Characterising the dynamics of repeat expansion in Huntington's disease using single-molecule long-read DNA sequencing

Anthony Lawrence Warland

A thesis submitted to Cardiff University for the degree of Doctor of Philosophy

February 2022

Thesis summary

Huntington's disease (HD) is a fatal neurodegenerative disease caused by the expansion of the CAG repeat in the huntingtin gene (*HTT*). The length of the CAG repeat is inversely correlated with the age at disease onset. However, onset varies considerably between individuals with the same repeat length, and other genetic variants have been identified as modifiers of age at onset of HD. These include SNPs in vicinity of *FAN1*, a nuclease involved in DNA repair, and changes to the sequence in and around the CAG repeat itself. Expansion of the *HTT* CAG tract from the inherited length is seen in both germline and somatic cells in HD. Striatal projection neurons exhibit the most somatic expansion and are also the cell type most susceptible to degeneration.

Repeat expansion is recapitulated in a neuronal cell model derived from an individual with juvenile HD and 109 CAGs, however, traditional methods of quantifying the repeat have limited accuracy at this size and provide no information about the sequence of the repeat. Short-read next-generation sequencing (NGS) platforms do not span repeats of this length and thus cannot provide the repeat size. Long-read NGS platforms can generate highly accurate reads of more than 20 kilobases, which is long enough to span the repeats found in these models.

In the first part of this thesis, I assess the utility of long-read PacBio sequencing in measuring the size and instability of the *HTT* CAG repeat in samples with various repeat lengths. In the second part of this thesis, I assess the utility of long-read PacBio sequencing in measuring the size, instability, and sequence of the *HTT* CAG repeat in a neuronal cell model of HD and conduct experiments looking at the effect of *FAN1* genotype and cell maturity on repeat length, instability, and sequence variation.

Acknowledgements

I would like to sincerely thank my supervisors Lesley Jones, Thomas Massey, and Nigel Williams. Lesley, for her unwavering support, encouragement, and guidance. Tom, for his helpful guidance, feedback and support, and Nigel, for his technical expertise and regular words of encouragement. I would also like to thank Nick Bray for overseeing and ensuring my progress in the project. Thanks also to the patients, families, carers, and clinical staff for the time and samples they donated that made this project possible. Thank you to the Medical Research Council who funded this studentship.

Many thanks go to members of the HD research community past and present without whose work mine would not have been possible, especially members of the Jones lab at Cardiff University, including Jasmine Donaldson for her fantastic work establishing the iPSC cell lines used in this project and training me in cell culture, Branduff McAllister for the MiSeq sequencing, analysis tools and all-round massive amount of help he gave me, Joseph Stone for keeping me sane while writing up, Peter Holmans and Sergey Lobanov for their statistical advice, Laura Heraty and Florence Gidney for showing me the art of small-pool PCR and Freja Sadler for joining me on the PhD journey. I owe much to the staff at the DPMCN, especially Jo Morgan for running my sequencing libraries, Ellis Pires for helping me to learn Python and troubleshooting my analysis pipeline. Thanks to you and to the rest of the MRC Centre core team for keeping the lab running.

I would also like to thank those people who have made Cardiff feel like home these last 4+ years, including my friends Ed, Sam, John, Nick, Anna, Jack and Tom. Thanks also to my friends and family back home and who are always there for me when I need them. A special thanks to my parents have always believed in me and supported me at every turn. Finally, thank you to Leona whose love and support has helped me more than I can say.

Table of Contents

Figure List

Table list

Chapter 1 : General introduction

1.1. A historical background to Huntington's disease

The first description of Huntington's disease was given by George Huntington in 1872 in a lecture to the Meigs and Mason Academy of Medicine in Middleport and was published two months later in the Philadelphia journal *The Medical and surgical Reporter* (Huntington 1872). In succinct detail, Huntington described the three key features of what he called 'hereditary chorea': (1) its pattern of inheritance, (2) its motor and psychiatric impairments and (3) its progressive nature and fatal outcome. While Huntington was not the first to describe the disease, his vivid description, based on several decades of continued contact with many affected individuals and their families, drew international attention to the disorder, which eventually became known as Huntington's disease (HD).

