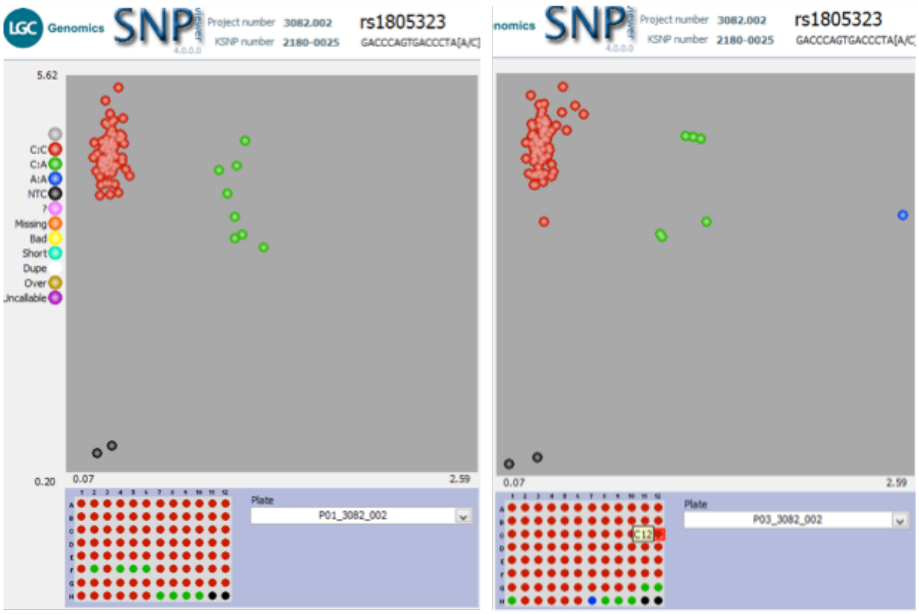

Figure 6-6 KASP assay result for rs1805323 on two plates, each contains 94 sample. Genotype with a "C:C" are represented by red dots across the upper left cluster and those with C:A genotypes in the middle representing heterozygous genotypes that are presented as green dots. The blue dot in the plot represents A:A genotypes. The two black dots are negative controls that have no fluorescence generated. The control samples have a higher concentration than the affected samples DNA that might cause the background signal from the DNA that causes the apparent heterozygous in the genotyping plot.

6.3.3 Identification of SNPs associated with somatic instability

The resulting genotypic data were used for this study to examine the effect of these SNPs on somatic instability. The respective genotypes were determined in all HD patients. In order to identify a possible modifying effect of SNPs on the somatic instability of the analysed HD patients, we determined the standardised residual variation of somatic instability for each individual after accounting for the effect of allele length, age at sampling and their interaction. Regression analysis was performed using the somatic instability measure as a dependent variable and the size of CAG repeat, age at sampling and their interaction as an independent variable (as described in Chapter 5). This residual variation represents a quantitative measure to test association with the *trans*-acting modifier gene genotypes. The residual variation values represent a specific measure of the tendency of the repeat to expand or contract for each individual. This residual variation measure ranged approximately from -2.323 to 4.523. The standardised residuals are normally distributed with a mean of 0. Individuals with a standardised residual greater than 0 have higher than the average somatic instability and are expected to develop the symptoms earlier than the average

for their CAG repeat length. In the same way individuals with standardised residuals, less than 0 have less somatic instability than the average and might develop the symptoms later than the average.

Association of allele dosage of each SNP was tested first by performing a linear regression with the standardised residual variation of somatic instability in gPLINK. The analysis of phenotype-genotype association is a regression of phenotype on genotype that ignores family structure in the given samples and analyses the data as if all individuals are unrelated. Failure to account for relatedness among families in linear regression analysis might result in false positive associations. As the relatedness is known in our data, a family-based association study can be used. Although the main focus of gPLINK is for population-based samples, there is some support for family-based association tests with quantitative phenotype (called the "QFAM" test within gPLINK), which uses permutation to account for the individuals' relatedness (Purcell *et al.*, 2007). This test combines a simple linear regression of phenotype on genotype with a permutation test, which corrects for family structure. Genotypes are divided into between and within family components, and these components are permuted independently at the level of the family, and the association analysis is performed on the within-family component, between family component or their sum for a total association test to form a new score for each individual. The family structure and family-based association test are investigated using both parental genotypes. In cases where parental genotypes are not available, siblings are used. The programme breaks down each pedigree into nuclear families and classifies them as those where both parents have genotypes and those in which they are not known. This approach provides a phenotype/genotype association test that accounts for the relatedness between individuals.

Association testing used an additive genetic model of allele dosage in gPLINK as the default for analysis. We can also assess tests that assume dominant or recessive genetic models of the minor allele. The additive model assumes two copies of minor allele have twice the effect on the phenotype of having a single copy. The dominant model assumes that an effect of phenotype is related to the presence of at least one copy of the minor allele and the recessive model

assumes that effect of phenotype is related to having two copies of the minor allele. Each bi-allelic locus was encoded using an additive genetic model in which genotypes were converted onto an ordinal scale of 0 for homozygous reference, 1 for heterozygous and 2 for the homozygous variant.

SNP	gene	Regression with residual variation in somatic instability		Regression with residual variation in age at onset	
		Beta	p value	Beta	p value
rs1800937	MSH6	0.144	0.361	-0.2995	0.1698
rs5742933	PMS1	0.151	0.164	-0.1019	0.495
rs1799977	MLH1	0.0348	0.781	0.06885	0.7037
rs6151792	MSH3	0.186	0.314	-0.3102	0.2431
rs61752302	UBR5	0.299	0.252	-0.03327	0.9284
rs16869352	UBR5	0.225	0.256	-0.2877	0.3303
rs1037699	RRM2B	0.174	0.332	-0.07593	0.7621
rs1037700	RRM2B	0.187	0.298	-0.1013	0.6903
rs3735721	RRM2B	-0.0711	0.729	0.2267	0.4645
rs175080	MLH3	-0.0624	0.515	-0.2177	0.1207
rs72734283	MLH3	-0.0588	0.675	-0.08078	0.7957
rs148491145	LOC105370558	-0.507	0.457	0.5289	0.3981
rs3512	FAN1	0.249	0.009	0.2014	0.1501
rs150393409	FAN1	0.476	0.185	-0.5577	0.3185
rs146353869	FAN1	0.484	0.176	-0.5657	0.3167
rs20579	LIG1	0.169	0.203	-0.03291	0.8581

Table 6-3 SNP association of candidate SNPs with a residual variation of somatic instability in HD patients. Beta and P value of the association analysis of candidates SNPs with somatic instability are estimated.

We did family-based association tests with quantitative phenotype, which was a residual variation in somatic instability for each individual, to determine the effect of all SNPs on somatic instability. Quantitative phenotype data were analysed under an additive model using the QFAM module of PLINK. We set the number of permutations to 100,000 and ran QFAM as a total association that combined within and between family tests. QFAM use a permutation test to obtain an empirical P-value while maintaining for family structure. Among the 16 SNPs, only one variant, rs3512 SNP (P= 0.009) was significantly associated with somatic instability (Table 6-3). The remaining SNPs were not associated with somatic instability in the examined HD patients (Table 6-3). Furthermore, the

presence of each additional minor allele G allele of the rs3512 variant was associated with an increase in the residual variation of somatic instability in HD patients (Figure 6-7). The mean standardised residual variation in somatic insatiability for rs3512 was -0.129 for individual homozygous for C reference allele, 0.0164 for C/G heterozygous and 0.270 for individual homozygous for G allele.

To control for testing multiple SNPs, a Bonferroni correction was applied across all SNPs in our analyses and considered P values of less than 0.05 as statistically significant. The Bonferroni correction adjusts P values when several dependent or independent statistical tests are being performed on a single dataset. The detected association was not significant after Bonferroni correction for multiple testing of 16 SNPs; this SNP (rs3512) was not statistically significant (P= 0.152).

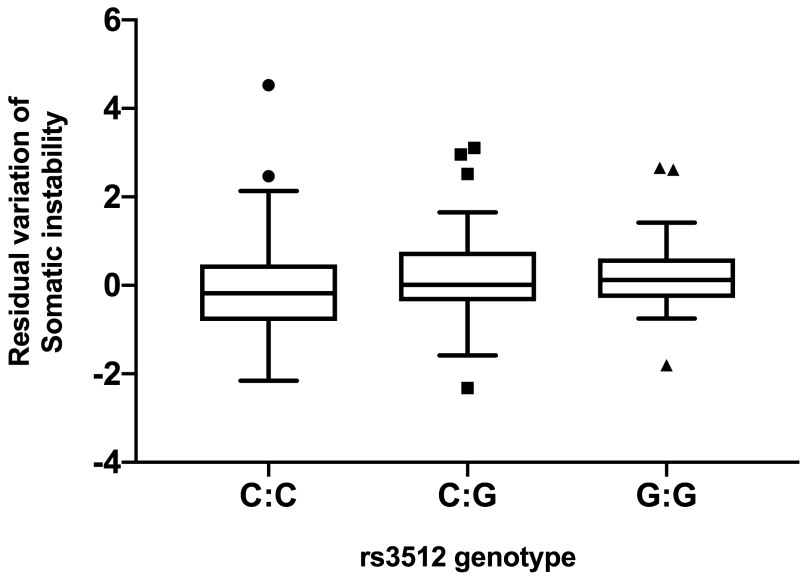

Figure 6-7 rs3512 genotype dependent standardised variation in somatic instability. Standardised residual variation in somatic instability that was corrected for by CAG repeat length, age at sampling and their interaction, was calculated. The box indicates the upper and lower quartiles of the standardised residual of somatic variation for each for rs3512 genotypes. The median is represented by a line dividing box. The lower values correspond to less somatic variation than expected and thus smaller residuals and the higher values correspond to more somatic variation than expected and thus higher residuals.

Although this cohort of HD patients is relatively large for a rare disease, it is nonetheless relatively small for detecting the contribution of modifier genes. Thus, in order to minimise the loss of power caused by a relatively small cohort and testing multiple SNPs, we selected a subset of SNPs for the analysis. Three

of these SNPs (rs1037700, rs1037699 and rs3735721) on chromosome 8 were found to be in high LD ($r^2 > 0.8$). Therefore, two of them were removed from further analysis. In addition, we excluded the SNPs that having a low minor allele frequency of <10%. Therefore, nine additional SNPs were removed from the analysis including the one SNP that was on LD on chromosome 8.The ability to detect association of rare variants (MAF <10%) is generally low. We restricted the analysis to SNPs that are not in LD ($r^2 < 0.8$) and also SNPs with high minor allele frequency (MAF>10%). This subset of data was selected for analysis because it might increase the significance of multiple corrections.

SNP ID	Gene	p value	adjusted p value
rs3512	*FAN1*	0.0091	0.0457
rs5742933	PMS1	0.1627	0.8137
rs20579	LIG1	0.2031	1
rs72734283	MLH3	0.6718	1
rs1799977	MLH1	0.7819	1

Table 6-4 Association between SNP genotypes and residual variation of somatic instability for SNPs with high minor allele frequency (MAF >10%) and that are not in LD (r^2<0.8). The association of SNPs with the residual variation was determined via family-based association test for quantitative trait in gPlink software and also the adjusted P values are given after Bonferroni correction. The observed rs3512 association (P = 0.0457) was significant after multiple test correction (Bonferroni corrected value <0.05).

The aim of these analyses was to evaluate the improvement of the association of these SNPs with variation in somatic instability. rs3512 remains significant (P = 0.0457) after multiple corrections (Bonferroni corrected P value<0.05), showing that the minor allele is associated with residual variation in somatic instability in HD (Table 6-4). This SNP's effect is expected to modify disease phenotype or progression, and the change in somatic instability depends on the SNP genotypes.

6.3.4 Identification of SNPs that modify age at disease onset

It has been shown that high levels of CAG repeat expansions in postmortem brain DNA is associated with an earlier age of disease onset, suggesting that somatic instability is a significant predictor of the age of onset and disease progression (Swami *et al.*, 2009). The main modifying factor in age at onset is the CAG

repeat length that contributes to approximately 70% of the variability, indicating that other modifying factors are likely to be involved in determining the age of onset (Wexler et al, 2004). We assume the genetic modifiers of somatic instability might also modify age at onset.

These candidate genes have shown an association with age at onset (GeM-HD Consortium, 2015; Bettencourt *et al.*, 2016). Our aim was to genotype these candidates SNPs and provide confirmation of suggested loci that modify the age of onset of HD. The variability in age at onset attributable to the size of CAG repeats was adjusted by linear regression. Thus, we performed regression model for an age of onset that accounted for allele length for each individual (as described in Chapter 5) with modifier genes. Regression analysis was performed using age at onset as a dependent variable, and the size of CAG repeats as an independent variable. We did association analysis for the age at onset in the same way as for somatic instability as described in section 6.3.3 except that included the residual variation in age at onset as a quantitative phenotype in order to evaluate the effect of candidate SNPs on the age of onset of Huntington's disease. This analysis is stratified by family structure. We used a subset of 135 affected individuals (43 males and 87 males), for whom we have a clinical age at onset, in the association statistical analyses. Ages of onset range from 23 to 65 years with an approximately normal distribution and a mean age of onset of 43.8 years. The residual variation in age at onset ranged from -3.65 to 2.71.

Surprisingly, the analysis did not show any significant result for the association of these candidate genes in explaining any variation in age at onset (Table 6-3). The sample size was small to detect any association of SNPs on the variation in age at onset of the disease. We didn’t find any association of these candidates’ loci in explaining the variation in age at onset.

6.4 Discussion

Our data from Chapter 5 enabled us to quantify the effect of the allele length and age at sampling, which interact to account for 90% of the variation in somatic instability of HD patients. We also confirmed that residual variation in somatic instability in buccal cells, which is independent of the allele length and

age at sampling, provides evidence that changes in somatic instability increased as a function of repeat length and time of sampling and their interaction. We hypothesised that individual-specific differences in somatic instability are mediated by genetic modifiers. The DNA repair machinery was implicated as a modifier of HD age of motor onset in GeM-HD (The GeM-HD Consortium, 2015) and the follow-up study (Bettencourt *et al.*, 2016).

In this study, we are developing a candidate gene-based approach that will enable the investigation of known candidate gene modifiers of HD age at onset as potential modifiers of somatic instability in HD patients. We sought to test whether these modifier genes in other HD populations could validate previous findings and ultimately lead to the identification of genetic factors contributing to the progression of HD. We tested our hypothesis by quantifying residual variation in somatic instability in HD patients from the Venezuelan cohort, and genotyping modifier gene polymorphism in the DNA repair pathway. We did family-based association tests with a quantitative phenotype, which was the residual variation in somatic instability for each individual, to determine the effect of all the *trans*-acting modifier genes on somatic instability.

The primary analysis for all SNP association showed only rs3512 in *FAN1* yielded a statistically significant association, but it did not survive after multiple corrections. Nonetheless, our secondary analysis examined only the SNPs not in LD and high MAF and showed a significant effect for the minor allele of rs3512 with somatic instability after multiple corrections. No statistically significant associations were detected for any of the remaining variants tested between the candidate genes and somatic instability.

The SNP (rs3512) was found to have age at onset modifying effect in HD (The GeM-HD Consortium, 2015). rs3512 is in the 3' UTR (the three prime untranslated region) of the *FAN1* (FANCD2 and FANCI associated nuclease1) gene on chromosome 15. FAN1 is an endo/exonuclease involved in DNA interstrand cross-link repair (Kratz *et al.*, 2010; Mackay *et al.*, 2010). This SNP (rs3512) was the most significant SNP in a candidate gene study (Bettencourt *et al.*, 2016) and was the second most significant signal in GeM-HD (The GeM-HD Consortium, 2015) and the minor allele was associated with later onset of HD in both studies. The association analysis demonstrated the presence of the minor allele of rs3512

could delay the age at onset by 1.325 years (The GeM-HD Consortium, 2015). Also, there is evidence to suggest the association of this SNP with later age at onset by 1.68 years for both HD and SCA patients, and 1.8 years later onset for all SCA patients alone (Bettencourt *et al.*, 2016). This effect is expected to modify disease phenotype or progression, and the change in age at onset depends on the SNP genotypes.

Individuals carrying minor allele of rs3512 have a higher level of somatic instability (Beta=0.249) than average. An increased level of somatic expansion is associated with larger CAG repeats, which is related to more severe pathology and earlier age at onset (Swami *et al.*, 2009). Therefore, a higher level of somatic expansion could be explained by the genetic variation in DNA mismatch repair genes that could alter age at onset in HD patients. Unexpectedly, the association in our result does not have the same effect as in the previously published data. This variant shows consistent replication effect across the previous studies. Our results showed this SNP display high MAF (0.26), and the minor allele has higher somatic instability (Beta=0.249) than average. An increased level of somatic expansion is associated with larger CAG repeats, which is associated with more severe pathology and earlier age at onset. Therefore, somatic expansion could explain the identification of genetic variation in DNA mismatch repair genes that could alter age at onset in HD patients. The association identified that increase of somatic instability is expected to be associated with earlier age at onset, rather than later as in previous studies (GeM-HD Consortium, 2015; Bettencourt *et al.*, 2016).

Although 412 individuals represent a relatively large cohort for a rare disease, it is nonetheless relatively small compared to previous studies and less power for detecting the contribution of modifier genes. Thus, undetectable association with DNA repair genes in no way excludes them as modifiers of instability in HD. Given the challenge to detect modifier gene associations in such rare disorder, we have shown evidence that rs3512 is associated with individual-specific variation in somatic instability in HD patients. These data suggest that *FAN1* is involved in somatic expansion in HD, with individuals with the homozygous minor allele genotype tending to show higher levels of somatic variation.

It would be worth repeating the analysis with larger samples as the power to detect the genetic modifiers association increases with increasing sample size. We didn't detect any association of these candidate loci in explaining the variation in age at onset. The sample size was relatively small (135 HD patients) for age at onset analysis resulting in less power to detect association with modifier genes.

The absence of any detectable association between candidate genes and age at onset and also the different association of the minor allele frequency of rs3512 in our data compared to previously published data could be due to the population stratification. This may bias analysis of rs3512 as a modifier of disease progression in HD. Also, the causative SNPs in somatic instability might be different in South American rather than European populations. These variants may be present in a specific gene-environmental interaction that is not shared across the different cohorts. Also, another explanation of different results could be due to a different expression level of variants in affected tissues, such as the brain, compared to other tissues (blood and buccal cells). The same allele might be driving higher expression in one tissue and lower expression in another tissue. Thus, the same allele could cause a high level of somatic instability in the affected tissue and lower somatic instability in unaffected tissues, like the brain and buccal cells. Therefore, we could expect to have opposite effect of one locus at the level of somatic instability and modifier genes. Thus, phenotype seems to be measuring a different aspect of disease somatic instability in different tissues. To increase the power of the study, it would be useful to obtain a more accurate measure of phenotype from the affected tissues, which is difficult to obtain from Huntington disease patients. The most affected area in HD is the brain, and notably, the somatic instability of the CAG repeat in the brain predicts the disease onset and progression (Swami *et al.*, 2009). Using more accurate phenotypes would be likely to enhance the power of genetic analyses with such a rare disease.

After designing the experiment, a second GWAS study identified a significant association with other mismatch repair genes (Hensman Moss *et al.*, 2017). They genotyped 218 HD patients from the TRACK-HD study using the Illumina Omni2.5v1.1 array, and they obtained genotypes for 1,773 HD patients from

REGISTRY study. They tested the association of these SNP genotypes with a residual variation of progression measures by meta-analysis. They found a significant association with the *MSH3* gene. Therefore, *MSH3* is probably a modifier of disease progression in HD that has highlighted the DNA repair pathway for therapeutic intervention. *MSH3* has also been implicated in the pathogenic process of HD and degree of somatic instability in striatal cells of a knock-in mouse model of HD (Dragileva *et al.*, 2009). Polymorphisms in the MSH3 gene were also associated with somatic instability in patients with DM1(Morales *et al.*, 2016). Notably, *MSH3* was included in our study, but we did not find an association.

Also, we did not replicate the strongest signal in GWAS, which is derived from rs146353869 in *FAN1* gene that is associated with 6 years earlier age at onset of HD. In our data, we identified 11 heterozygous individuals and no homozygous for the minor allele. No statistically significant association were detected between the rs146353869 and residual variation in somatic instability. Also, to confirm the association with age at onset, we have a subset of 135 patients that we have a clinical age at onset details. The analysis did not show any significant association between this SNP in explaining any variation in age at onset. This is likely because we are genotyping a relatively small number of samples, which is less powered to find an association with an SNP with a low MAF (0.019).