Huntington described the core features of HD's inheritance pattern in his original paper, however, it was only after the rediscovery of Gregor Mendel's work on inheritance in 1900 that the inheritance of HD was first described as an autosomal dominant disorder by Charles Davenport (Davenport 1912). In 1983, genetic linkage analysis narrowed down the location of the causative gene for HD to the short arm of chromosome 4 (4p16.3) (Gusella et al. 1983). In the 10 years that followed, further genetic mapping - made possible with the help of many HD families, including an especially large HD kindred from Venezuela (MacDonald et al. 1993; Wexler 2004), enabled the causative mutation for HD to be identified as an expanded trinucleotide (CAG) repeat in the huntingtin gene *HTT* (MacDonald et al. 1993). Since the discovery of the first repeat expansion disorder, SBMA, in 1991 (La Spada et al. 1991), many other DNA repeat expansion diseases, including several CAG repeat expansion diseases, have been discovered and characterised (Khristich and Mirkin 2020) and new sequencing technologies are allowing more repeat disorders to be discovered, e.g. CANVAS (Huin et al. 2021).

In the decades that followed the identification of the HD-causing mutation, much insight has been gained into the genetic factors which influence the timing of motor onset and clues as to how these result in the neuropathology of HD have emerged.

1.2. Clinical features of Huntington's disease

Huntington's Disease (HD) is a progressive neurodegenerative disease characterised by involuntary jerky movements, psychiatric disturbances and dementia (Bruyn 1968). Onset of motor symptoms is most commonly between 30 to 50 years of age (Bates et al. 2014) but has been observed much earlier and later, with juvenile cases (age at onset < 20 years) occurring at a rate of approximately 5% (Quarrell et al. 2012). Life expectancy is highly variable but is typically around 10-30 years from the age at motor onset (AMO). Clinical diagnosis of HD relies on a family history of the disease or by genetic analysis (Craufurd et al. 2015).

Motor onset and clinical diagnosis is often preceded by a variable 'premanifest' period of up to 15 years which is characterised by a subtle decline in functional abilities (Tabrizi et al. 2013; Scahill et al. 2020). Motor, cognitive and psychiatric disturbances worsen gradually but irreversibly until death which is, on average, 18 years after motor onset (Douglas et al. 2013). There are currently no treatments which slow the progression of HD, however, patients are offered care for the management of symptoms to maximise quality of life.

The motor symptoms of HD are the most characteristic sign of the disease and the most common form of unwanted movements in HD is chorea. Derived from the Greek term for "dance", chorea are erratic, rapid, involuntary movements of the limbs, face and trunk (Bates et al. 2014) and manifest in most but not all HD patients. Other motor symptoms common to HD include some form of hypokinesia, i.e. decreased overall bodily movement, including akinesia (slowness in starting a movement) and bradykinesia (slowness in executing a movement). Patients often exhibit a combination of hypokinesia and the hyperkinesia of choreatic movements. Choreatic movements and the resulting increased muscle tone can lead to twisting and turning in all voluntary muscles. This is known as dystonia which, in combination with chorea, can significantly impair other motor activities such a walking. Speech and swallowing difficulties are also a common symptom of HD as involuntary motor control is disrupted (Skodda et al. 2014). Tics, which are rapid movements mainly of the arms and face, are sometimes present and often co-occur with of one or more of the unwanted movements described above. The exact pattern of motor symptoms and their timing is different in every HD patient though some, or all, of the above symptoms usually manifest at some point throughout the disease course.

Cognitive impairment, i.e., slowed thinking is often the first symptom to manifest in HD patients (Langley et al. 2021). A decline in cognitive capacities is followed by reduced attention control and memory function. Additional to this, a loss of insight into one's own function, time and location can take place resulting in difficulties in planning. It is thought that complete dementia can occur in the later stages of the disease although the communication problems of HD patients make this difficult to establish reliably (Lemiere et al. 2004; Paulsen and Conybeare 2005).

Figures for the prevalence of psychiatric and behavioural symptoms in HD are highly variable, however the most common are depression, anxiety, irritability, and apathy. Less common, although still elevated compared to the general population, are obsessive compulsive disorder and psychosis. Neuropsychiatric symptoms often occur prior to the motor onset in HD (van Duijn et al. 2007; van Duijn et al. 2008). Due to the elevated risk of suicide in HD patients, occurring in up to 7% of cases compared to 1% in non-mutation carriers (Cardoso 2017), and the loss of function associated with them, it is thought that the psychiatric symptoms of HD have the biggest impact on the quality of life of patients and their caregivers.

In addition to the three core symptom domains in HD, further secondary symptoms are associated with the disease. These include loss of body mass, sleep disruption and autonomic nervous system dysfunction (Bates et al. 2014). Weight loss is the most commonly reported of these and can occur prior to motor onset (Mochel et al. 2007; Cardoso 2017). The rate of weight loss is positively associated with the *HTT* CAG length (Aziz et al. 2008), and a higher body mass is associated with a slower rate of disease progression (Myers et al. 1991). Sleep disturbance is also very common with an estimated prevalence of up to 70% among HD patients (Arnulf et al. 2008) and often starts before the onset of motor symptoms (Lazar et al. 2015). Symptoms related to autonomic nervous system dysregulation are numerous and present at a range of rates. These include hyperhidrosis, incontinence, chronic pain, difficulty swallowing, and heat and cold intolerance (Bates et al. 2014).