Our finding showed that a small amount of variation in HD somatic instability could be accounted for by genetic variation in DNA repair pathway. The search for modifier genes might have consequences in the understanding of not only Huntington's disease but also diseases such as spinocerebellar ataxia and myotonic dystrophy. Further study to identify these modifiers might provide insight into the pathological process in HD and therefore into therapeutic targets for future investigations.

Chapter 7 Final discussion and conclusion

The overall aim of this project was to evaluate the ability of NGS technology to genotype CAG repeats of the *HTT* gene, including the polymorphic CCG repeats and the flanking sequences, and to detect the distribution and pattern of additional sequence variants within or flanking the repeat. Also, investigations of the use of NGS to measure somatic instability was another aim of this study, as well as to detect *trans*-acting modifiers of HD instability and symptomatic variability.

In this study, we report the analysis of the normal alleles in the general Scottish population and Venezuelan HD families using NGS approaches and the MiSeq platform. In addition, genotyping of the expanded alleles were described from the Venezuelan patients by MiSeq sequencing. We report the first sequence data for normal and expanded *HTT* CAG repeats using optimised library preparation for MiSeq sequencing. From this data, we established that it was possible to sequence and genotype *HTT* CAG alleles, including the polymorphic CCG repeats and the flanking sequences, by amplifying the region using locus-specific primers combined with MiSeq sequencing adapter.

Using this approach we sequenced *HTT* CAG/CCG repeat from blood DNA samples of 210 unaffected individuals from the Scottish population and 742 buccal cell swab DNAs from 333 unaffected and 409 affected individuals from the Venezuelan HD cohort. Our data have revealed that the distribution of the CAG repeat size in the Scottish and Venezuelan population was similar to that previously published in the same populations. CCG alleles of 7 and 10 repeats account for the majority of normal chromosomes in both populations. The carriers of 7 CCG repeats in normal chromosomes were associated with a higher mean CAG length and longer CAG repeat lengths on normal chromosomes in both populations. All expanded alleles were associated with 7 CCG repeats in the Venezuelan population, which reveals that the CCG repeat is stably transmitted. In addition, we revealed the existence of atypical allele structures by sequencing normal *HTT* alleles in both populations. No atypical allele structures were identified by sequencing expanded alleles. Thus, the lack of variation in CCG repeats and the typical allele structures support a possible common source

of mutation in Venezuela due to a founder effect, inheriting 7 CCG repeats on the expanded alleles and also the typical allele structure.

From these data, we were able to sequence CAG repeats at a mode of up to 74 repeats in the presence of the normal alleles for most samples. However, we were unable to reveal the 3'-flanking region using 400 bp forward reads for patients with 74 CAG repeats or longer. For such long alleles, we investigated the utility of reverse reads and found that these can be used to improve the efficiency of genotyping of the CCG repeat and the 3'-flanking DNA sequence. We thus used the reverse reads to genotype the CCG repeat and the 3'-flanking DNA sequence, to detect any variants in that region and for haplotype analysis. Sequencing of alleles longer than 74 CAG repeats may be done using that approach, i.e. using the forward read to genopype the CAG repeat and the reverse reads to genoype the 3'-flank and the CCG repeats.

Analysis of the haplotypes comprising the polymorphic CAG and CCG repeats has revealed that CAG 17-CCG 7 and CAG 17-CCG 10 are the most common haplotypes on their normal chromosomes in the Scottish population. Haplotype analysis in the Venezuelan population showed that the CAG 17-CCG 7 and CAG 18-CCG 9 are the most common haplotypes on their normal chromosomes. Therefore, the Scottish and Venezuelan populations share the same major HD haplotype (CAG 17-CCG 7) that account for ~14 to 19% of their normal chromosomes.

Our analysis of homozygous cases showed that homozygous individuals had an earlier age at onset of disease than heterozygotes. Therefore, homozygosity for HD is expected to result in a more severe phenotype and rapid disease progression. In contrast, an earlier study reported a similar age at onset for homozygous and heterozygous individuals, but homozygotes had a more severe phenotype of HD (Squitieri *et al.*, 2003). These different findings could be because the previously reported study was conducted on a small number of HD patients (eight homozygotes and 75 heterozygotes). In addition, it may be difficult to determine the accurate age at onset in HD due to an overlap of cognitive, motor and psychiatric symptoms in the early stages of HD. Thus, patients with homozygous HD mutations are an important aspect for genetic counselling, given the fact that all their offspring will inherit the HD mutation,

unless a contraction of the CAG repeat tract occurs, which is highly unlikely. Counselling is very important for the patients to know about the increased risk of transmitting the mutation to their offspring.

Our data revealed the high frequency of atypical allele structures in normal *HTT* alleles. Most of the atypical alleles have been reported in published data, suggesting that we accurately determined the genotype for HD alleles and that these atypical alleles are genuine germline variants existing in normal alleles, not PCR and/or sequencing errors. Five atypical allele structures were analysed from 27 normal alleles in the Scottish population. We also observed three atypical normal allele structures from 107 normal alleles from the Venezuelan population, which were also found in the Scottish population. Interestingly, among the Scottish population, only one of these atypical alleles was not a previously described *HTT* allele. This allele is associated with interruption of CCG repeats into the CAG repeat tract, deletion of CAA in the intervening sequence, along with an additional CCT sequence to the CCT duplication present at the end of the CCG repeat. It must be noted that no expanded atypical alleles were detected in the Venezuelan HD patients compared to the analysis of normal *HTT* alleles from unaffected individuals from Scottish and Venezuelan populations that revealed atypical allele haplotypes in the normal alleles.

In parallel to my work, a large screen of HD patients from the Enroll-HD cohort, TRACK-HD, and the European Huntington's Disease Network (EHDN) REGISTRY project samples was conducted using MiSeq sequencing (Dr. Marc Ciosi, unpublished data). Dr. Marc Ciosi sequenced these alleles and described (22/~1,300) mutant atypical alleles in these samples. Most of these variants had already been described in normal alleles in the Scottish and Venezuelan samples. However, he found one atypical expanded allele structure with no interruption between CAG and CCG repeats. Such a lack of interruption between CAG and CCG repeats has previously been reported only in intermediate alleles (Goldberg et al. 1995, Chong et al 1997). We therefore predict that these atypical alleles in HD arise from the general population. Those atypical alleles in HD are present in the general population and then they expand to the disease size range. However, there were no atypical alleles in general population for DM1, while atypical alleles are sometimes seen in DM1 affected families (Dr.

Sarah Cumming, Submitted Jun, 2018). Therefore, the atypical expanded alleles in DM1 most likely arise from a *de novo* mutation in patients. Variant repeats in DM1, SCA1, and fragile X syndrome have been shown to modify the mutational dynamics by reducing the amount of germline and somatic instability (Chung *et al.*, 1993; Zhong *et al.*, 1995; Musova *et al.*, 2009; Braida *et al.*, 2010). In addition, Goldberg et al. 1995 and Chong et al.1997 have shown that intermediate HTT CAG alleles with no interpretation between CAG and CCG repeats have a higher germline instability than other intermediate alleles.

These recent data from our group suggests that in addition to the polyglutamine repeat length, sequence variants within the *HTT* repeat tract may affect disease severity (Dr. Marc Ciosi, unpublished data). Atypical alleles may potentially expand less or more in both the soma and germline, possibly modifying the disease severity. Some atypical alleles may have an impact leading to increased severity and level of instability. For instance, individuals having atypical expanded alleles with loss of CAA from the CAG repeats tract have more severe symptoms than expected based on the number of glutamines (Dr. Marc Ciosi, unpublished data). Similarly, four HD patients were described with an atypical allele, which is characterised by loss of intervening sequences in the expanded alleles, leading to pure tracts of CAG and CCG repeats. These patients are also more severely affected than expected (Dr. Marc Ciosi, unpublished data). Therefore, pathogenesis and somatic instability are best predicted by the number of CAG repeat, not polyglutamine length. Individuals carrying atypical alleles with fewer variants CAA are more severely affected and those alleles are likely to be unusually unstable through a transmission to offspring relative to the number of total glutamine (Goldberg et al. 1995, Chong et al 1997).

Individuals having atypical alleles with duplication of CAACAG repeats within the CAG repeats tracts have less severe symptoms than expected relative to the number of polyglutamines (Dr. Marc Ciosi, unpublished data). Atypical alleles with more variant repeats would be expected to be less prone to expansion when transmitted to the next generation, according to data available for other triplet repeat disorders (Chung *et al.*, 1993; Zhong *et al.*, 1995; Musova *et al.*, 2009; Braida *et al.*, 2010). Notably, the presence of interruptions within atypical normal alleles which break the CAG repeat into three smaller repeat tracts in

our data (the one atypical normal allele that was first seen in the Scottish population, our data from Chapter 3) may stabilise the allele and decrease the severity of symptoms if found on the mutant chromosome.

Thus, we expect atypical allele with more variant repeats to be more stable, and those with less variant repeats to be more unstable, predisposing alleles to expansion and eventually to disease status. Therefore, for the trinucleotide repeats disorders, the interruptions provide genetic stability to the repeat tract and the interrupted tracts are less likely to expand upon transmission (Chung *et al.*, 1993; Zhong *et al.*, 1995; Musova *et al.*, 2009; Braida *et al.*, 2010).The identification of those atypical alleles can be difficult in the HD diagnostic test using PCR analysis, as it does not take into account the allele structure. From M. Ciois's data, atypical alleles with loss of CAA are associated with a motor onset arising ~10 years earlier and individuals with a duplication of CAACAG within the CAG repeats track have a age at motor onset delayed by ~5 years based on total number of glutamine (unpublished data). Thus, sequencing those alleles is necessary as they have an implication for the individual's risk, as well as allele stability through life and transmission to offspring.

The NGS approach represents a rapid and a higher throughput method, compared to other traditional diagnostic and research methods. These traditional approaches used in genotyping HD alleles have some limitations in terms of sensitivity, accuracy and measuring somatic instability because they are not precise in determining CAG allele sizes and fail to define the extent of somatic instability. In addition, these approaches cannot detect any variants repeats within or flanking the repeats. Thus, replacing these conventional approaches with NGS would allow for a rapid analysis and accurate quantification of allele length variation. In addition, it is important to note the sensitivity of this sequencing method and the ability to detect insertions, deletions or other sequence variants within or flanking the repeat. Therefore, NGS is likely to have significant clinical utility. We illustrate the success and challenges of using NGS in trinucleotide disorders and suggest that our experience is likely to be applicable to many other expanded repeat alleles.

The present protocol has the potential to improve HD diagnostics as it can reveal *HTT* variants that have been ignored by standard diagnostic protocols. This

developed approach can improve the accuracy of diagnostics based on the CAG repeat genotyping. It also allows the quantification of CAG repeat somatic instability. Our protocol has the advantage of using the dual index approach in which 96 samples can be sequenced using MiSeq sequencing. This approach can be used to multiplex larger number of samples up to 384 samples using 96 barcode kits such as the Nextera Index kit, which also reduces the cost per sample. A large-scale genotyping study would be useful for a population screening approach for HD. Population screening is a necessary step in identifying individuals who may be pre-symptomatic or asymptomatic for a disease, particularly one with a late age at disease onset. Genetic screening could be useful in decisions about lifestyle and reproductive choices. Screening for HD in the general population will allow for early and accurate diagnosis. Also, that can help in providing early intervention and management of HD symptoms. Furthermore, with increasing clinical trials and accelerating therapeutic development, affected individuals could start new therapeutics early as they become available. Any possible therapy is more likely to slow or stabilise the disease rather than to reverse the symptoms, therefore it would be useful to identify at risk individuals in the population before they develop symptoms.

We have analysed the pattern of somatic instability in buccal cell DNA in the normal and expanded *HTT* alleles from unaffected and affected individuals of the Venezuelan kindred. There was no significant level of instability observed in the normal alleles, but with mutant alleles, instability was observed in all expanded alleles and an even greater level was observed with older age. These data further suggest the allele length and age are major modifiers of somatic mosaicism and that the normal alleles are highly stable.

It has been shown that, in HD, there is a noticeable correlation between the tissues vulnerable to neuropathology and those exhibiting a high level of instability (Kennedy *et al.*, 2003). This observation has led to the hypothesis that somatic instability contributes to the disease pathogenesis and progression in HD. Somatic instability is tissue-specific, age-dependent, and expansion-biased (Kennedy *et al.*, 2003). In order to test whether the somatic instability of HD CAG repeats could be quantified and to understand the role of somatic

instability in HD, somatic instability was investigated in HD patients from the Venezuelan HD families. We thus used MiSeq read length distributions to quantify the degree of somatic mosaicism of the expanded CAG repeats for each individual. Our hypothesis was that the main drivers for somatic mosaicism are CAG repeat length and age at sampling and that individual-specific residual variation for both effects will also be associated with disease severity.

Our results demonstrated that CAG repeat length is a major determinant of somatic instability that accounts for 64% of the variation in somatic instability. Age at sampling was also confirmed as a major factor in determining the level of somatic instability by using multivariate analyses, confirming the interaction between age at sampling and allele length. 90% of the variation in somatic instability was explained by CAG repeat length, age at sampling and their interaction. The remaining variation could be explained by environmental factors and other genetic modifiers. We also observed a significant correlation between the expanded CAG repeat length and age at disease onset in HD patients. The expanded CAG repeat explains 56% of the variation in age at onset. This suggests expanded allele length is the initial trigger of HD pathogenesis and a major factor determining the rate of the process.

Further analysis was carried out to test whether age at onset and disease severity are influenced by somatic instability, and to determine whether the pathogenic effect of expanded alleles can be modulated by CAG repeat instability. We evaluated the residual variation of somatic instability not explained by allele length and age at sampling that likely reflects the specific mutation rate variation between individuals. We expect those individuals' specific differences will be associated with disease severity and that somatic instability will have an influence on age at onset. Our analysis showed no association between that the residual variation in somatic instability and the variation in age at disease onset that is not explained by the allele length. In contrast, there was evidence from a previous study showing a negative correlation (r^2=0.068, P=0.002) between repeat instability and age at onset of DM1 patients (Morales *et al.*, 2012). The evidence from this study suggests that the residual variation of somatic instability is an individual-specific and heritable quantitative trait and that residual variation in somatic instability is associated

with residual variation in age at onset in DM1 patients (Morales *et al.*, 2012). Moreover, data from our group on ~400 HD patients from the Enroll-HD cohort showed that there is a correlation between repeat instability and age at onset of disease (r^2=0.02, P=0.002) (Dr. Marc Ciosi). These data revealed the somatic instability accounted for some of the variation in age at onset and the repeat expands more rapidly for individuals who have an earlier age at onset in both DM1 and HD patients. However, it is important to note that the correlation between residual variations of somatic instability and age at onset is relatively low for both HD patients (r^2=0.02) and DM1 patients (r^2=0.068). That implicates the role of individual-specific environmental or genetic factors as a modifier of the individual-specific differences in the level of somatic instability. Furthermore, it has been reported that CAG repeat expansion in postmortem DNA is associated with an earlier age of disease onset, suggesting that somatic instability is a significant predictor of the age of onset and disease progression (Swami *et al.*, 2009). Although there are data implicating somatic mosaicism as an important component of the disease pathway and that somatic instability play a role in age at onset, we could not verify it in our study. This could result from the relatively low number of samples in our analysis (135 individuals) that have age at onset details and the difficulties of precisely assigning age at onset to a patient. In contrast to the Enroll-HD cohort, the phenotype in our study is less well characterised and also we tested a relatively small number of patients for the analysis. Furthermore, it is important to note the measurement of somatic mosaicism from the non-target tissue is likely underestimating the real repeat dynamics in the brain.

Because of PCR slippage, our method has a limitation in detecting and measuring somatic contractions. Data from single molecule experiments on HD patients showed that there was a low frequency of contractions (Veitch *et al.*, 2007). Understanding somatic contractions is important, enabling us to provide insight into the contraction mechanisms to study triplet disease disorders and also to understand the phenotype-genotype correlation. It might be that expansions and contractions involve different mechanisms and also different genetic modifiers. A high frequency of contractions might result in individuals with less severe symptoms. If so, enhancing repeat contraction could be clinically beneficial. Understanding the mechanisms of instability in humans is crucial for the

development of therapeutic approaches that target repeat instability and to identify genetic factors influencing it. A therapeutic strategy for the expansion disorders could work by inhibiting expansions, or by inducing contractions of the mutation by reversing the repeat expansion in the mutant genes to the shorter length present in the non-disease chromosomes. This targeting therapy may have consequences on the downstream pathology and symptoms in HD.

Given the limitation of our study in addressing the contraction process in HD, our group has shown expanded HD allele from a single molecule could be sequenced using NGS approach. A single molecule sequencing experiment was conducted from a mix DNAs from 25 HD patients carrying between 40 to 67 CAG repeats that had been previously sequenced and sized by MiSeq sequencing (Dr. Sarah Cumming, unpublished data). MiSeq sequencing of PCR products derived from single expanded alleles was carried out to examine the read length distribution. The numbers of reads longer than the mode were very small (2-3%), while reads shorter than the mode were still very numerous. These data showed that in a bulk PCR sequencing experiment the reads shorter than the inherited allele length are derived from PCR slippage, and reads larger than inherited allele length are mainly due to somatic mosaicism. Although this experiment provides evidence for our ability to sequence a single HD molecule, it was time-consuming and labour-intensive as it involved an SP-PCR approach and detection of products by Southern blot hybridisation.

Avoiding these the labour-intensive preparation steps required for that experiment, it could be replaced by a rapid and high-throughput method. Interesting data from our group has shown the sequencing of DM1 single molecules is possible by tagging each parental molecule with unique sequences via PCR followed by MiSeq sequencing (Dr. Khaldah Nasser, unpublished data). The aim of this experiment was to distinguish somatic mosaicism from PCR slippage and sequencing errors. The idea was to tag each parental molecule with unique sequences of 14 random nucleotides in order to distinguish each template (Kinde *et al.*, 2011; Kou *et al.*, 2016). Then, these tagged molecules were subsequently amplified to generate daughter molecules that were derived from parental template molecules. This method allows for the somatic variants

present in the parental templates to be copied in derivative molecules and therefore discriminated from PCR or sequencing errors.

Therefore, by applying the same method by tagging each parental HD molecule with unique barcoded primers by PCR followed by sequencing, we could more accurately evaluate the level of contractions and expansions by comparing the read length distribution obtained by single molecule to the sequencing of a bulk DNA. This method could be used to provide an understanding of the underlying dynamics of the contraction and expansions in HD single molecules, although the method could be challenging to amplify a single HD mutant allele as it had showed sequencing DM1 normal alleles.

There is strong evidence that somatic expansions in HD occur in non-dividing cells such as neurons (Shelbourne *et al.*, 2007). The presence of expansions in non-dividing somatic cells indicates that somatic instability can be induced during metabolic pathways other than DNA replication and cell division. The data from the HD knock out mice of the mismatch repair genes eliminate the somatic expansions and delay the phenotype (Wheeler *et al.*, 2003; Dragileva *et al.*, 2009; Pinto *et al.*, 2013). These data support the involvement of the DNA repair pathway in somatic instability. However, during gametogenesis, cell divisions occur, and these are particularly numerous during sperm production. The CAG repeat length may increase during germline transmission of premutation and disease-causing alleles, accounting for the phenomenon of anticipation, which is mostly seen during paternal transmission. This could suggest that cell division may result in repeat length changes in gametes. During early development, numerous cell divisions occur, but by birth many tissues will be post-mitotic. However, due to the difficulty in obtaining tissues, the level of CAG repeat instability during development has not been widely studied, therefore it is difficult to determine the mechanisms that may underlie repeat instability during development.

Studies in HD mouse models confirmed the role of *trans* acting modifiers in HD CAG repeat instability and HD CAG-dependent phenotypes (Wheeler *et al.*, 2003; Pinto *et al.*, 2013). These modifier genes have been shown to be critical in generating somatic repeat expansion in mice. The most obvious candidates for *trans* acting modifiers of somatic instability are the DNA mismatch repair genes.

In humans, several analyses have identified polymorphisms in DNA repair genes that may help to explain variable disease onset and progression, such as *MSH3* and *PMS2* (GeM-HD Consortium, 2015; Bettencourt *et al.*, 2016). Therefore, we investigated whether known DNA mismatch repair (MMR) gene modifiers of HD age at onset were also modifiers of somatic instability in HD patients. These modifier genes in HD could ultimately lead to an identification of genetic factors contributing to the progression of HD.