1.3. The genetics of Huntington's disease

1.3.1. The huntingtin gene and mutation

HD is caused by a trinucleotide (CAG) repeat expansion in exon 1 of the *HTT* gene. A repeat size of 40 CAGs or more is completely penetrant for HD, with longer repeats

generally resulting in earlier motor onset (Andrew et al. 1993) (Figure 1.1). Repeat size explains approximately 50% of the variation in AMO (Andrew et al. 1993), though this association is as high as 70% in Venezuelan kindreds (Wexler et al. 2004). Despite this, patients with the same number of repeats can vary in onset age by more than 80 years (source REGISTRY-HD database: https://clinicaltrials.gov/ct2/show/NCT01590589 accessed 05/01/2020). Of the remaining variation in AMO, approximately 40% is heritable (Wexler 2004).

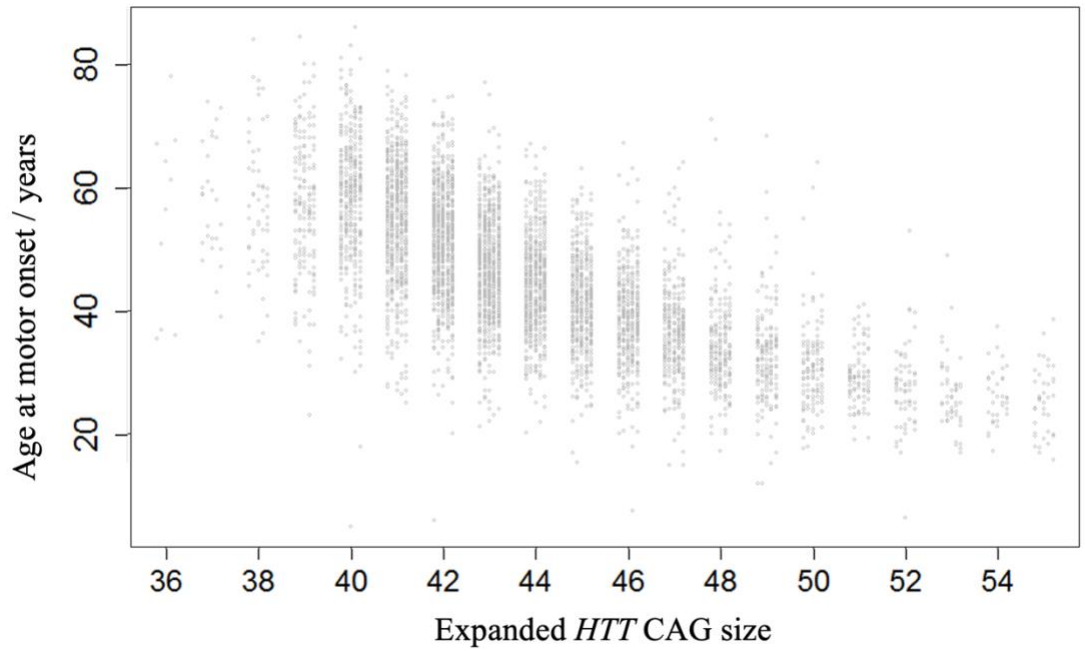


Figure 1.1. Age at motor onset is inversely correlated to *HTT* CAG repeat size. N = 6313. Each dot represents one Huntington's disease patient from the REGISTRY-HD database. Figure adapted with permission from McAllister 2021.

Below 40 repeats, the penetrance of HD depends on the uninterrupted CAG length, with 36-39 repeats resulting in partial penetrance (see Figure 1.2), which only affects a subset of individuals and often with a milder HD phenotype in those that do get the disease (Rubinsztein et al. 1996).

27-35 repeats represents an intermediate length which does not result in the disease but is associated with an elevated risk of entering the pathogenic range in subsequent generations (Migliore et al. 2019). Such intergenerational expansion, or sporadic *de novo* HD, is also dependent on the exact repeat length in the parent, with longer CAGs more liable to expand in the gametes (Semaka and Hayden 2014).