We did family-based association tests with a quantitative phenotype, which used the residual variation of somatic instability for each individual to determine the effect of all these *trans*-acting modifier genes on the somatic instability of the analysed HD patients. This residual variation represents a quantitative measure to test association with the *trans*-acting modifier gene genotypes. Our data demonstrated a significant effect for the minor allele of rs3512 in *FAN1* with somatic instability after multiple corrections for SNPs that were not in LD and which have a high minor allele frequency. The association analysis demonstrated that individuals carrying the minor allele of rs3512 have a higher level of somatic instability than average (Beta =0.23). A similar study in our group was carried out by Dr. Marc Ciosi in which he identified an effect for the minor allele of rs3512 in *FAN1* with somatic instability (Beta =0.29) in TRACK-HD patients (Dr.Marc Ciosi, unpublished data). In addition, Marc Ciosi had observed a similar effect in Enroll-HD patients; there was a significant association of the minor allele of rs3512 with somatic instability (Beta =0.018). Our group's results were consistent in demonstrating the individuals carrying the minor allele at rs3512 have a higher level of somatic instability than average. We were expecting a high level of somatic instability to be associated with an earlier age at onset. Surprisingly, the association obtained in our results does not have the same effect regarding the expected effect on age at disease onset as in the previously published data, in which the minor allele was associated with later age at onset (GeM-HD Consortium, 2015; Bettencourt *et al.*, 2016). It is worth mentioning that our result is the first data to show that the minor allele in *FAN1* is associated with somatic mosaicism in human HD patients. Furthermore, there is no mice data on the effect of *FAN1*. Our data directly implicates *FAN1* in having effects on somatic instability and therefore playing a role in expansion pathway.

Nevertheless, GWAS and also Marc Ciosi's data have shown the DNA mismatch repair genes modify the age at onset, and that these MMR genes affect the level of somatic expansion. Also, the pathway analyses from GWAS highlighted DNA mismatch repair as a likely mechanism to modify disease phenotype. While the role of *FAN1* in mediating somatic expansion in HD was not observed, there was suggestive evidence that this gene interacts with *MLH1* in the GeM-HD GWAS. FAN1 is involved in interstrand DNA cross-link repair (Jin and Cho, 2017). FAN1 might recruited to the abnormal DNA structures and the mismatch repair gene, such as MLH1 is recruited and interacts with FAN1 to recognises and resolve the abnormal structures (Smogorzewska *et al.*, 2010). This repair of the strand may lead to an expansion of CAG repeats. FAN1 might target these structures and interacts with MLH1 that may play a role in modulating repeat instability. This SNP in FAN1 is in LD with other SNPs in different genes. Therefore, the causative SNP is not known. The SNP rs3512 lies in the 3'UTR of FAN1. The 3'UTR sequence contains regulatory elements that determine mRNA stability and localization (Arnold *et al.*, 2012). If the SNP is causative, since it lies in the 3' UTR of the gene, it could interfere with mRNA stability and translation through effects on polyadenylation, or interactions with regulatory proteins or interactions of the transcript with miRNA. Variations in the 3' UTR of genes could also potentially inhibit nuclear-cytoplasmic transport of an RNA, which would lead to reduced translation. The mutation found in 3'UTR of FAN1 might therefore alter gene expression by various possible mechanisms.

The absence of any detectable association between candidate genes and age at onset and also the different association of the minor allele frequency of rs3512 in our data compared to previously published data could be due to the population stratification. The causative SNPs in somatic instability might be different in South American rather than European populations. Also, these variants may be present in specific gene-environmental interactions that are not shared across the different cohorts. However, this does not seem to be a credible explanation, given that Marc Ciosi's data gave the opposite effect of the variant effect than GWAS analysis but both included patients from Caucasian populations.

Another explanation of different results could be due to different expression levels of variants in affected tissues, such as the brain, compared to other tissues e.g. blood and buccal cells. The same allele might be driving higher expression in one tissue and lower expression in another tissue. Thus, the same allele could cause a high level of somatic instability in the affected tissue and lower somatic instability in unaffected tissues, like blood and buccal cells. To study the tissue-specific gene expression of rs3512 on *FAN1* across many tissues, we associate this variant with gene expression levels from the Genotype-Tissue Expression (GTEx) programme - v7 release (https://www.gtexportal.org/home/). We found the highest expression level of the variant in brain cortex as compared to other tissues types such as blood. Therefore, the minor allele of rs3512 is associated with the same effect of expression level in cortex and blood, but it is higher in the brain. We expected the minor allele to be overexpressed in affected tissues and underexpressed on the unaffected tissues, to clarify our finding that different aspect of disease somatic instability is measured in different tissues. The GTEx data did not support the lack of correlation between our data and the GWAS data that expected to be due to expression changes in *FAN1* across unaffected and affected tissues. Notably, the GTEx study included complex tissues but does not account for the cellular heterogeneity of the tissue, which could mask the biological importance of cellular expression and lead to misinterpretation of gene expression association.

Understanding the role of rs3512 in *FAN1* could be addressed by using brain samples from patients to evaluate the somatic instability and SNP genotypes in a human brain, as it is involved in the pathological process of HD. The somatic instability of the CAG repeat in the brain predicts the disease onset and progression (Swami *et al.*, 2009). However, brain samples are difficult to obtain from living patients, which makes it hard to study changes in the brain and impossible to perform longitudinal studies. Also, it would be useful to obtain a more accurate measure of phenotype to enhance the power of understanding the role of somatic instability. For instance, the progression score from TRACK-HD is considered as a good measure for HD patient phenotype, which reflects the progression in the motor, cognitive and imaging measures (Hensman Moss *et al.*,

2017). Therefore, obtaining a measure of progression score from HD patients could help in assessing somatic instability.

The power of our study in detecting the association between somatic instability and age at onset may have been reduced due to the relatively small samples size. It would be worth repeating the analysis with a larger number of samples to confirm our results as the power to detect the genetic modifiers association increases with increasing sample size. In addition, no statistically significant associations were detected for any of the remaining variants tested between the candidate genes and somatic instability. The sample size was relatively small (135 HD patients) for age at onset analysis, resulting in less power to detect associations with modifier genes.

After designing the experiment, a second GWAS identified a significant association with other mismatch repair genes (Hensman Moss *et al.*, 2017). They genotyped 218 HD patients from the TRACK-HD study using the Illumina Omni2.5v1.1 array and obtained genotypes for 1,773 HD patients from the REGISTRY study. They tested the association of these SNP genotypes with the residual variation of progression measures by meta-analysis. They found a significant association with the *MSH3* gene. Therefore, *MSH3* is probably a modifier of disease progression in HD, highlighting the DNA repair pathway for therapeutic intervention. Their data from TRACK-HD was able to give a significant association in only 216 subjects. Therefore, a well-characterised phenotype is powerful to detect association even with a small number of patients. *MSH3* has also been implicated in the pathogenic process of HD and degree of somatic instability in striatal cells of a knock-in mouse model of HD (Dragileva *et al.*, 2009). Polymorphisms in the *MSH3* gene were also associated with somatic instability in patients with DM1 (Morales *et al.*, 2016). Notably, *MSH3* was included in our study, but we did not find an association. This is likely due to our sample being much smaller than that GWAS used in the study and thus less powered to identify the association with this gene. Identification of modifier genes for HD somatic instability would provide new targets for therapeutic interventions in HD that alter the HD disease in human patients, with the potential to delay disease onset and progression.

In conclusion, our findings highlighted the importance of the utility of using NGS approaches to genotype, characterise the distribution pattern of HD alleles and also to quantify somatic instability in the number of CAG repeats. Our study offers further relevance to the hypothesis that the repeat instability is age and allele length dependent and expansion-biased. The longest CAG repeats show wide somatic instability and may offer a mechanistic model to study triplet drug-controlled instability and genetic factors influencing it. Elucidating the mechanisms of repeat instability is crucial for understanding disease pathogenesis, and also for the development of therapeutic approaches aimed at preventing CAG repeat expansion. If suppressing somatic expansion substantially delays the onset of disease, treatment may be possible during a lifetime. Identification of interesting variants in the *trans*-acting modifier gene *FAN1* that modifies repeats instability of HD and could be used as a therapeutic target. Our data provided insight into the role of variants in mismatch repair genes in repeat instability in HD. These data can be used for large-scale analysis using well phenotyped large cohorts for HD such as Enroll-HD, to address the possible dependence of somatic variation of CAG repeats on disease pathology and to identify factors acting in *cis* or in *trans* to the mutation. In addition, identification of the genetic and environmental factors that may influence the age at onset could have an impact on patient's survival or disease risk. A large-scale analysis could provide a better understanding of phenotype-genotype correlation in HD, stratify HD patients to allow the appropriate design of clinical trials, and may reveal novel therapeutic interventions.

Bibliography

Agostinho, L. A., dos Santos, S. R., Alvarenga, R. M. P. and Paiva, C. L. A. (2013) 'A systematic review of the intergenerational aspects and the diverse genetic profiles of Huntington's disease', *Genetics and Molecular Research*, 12(2), pp. 1974–1981. doi: 10.4238/2013.June.13.6.

Agostinho, L. de A., Rocha, C. F., Medina-Acosta, E., Barboza, H. N., Da Silva, A. F. A., Pereira, S. P. F., Da Silva, I. D. S., Paradela, E. R., Figueiredo, A. L. D. S., Nogueira, E. D. M., Alvarenga, R. M. P., Hernan Cabello, P., Dos Santos, S. R. and Paiva, C. L. A. (2012) 'Haplotype analysis of the CAG and CCG repeats in 21 Brazilian families with Huntington's disease', *Journal of Human Genetics*, 57(12), pp. 796–803. doi: 10.1038/jhg.2012.120.

Alonso, M. E., Yescas, P., Rasmussen, A., Ochoa, A., Macías, R., Ruiz, I. and Suástegui, R. (2002) 'Homozygosity in Huntington's disease: new ethical dilemma caused by molecular diagnosis', *Clinical Genetics*, 61(6), pp. 437–442.

Andresen, J. M., Gayán, J., Djoussé, L., Roberts, S., Brocklebank, D., Cherny, S. S., The US-Venezuela Collaborative Research Group, Group, T. H. M. C. R., Cardon, L. R., Gusella, J. F., MacDonald, M. E., Myers, R. H., Housman, D. E. and Wexler, N. S. (2007) 'The relationship between CAG repeat length and age of onset differs for Huntington's disease patients with juvenile onset or adult onset', *Annals of Human Genetics*, 71, pp. 295–301. doi: 10.1111/j.1469-1809.2006.00335.x.

Andrew, S. E., Goldberg, Y. P., Theilmann, J., Zeisler, J. and Hayden, M. R. (1994) 'A CCG repeat polymorphism adjacent to the CAG repeat in the huntington disease gene: Implications for diagnostic accuracy and predictive testing', *Human Molecular Genetics*, 3(1), pp. 65–67.

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

Arnold, M., Ellwanger, D. C., Hartsperger, M. L., Pfeufer, A. and Stu, V. (2012) 'Cis-Acting Polymorphisms Affect Complex Traits through Modifications of MicroRNA Regulation Pathways', 7(5), pp. 1–12. doi: 10.1371/journal.pone.0036694.

Barron, L. H., Rae, A., Holloway, S., Brock, D. J. H. and Warner, J. P. (1994) 'A single allele from the polymorphic CCG rich sequence immediately 3' to the unstable CAG trinucleotide in the IT15 cDNA shows almost complete disequilibrium with huntington's disease chromosomes in the Scottish population', *Human Molecular Genetics*, 3(1), pp. 173–175. doi: 10.1093/hmg/3.1.173.

Bates, G. P. (2005) 'The molecular genetics of Huntington disease — a history', *Nature*, 6, pp. 766–773. doi: 10.1038/nrg1686.

Bellacosa, A. (2001) 'Functional interactions and signaling properties of mammalian DNA mismatch repair proteins', *Cell Death and Differentiation*, 8(11), pp. 1076–1092. doi: 10.1038/sj.cdd.4400948.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R. and et al. (2008) 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456(7218), pp. 53–59. doi: 10.1038/nature07517.

Bettencourt, C., Hensman-Moss, D., Flower, M., Wiethoff, S., Brice, A., Goizet, C., Stevanin, G., Koutsis, G., Karadima, G., Panas, M. and et al. (2016) 'DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine Diseases', *Annals of Neurology*, 79(6), pp. 983–990. doi: 10.1002/ana.24656.

Braida, C., Stefanatos, R. K. A., Adam, B., Mahajan, N., Smeets, H. J. M., Niel, F., Goizet, C., Arveiler, B., Koenig, M., Lagier-Tourenne, C., Mandel, J. L., Faber, C. G., de Die-Smulders, C. E. M., Spaans, F. and Monckton, D. G. (2010) 'Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients', *Human Molecular Genetics*, 19(8), pp.

1399–1412. doi: 10.1093/hmg/ddq015.

Brocklebank, D., Gayán, J., Andresen, J. M., Roberts, S. A., Young, A. B., Snodgrass, S. R., Penney, J. B., Ramos-Arroyo, M. A., Cha, J. J., Rosas, H. D., Hersch, S. M., Feigin, A., Cherny, S. S., Wexler, N. S., Housman, D. E. and Cardon, L. R. (2009) 'Repeat Instability in the 27–39 CAG Range of the HD Gene in the Venezuelan Kindreds: Counseling Implications', *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 5(150B), pp. 425–429. doi: 10.1002/ajmg.b.30826.

van den Broek, W. ., Nelen, M. R., Wansink, D. G., Coerwinkel, M. M., Riele, H. te, Groenen, P. J. and Wieringa, B. (2002) 'Somatic expansion behaviour of the (CTG) n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch – repair proteins', *Human Molecular Genetics*, 11(2), pp. 191–198.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N. and Knight, R. (2011) 'Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample', *Proceedings of the National Academy of Sciences*, 108(Supplement 1), pp. 4516–4522. doi: 10.1073/pnas.1000080107.

Castilhos, R., Augustin, M., Santos, J., Perandones, C., Saraiva-Pereira, M. and Jardim, L. (2016) 'Genetic aspects of Huntington's disease in Latin America. A systematic review', *Clinical Genetics*, 89, pp. 295–303. doi: 10.1111/cge.12641.

Chi, L. M. and Lam, S. L. (2005) 'Structural roles of CTG repeats in slippage expansion during DNA replication', *Nucleic Acids Research*, 33(5), pp. 1604–1617. doi: 10.1093/nar/gki307.

Chung, M., Ranum, L. P. W., Duvick, L. A., Servadio, A., Zoghbi, H. Y. and Orr, H. T. (1993) 'Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I', *Nature genetics*, 5, pp. 254–258. doi: 10.1038/ng0293-165.

Cleary, J. D. and Pearson, C. E. (2003) 'The contribution of cis-elements to

disease-associated repeat instability: Clinical and experimental evidence', *Cytogenetic and Genome Research*. doi: 10.1159/000072837.

Costa, M. D. C., Magalhães, P., Guimarães, L., Maciel, P., Sequeiros, J. and Sousa, A. (2006) 'The CAG repeat at the Huntington disease gene in the Portuguese population: Insights into its dynamics and to the origin of the mutation', *Journal of Human Genetics*, 51(3), pp. 189–195. doi: 10.1007/s10038-005-0343-8.

Djoussé, L., Knowlton, B., Hayden, M., Almqvist, E. W., Brinkman, R., Ross, C., Margolis, R., Rosenblatt, A., Durr, A., Dode, C., Morrison, P. J., Novelletto, A., Frontali, M., Trent, R. J. A., Mccusker, E., Macdonald, M. E., Myers, R. H. and Myers, R. H. (2003) 'Interaction of Normal and Expanded CAG Repeat Sizes Influences Age at Onset of Huntington Disease', *American Journal of Medical Genetics*, 119, pp. 279–282. doi: 10.1002/ajmg.a.20190.

Dragileva, E., Hendricks, A., Teed, A., Gillis, T., Lopez, E. T., Friedberg, E. C., Kucherlapati, R., Edelmann, W., Lunetta, K. L., MacDonald, M. E. and Wheeler, V. C. (2009) 'Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes', *Neurobiology of Disease*, 33(1), pp. 37–47. doi: 10.1016/j.nbd.2008.09.014.

Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M. and et al. (1993) 'Trinucleotide repeat length instability and age of onset in Huntington's disease', *Nature Genetics*. Nature Publishing Group, 4(4), pp. 387–392. doi: 10.1038/ng0893-387.

Ewing, B. and Green, P. (1998) 'Base-calling of automated sequencer traces using phred. II. Error probabilities.', *Genome Research*, 8, pp. 186–194. doi: 10.1101/gr.8.3.175.

Ewing, B., Hillier, L., Wendl, M. and Green, P. (1998) 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment.', *Genome Research*, 8, pp. 175–185. doi: 10.1101/gr.8.3.175.

Falush, D., Almqvist, E. W., Brinkmann, R. R., Iwasa, Y. and Hayden, M. R.

(2000) ‘Measurement of Mutational Flow Implies Both a High New-Mutation Rate for Huntington Disease and Substantial Underascertainment of Late-Onset Cases’, *American Journal of Human Genetics*, 68, pp. 373–385.

García-Planells, J., Burguera, J. A., Solís, P., Millán, J., Ginestar, D., Palau, F. and Espinós C. (2005) ‘Ancient Origin of the CAG Expansion Causing Huntington Disease in a Spanish Population’, *Human mutation*, 25(5), pp. 453–459. doi: 10.1002/humu.20167.

Gayán, J., Brocklebank, D., Andresen, J. M., Alkorta-Aranburu, G., Cader, M. Z., Roberts, S. A., Cherny, S. S., Wexler, N. S., Cardon, L. R. and Housman, D. E. (2008) ‘Genomewide linkage scan reveals novel loci modifying age of onset of Huntington’s disease in the Venezuelan HD kindreds’, *Genetic Epidemiology*, 32, pp. 445–453. doi: 10.1002/gepi.20317.

GeM-HD Consortium (2015) ‘Identification of Genetic Factors that Modify Clinical Onset of Huntington’s Disease’, *Cell*, 162(3), pp. 516–526. doi: 10.1016/j.cell.2015.07.003.

Goldberg, Y. P., Andrew, S. E., Clarke, L. A. and Hayden, M. R. (1993) ‘A PCR method for accurate assessment of trinucleotide repeat expansion in Huntington disease’, *Human Molecular Genetics*, 2(6), pp. 635–636.

Goldberg, Y. P., Mcmurray, C. T., Zeisler, J., Almqvist, E., Slllence, 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(10), pp. 1911–1918.

Gomes-Pereira, M., Bidichandani, S. I. and Monckton, D. G. (2004) ‘Analysis of unstable triplet repeats using small-pool polymerase chain reaction’, *Trinucleotide Repeat Protocols*, 277, pp. 61–76. doi: 10.1385/1-59259-804-8:061.

Gomes-Pereira, M., Fortune, M. T., Ingram, L., McAbney, J. P. and Monckton, D. G. (2004) ‘Pms2 is a genetic enhancer of trinucleotide CAG . CTG repeat somatic mosaicism : implications for the mechanism of triplet repeat expansion’, *Human*

Molecular Genetics, 13(16), pp. 1815–1825. doi: 10.1093/hmg/ddh186.

Gomes-Pereira, M. and Monckton, D. G. (2006) 'Chemical modifiers of unstable expanded simple sequence repeats: What goes up, could come down', *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 598(1), pp. 15–34. doi: 10.1016/j.mrfmmm.2006.01.011.

Gusella, J. F. and MacDonald, M. E. (2000) 'Molecular genetics: Unmasking polyglutamine triggers in neurodegenerative disease', *Nature Reviews Neuroscience*, 1(2), pp. 109–115. doi: 10.1038/35039051.

Gusella, J. F. and MacDonald, M. E. (2006) 'Huntington's disease: seeing the pathogenic process through a genetic lens', *Trends in Biochemical Sciences*, 31(9), pp. 533–540. doi: 10.1016/j.tibs.2006.06.009.

Gusella, J. F. and MacDonald, M. E. (2009) 'Huntington's disease: the case for genetic modifiers.', *Genome medicine*. BioMed Central, 1(8), p. 80. doi: 10.1186/gm80.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. and Martin, J. B. (1983) 'A polymorphic DNA marker genetically linked to Huntington's disease', *Nature*, 306(5940), pp. 234–238. doi: 10.1038/306234a0.

Harper, P. S. (1992) 'The epidemiology of Huntington ' s disease', *Human Genetics*, 89, pp. 365–376.

Hensman Moss, D. J., Pardiñas, A. F., Langbehn, D., Kitty, L., Leavitt, B. R., Roos, R., Durr, A., Mead, S., Investigators, T.-H., Investigators, R., Holmans, P., Jones, L. and Tabrizi, S. J. (2017) 'Identification of genetic variants associated with Huntington ' s disease progression : a genome-wide association study', 16(9), pp. 701–711. doi: 10.1016/S1474-4422(17)30161-8.

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), pp. 971–983. doi: 10.1016/0092-8674(93)90585-E.

Jin, H. and Cho, Y. (2017) 'Structural and functional relationships of FAN1', *DNA Repair*, 56, pp. 135–143. doi: 10.1016/j.dnarep.2017.06.016.

Kay, C., Collins, J. A., Miedzybrodzka, Z., Madore, S. J., Gordon, E. S., Gerry, N., Davidson, M., Slama, R. A. and Hayden, M. R. (2016) 'Huntington disease reduced penetrance alleles occur at high frequency in the general population', *Neurology*, 87(3), pp. 282–288. doi: 10.1212/WNL.0000000000002858.

Kennedy, L., Evans, E., Chen, C.-M., Craven, L., Detloff, P. J., Ennis, M. and Shelbourne, P. F. (2003) 'Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis', *Human Molecular Genetics*, 12(24), pp. 3359–3367. doi: 10.1093/hmg/ddg352.

Kennedy, L. and Shelbourne, P. F. (2000) 'Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease?', *Human molecular genetics*, 9(17), pp. 2539–44.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. and Vogelstein, B. (2011) 'Detection and quantification of rare mutations with massively parallel sequencing.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(23), pp. 9530–5. doi: 10.1073/pnas.1105422108.

Kou, R., Lam, H., Duan, H., Ye, L., Jongkam, N., Chen, W., Zhang, S. and Li, S. (2016) 'Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations', *PLoS ONE*, 11(1), pp. 1–15. doi: 10.1371/journal.pone.0146638.

Kratz, K., Schö, B., Kaden, S., Sendoel, A., Eberhard, R., Lademann, C., Cannavó, E., Sartori, A. A., Hengartner, M. O. and Jiricny, J. (2010) 'Deficiency of FANCD2-Associated Nuclease KIAA1018/FAN1 Sensitizes Cells to Interstrand Crosslinking Agents', *Cell*, 142, pp. 77–88. doi: 10.1016/j.cell.2010.06.022.

Kremer, B., Almqvist, E., Theilmann, J., Spence, N., Telenius, H., Goldberg, Y.

P. and 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), pp. 343–50. doi: .

Kremer, B., Goldberg, P., Andrew, S. E., Theilmann, J., Telenius, H. and Jutta Zeisler, Ferdinando Squitieri, Biaoyang Lin, Ann Bassett, Elizabeth Almqvist, Thomas D. Bird, and M. R. H. (1994) ‘A Worldwide Study of the Huntington’s Disease Mutation: The Sensitivity and Specificity of Measuring CAG Repeats’, *The New England Journal of Medicine*, 331(1), pp. 5–9. doi: 10.1056/NEJM199409293311301.

Kunkel, T. a (1993) ‘Slippery DNA and diseases’, *Nature*, 365(6443), pp. 207–208. doi: 10.1038/365207a0.

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. and et al. (2012) ‘CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion’, *Neurology*, 78, pp. 690–695. doi: 10.1212/WNL.0b013e318249f683.

Leeflang, E. P., Tavaré, S., Marjoram, P., Neal, C. O. S., Srinidhi, J., MacDonald, M. E., De Young, M., Wexler, N. S., Gusella, J. F. and Arnheim, N. (1999) ‘Analysis of germline mutation spectra at the Huntington’s disease locus supports a mitotic mutation mechanism’, *Human Molecular Genetics*, 8(2), pp. 173–183. doi: 10.1093/hmg/8.2.173.

Leeflang, E. P., Zhang, L., Tavarev, S., Hubert, R., Srinidhi, J., Macdonald, M. E., Myers, R. H., Young, M. De, Wexlee, N. S., Gusella, J. F. and Arnheimls, N. (1995) ‘Single sperm analysis of the trinucleotide repeats in the Huntington ’ s disease gene : quantification of the mutation frequency spectrum’, *Human Molecular Genetics*, 4(9), pp. 1519–1526.

Leung, C. M., Chan, Y. W., Chang, C. M., Lyu, Y. and Chen, C. N. (1992) ‘Huntington ’ s disease in Chinese : a hypothesis of its origin’, *Neurology*, 55, pp. 681–684.

Li, G. M. and Modrich, P. (1995) ‘Restoration of mismatch repair to nuclear

extracts of H6 colorectal tumor cells by a heterodimer of human MutL homologs.', *Proceedings of the National Academy of Sciences of the United States of America*, 92(6), pp. 1950–1954. doi: 10.1073/pnas.92.6.1950.

Li, J.-L., Hayden, M. R., Almqvist, E. W., Brinkman, R. R., Durr, A., Dodé, C., Morrison, P. J., Suchowersky, O., Ross, C. A., Margolis, R. L. and et al. (2003) 'A Genome scan for modifiers of age at onset in Huntington disease: The HD MAPS study', *American journal of human genetics*. Elsevier, 73(3), pp. 682–687. doi: 10.1086/378133.

Li, J.-L., Hayden, M. R., Warby, S. C., Durr, A., Morrison, P. J., Nance, M., Ross, C. A., Margolis, R. L., Rosenblatt, A., Squitieri, F. and et al. (2006) 'Genome-wide significance for a modifier of age at neurological onset in Huntington's disease at 6q23-24: the HD MAPS study.', *BMC medical genetics*, 7(71). doi: 10.1186/1471-2350-7-71.

Lipkin, S. M., Wang, V., Jacoby, R., Banerjee-Basu, S., Baxevanis, A. D., Lynch, H. T., Elliott, R. M. and Collins, F. S. (2000) 'MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability', *Nature Genetics*, p. 27–35. doi: 10.1038/71643.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Danni, L., Lu, L. and Law, M. (2012) 'Comparison of Next-Generation Sequencing Systems', *Journal of Biomedicine and Biotechnology*, 2012. doi: 10.1155/2012/251364.

MacDonald, M. E., Barnes, G., Srinidhi, J., Duyao, M. P., Ambrose, C. M., Myers, R. H., Gray, J., Conneally, P. M., Young, A. and Penney, J. (1993) 'Gametic but not somatic instability of CAG repeat length in Huntington's disease.', *Journal of medical genetics*, 30(12), pp. 982–6. doi: 10.1136/jmg.30.12.982.

Mackay, C., Cile, A.-C., Clais, D., Lundin, C., Agostinho, A., Deans, A. J., Macartney, T. J., Hofmann, K., Gartner, A., West, S. C., Helleday, T., Lilley, D. M. J. and Rouse, J. (2010) 'Identification of KIAA1018/FAN1, a DNA Repair Nuclease Recruited to DNA Damage by Monoubiquitinated FANCD2', *Cell*, 142, pp. 65–76. doi: 10.1016/j.cell.2010.06.021.

Manley, K., Shirley, T. L., Flaherty, L. and Messer, A. (1999) 'Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice', *Nature Genetics*, 23(4), pp. 471–473. doi: 10.1038/70598.

Marti, E. (2016) 'RNA toxicity induced by expanded CAG repeats in Huntington ' s disease', *Brain pathology*, 26, pp. 779–786. doi: 10.1111/bpa.12427.

McMurray, C. (2010) 'Mechanisms of trinucleotide repeat instability during human development', *Nature Reviews Genetics*, 11(11), pp. 786–799. doi: 10.1038/nrg2828.Mechanisms.

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shawand, P. D. and Marshall, D. (2013) 'Using tablet for visual exploration of second-generation sequencing data', *Briefings in Bioinformatics*, 14(2), pp. 193–202. doi: 10.1093/bib/bbs012.

Morales, F., Couto, J. M., Higham, C. F., Hogg, G., Cuenca, P., Braida, C., Wilson, R. H., Adam, B., Del Valle, G., Brian, R., Sittenfeld, M., Ashizawa, T., Wilcox, A., Wilcox, D. E. and Monckton, D. G. (2012) 'Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity', *Human Molecular Genetics*, 21(16), pp. 3558–3567. doi: 10.1093/hmg/dds185.

Morales, F., Vásquez, M., Santamaría, C., Cuenca, P., Corrales, E. and Monckton, D. G. (2016) 'A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients', *DNA Repair*, 40, pp. 57–66. doi: 10.1016/j.dnarep.2016.01.001.

Morovvati, S., Nakagawa, M., Osame, M. and Karami, A. (2008) 'Analysis of CCG Repeats in Huntingtin Gene among HD Patients and Normal Populations in Japan', 39, pp. 2007–2009. doi: 10.1016/j.arcmed.2007.06.015.

Musova, Z., Mazanec, R., Krepelova, A., Ehler, E., Vales, J., Jaklova, R., Prochazka, T., Koukal, P., Marikova, T., Kraus, J., Havlovicova, M. and Sedlacek, Z. (2009) 'Highly unstable sequence interruptions of the CTG repeat in the

myotonic dystrophy gene', *American Journal of Medical Genetics, Part A*, 149A(7), pp. 1365–1374. doi: 10.1002/ajmg.a.32987.

Myers, R. H., MacDonald, M. E., Koroshetz, W. J., Duyao, M. P., Ambrose, C. M., Taylor, S. A. M., Barnes, G. and J. Srinidhi, C. S. Lin, W. L. Whaley, A. M. Lazzarini, M. Schwarz, G. Wolff, E. D. Bird, J.-P. G. V. & J. F. G. (1993) 'De novo expansion of a (CAG)n repeat in sporadic Huntington's disease', *Nature genetics*, 3, pp. 73–96. doi: 10.1038/ng0293-165.

Nahhas, F. A., Garbern, J., Krajewski, K. M., Roa, B. B. and Feldman, G. L. (2005) 'Juvenile onset Huntington disease resulting from a very large maternal expansion', *American Journal of Medical Genetics*. doi: 10.1002/ajmg.a.30891.

Nance, M. A., Mathias-Hagen, V., Breningstall, G., Wick, M. J. and McGlennen, R. C. (1999) 'Analysis of a very large trinucleotide repeat in a patient with juvenile Huntington ' s disease', *Neurology*, 52(2), pp. 392–4.

Nance, M. a and Myers, R. H. (2001) 'Junvenile Onset Huntington's Disease-Clinical and Research Perspectives', *Mental Retardation and Developmental Disabilities Research Reviews*, 7(3), pp. 153–157. doi: 10.1002/mrdd.1022.

Palombo, F., Iaccarino, I., Nakajima, E., Ikejima, M., Shimada, T. and Jiricny, J. (1996) 'hMutS$_{NL}$, a heterodimer of hM SH2 and hM SH3, binds to insertion / deletion loops in DNA', *Current Biology*, 6(9), pp. 1181–1184.

Paradisi, I., Hernández, A. and Arias, S. (2008) 'Huntington disease mutation in Venezuela: Age of onset, haplotype analyses and geographic aggregation', *Journal of Human Genetics*, 53, pp. 127–135. doi: 10.1007/s10038-007-0227-1.

Pearson, C. E. (2003) 'Slipping while sleeping? Trinucleotide repeat expansions in germ cells', *TRENDS in Molecular Medicine*, 9(11), pp. 490–5. doi: 10.1016/j.molmed.2003.09.006.

Pearson, C. E., Tam, M., Wang, Y.-H., Montgomery, S. E., Dar, A. C., Cleary, J. D. and Nichol, K. (2002) 'Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions.', *Nucleic acids research*.

Oxford University Press, 30(20), pp. 4534–47.

Pêcheux, C., Mouret, J. F., Dürr, A., Agid, Y., Feingold, J., Brice, A., Dodé, C. and Kaplan, J. C. (1995) 'Sequence analysis of the CCG polymorphic region adjacent to the CAG triplet repeat of the HD gene in normal and HD chromosomes.', *Journal of medical genetics*, 32(5), pp. 399–400.

Pennuto, M. (2010) 'Pathogenesis of Polyglutamine Diseases', (December). doi: 10.1002/9780470015902.a0021486.

Pinto, R. M., Dragileva, E., Kirby, A., Lloret, A., Lopez, E., St. Claire, J., Panigrahi, G. B., Hou, C., Holloway, K., Gillis, T., Guide, J. R., Cohen, P. E., Li, G. M., Pearson, C. E., Daly, M. J. and Wheeler, V. C. (2013) 'Mismatch Repair Genes Mlh1 and Mlh3 Modify CAG Instability in Huntington's Disease Mice: Genome-Wide and Candidate Approaches', *PLoS Genetics*, 9(10). doi: 10.1371/journal.pgen.1003930.

Pramanik, S., Basu, P., Gangopadhaya, P. K., Sinha, K. K., Jha, D., Sinha, S., Das, S. K., Maity, B. K., Mukherjee, S. C., Roychoudhuri, S., Majumder, P. P. and Bhattacharyya, N. P. (2000) 'Analysis of CAG and CCG repeats in Huntingtin gene among HD patients and normal populations of India', *European Journal of Human Genetics*, 8(678–682).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J. and Sham, P. C. (2007) 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *The American Journal of Human Genetics*, 81, pp. 559–575. doi: 10.1086/519795.

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. and Gu, Y. (2012) 'A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers', *BMC Genomics*. BMC Genomics, 13(341). doi: 10.1186/1471-2164-13-341.

Raskin, S., Allan, N., Teive, H. A., Cardoso, F., Haddad, M. S., Levi, G., Boy, R.,

Junior, J. L., Sotomaior, V. S., Janzen-Dück, M., Jardim, L. B., Fellander, F. R. and Andrade, L. A. (2000) 'Huntington disease: DNA analysis in Brazilian population.', *Arquivos de neuro-psiquiatria*, 58(4), pp. 977–985.

Rawlins, M. D., Wexler, N. S., Wexler, A. R., Tabrizi, S. J., Douglas, I., Evans, S. J. W. and Smeeth, L. (2016) 'The prevalence of huntington's disease', *Neuroepidemiology*, 46(2), pp. 144–153. doi: 10.1159/000443738.

De Rooij, K. E., De Koning Gans, P. A., Roos, R. A., Van Ommen, G. J. and Den Dunnen, J. T. (1995) 'Somatic expansion of the (CAG)n repeat in Huntington disease brains.', *Human genetics*, 95(3), pp. 270–4.

De Rooij, K. E., De Koning Gans, P. a, Skraastad, M. I., Belfroid, R. D., Vegter-Van Der Vlis, M., Roos, R. a, Bakker, E., Van Ommen, G. J., Den Dunnen, J. T. and Losekoot, M. (1993) 'Dynamic mutation in Dutch Huntington's disease patients: increased paternal repeat instability extending to within the normal size range.', *Journal of medical genetics*, 30(12), pp. 996–1002. doi: 10.1136/jmg.30.12.996.

Ross, C. A. and Tabrizi, S. J. (2011) 'Huntington ' s disease : from molecular pathogenesis to clinical treatment', *The Lancet Neurology*. Elsevier Ltd, 10(1), pp. 83–98. doi: 10.1016/S1474-4422(10)70245-3.

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

Semagn, K., Babu, R., Hearne, S. and Olsen, M. (2014) 'Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): Overview of the technology and its application in crop improvement', *Molecular Breeding*, 33, pp. 1–14.

Semaka, A., Collins, J. A. and Hayden, M. R. (2010) 'Unstable familial

transmissions of huntington disease alleles with 27-35 CAG repeats (intermediate alleles)', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 153(1), pp. 314–320.

Semaka, A., Kay, C., Doty, C. N., Collins, J. A., Tam, N. and Hayden, M. R. (2013) 'High frequency of intermediate alleles on huntington disease-associated haplotypes in British Columbia's general population', *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 162B(8), pp. 864–871. doi: 10.1002/ajmg.b.32193.

Shelbourne, P. F., Keller-McGandy, C., Bi, W. L., Yoon, S.-R., Dubeau, L., Veitch, N. J., Vonsattel, J. P., Wexler, N. S., Arnheim, N., Augood, S. J. and Augood, S. J. (2007) 'Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain', *Human Molecular Genetics*, 16(10), pp. 1133–1142. doi: 10.1093/hmg/ddm054.

Smogorzewska, A., Desetty, R., Saito, T. T., Schlabach, M., Lach, F. P., Sowa, M. E., Clark, A. B., Kunkel, T. A., Harper, J. W., Colaiácovo, M. P. and Elledge, S. J. (2010) 'A Genetic Screen Identifies FAN1, a Fanconi Anemia-Associated Nuclease Necessary for DNA Interstrand Crosslink Repair', *Molecular Cell*, 39(1), pp. 36–47. doi: 10.1016/j.molcel.2010.06.023.

Snell, R. G., MacMillan, J. C., Cheadle, J. P., Fenton, I., Lazarou, L. P., Davies, P., MacDonald, M. E., Gusella, J. F., Harper, P. S. and Shaw, D. J. (1993) 'Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease', *Nature Genetics*, 4(4), pp. 393–397. doi: 10.1038/ng0893-393.

Squitieri, F., Andrew, S. E., Goldberg, Y. P., Kremer, B., Spence, N., Zelsler, J., Nichol, K., Theilmann, J., Greenberg, J., Goto, J., Kanazawa, I., Vesa, J., Peltonen, L., Almqvist, E., Anvret, M., Telenius, H., Lin, B., Napolitano, G., Morgan, K. and Hayden, M. R. (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), pp. 2103–2114. doi: 10.1093/hmg/3.12.2103.

Squitieri, F., Gellera, C., Cannella, M., Mariotti, C., Cislaghi, G., Rubinsztein, D. C., Almqvist, E. W., Turner, D., Bachoud-Lévi, A. C., Simpson, S. A., Delatycki, M., Maglione, V., Hayden, M. R. and Di Donato, S. (2003) 'Homozygosity for CAG mutation in Huntington disease is associated with a more severe clinical course', *Brain*, 126(4), pp. 946–955. doi: 10.1093/brain/awg077.

Swami, M., Hendricks, A. E., Gillis, T., Massood, T., Mysore, J., Myers, R. H. and Wheeler, V. C. (2009) 'Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset', *Human Molecular Genetics*, 18(16), pp. 3039–3047. doi: 10.1093/hmg/ddp242.

Telenius, H., Kremer, B., Goldberg, Y. P., Theilmann, J., Andrew, S. E., Zeisler, J., Adam, S., Greenberg, C., Ives, E. J., Clarke, L. A. and Hayden, M. R. (1994) 'Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm.', *Nature genetics*, 6(4), pp. 409–414. doi: 10.1038/ng0494-409.

Telenius, H., Kremer, H. P. H., Theilmann, J., Andrew, S. E., Aimqvist, E., Anvret, M., Greenberg, C., Greenberg, J., Lucotte, G., Squitieri, F., Starr, E., Goldberg, Y. P. and Hayden, M. R. (1993) 'Molecular analysis of juvenile Huntington disease : the major influence on (CAG) n repeat length is the sex of the affected parent', *Human Molecular Genetics*, 2(10), pp. 1535–1540.

The U.S.–Venezuela Collaborative Research Project and 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), pp. 3498–3503. doi: 10.1073/pnas.0308679101.

Veitch, N. J., Ennis, M., McAbney, J. P., Shelbourne, P. F. and Monckton, D. G. (2007) 'Inherited CAG.CTG allele length is a major modifier of somatic mutation length variability in Huntington disease', *DNA Repair*, 6(6), pp. 789–796. doi: 10.1016/j.dnarep.2007.01.002.

Vonsattel, J. P., Myers, R. H., Stevens, T. J., Ferrante, R. J., Bird, E. D. and Richardson, E. P. (1985) 'Neuropathological classification of Huntington's disease.', *Journal of neuropathology and experimental neurology*, 44(6), pp.

559–77.

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(10), pp. 561–566. doi: 10.1038/ejhg.2010.229.

Warner, J. P., Barron, L. H. and Brock, D. J. (1993) 'A new polymerase chain reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded on Huntington's disease chromosomes', *Molecular and Cellular Probes*, 7(3), pp. 235–239. doi: 10.1006/mcpr.1993.1034.

Warner, J. P., Barron, L. H., Goudie, D., Kelly, K., Dow, D., Fitzpatrick, D. R. and Brock, D. J. (1996) 'A general method for the detection of large CAG repeat expansions by fluorescent PCR', *Journal of medical genetics*, 33(12), pp. 1022–1026.

Wheeler, V. C., Lebel, L. A., Vrbanac, V., Teed, A., te Riele, H. T. and MacDonald, M. E. (2003) 'Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum', *Human Molecular Genetics*, 12(3), pp. 273–281. doi: 10.1093/hmg/ddg056.

Zhong, N., Yang, W., Dobkin, C. and Brown, W. T. (1995) 'Fragile X gene instability: anchoring AGGs and linked microsatellites.', *American journal of human genetics*, 57(2), pp. 351–61.