Gametic CAG repeat instability also results in the phenomenon of anticipation in HD, whereby the affected offspring of those with a disease allele tend to have an earlier

disease onset that their parents (Ranen et al. 1995). Alleles transmitted by the male parent showed the largest repeat number increases (Aziz et al. 2011). Indeed, analysis of single sperm found that > 90% of disease-range alleles showed an increase in repeat size (Leeflang et al. 1995), suggesting that repeat instability arises in the germline rather than post-zygotically. Yoon *et al.* show that repeat instability in HD can occur before, during and after meiosis, with the largest expansions observed in post-meiotic germline cells (Yoon et al. 2003). Somatic instability also occurs in HD and other repeat disorders and is thought to be a key driver of pathogenesis. The evidence for this is explored in section 1.3.3.

Below the intermediate range lies the normal allele range at 13-26 CAGs, as shown in Figure 1.2. Carriers of two normal alleles are unaffected by HD. HD patients typically have one normal length allele and one allele of 40 repeats or more, although homozygosity for the expanded repeat has been observed. Squitieri et al. found that homozygosity did not lower age of onset, but was associated with a more severe phenotype and an accelerated rate of disease progression (Squitieri et al. 2003).

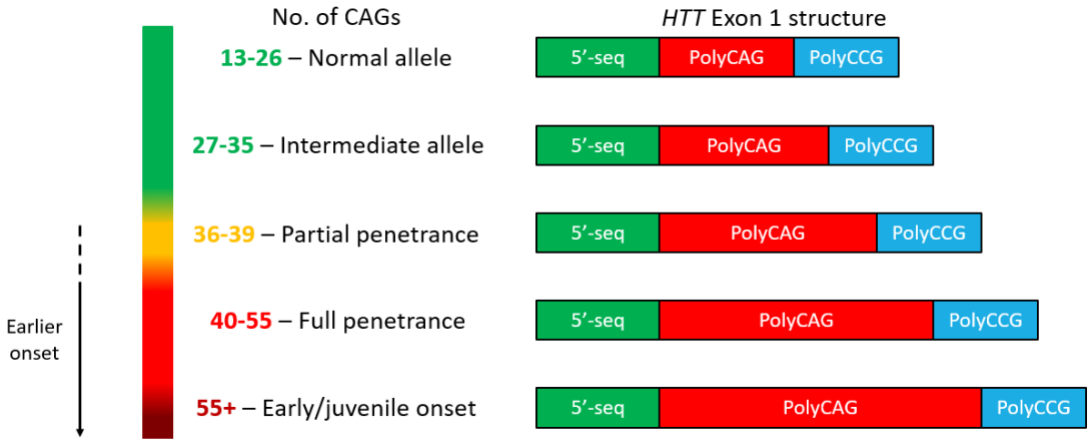


Figure 1.2. *HTT* CAG repeat length and disease penetrance in HD.

The prevalence of HD is highly variable between different ethnic and geographic populations; worldwide prevalence is estimated to be approximately 3 individuals per 100,000, while prevalence in Western populations are estimated at 10.6-13.7 individuals per 100,000 (Pringsheim et al. 2012; Bates et al. 2015).

1.3.2. Genetic modifiers of HD

A study by Wexler et al. of 83 Venezuelan HD kindreds encompassing 18,149 individuals found that ~40% of the variation in AMO not caused by CAG repeat length is heritable. A subset of ~4000 of the most-at risk individuals was genotyped at the

HTT repeat locus, with ~1000 identified as carriers of an HD-conferring expanded repeat. The log age of HD onset was regressed on repeat length to generate a residual age at onset. Variance-components analysis found that 38% of the residual age at onset was attributable to genes other than the HD gene (Wexler et al. 2004).

A number of genome-wide significant single-nucleotide polymorphisms (SNPs) are associated with altered AMO after accounting for inherited uninterrupted CAG length (Lee et al. 2015; Moss et al. 2017; Lee et al. 2019), many of which are found in the DNA damage response network. The most recent genome-wide association study (GWAS) on 9,064 individuals affected by HD identified 21 genome-wide significant SNPs including 13 candidate genes, of which 11 are involved in DNA repair (Lee et al. 2019). The most significant hit was in *FAN1*, a DNA endo/exonuclease involved in inter-strand crosslink repair and the recovery of stalled replication forks (Liu et al. 2010; Chaudhury et al. 2014). SNPs in *MLH1* and *PMS1* were also among the top 5 most significant loci. *MLH1* partners with *PMS2* or *MLH3* – both of which are also tagged by genome-wide significant modifying SNPs – to form mismatch repair complexes MutLα and MutLβ respectively (Kadyrov et al. 2006; Lee et al. 2019). *PMS1* combines with *MLH1* in the MutLβ complex, however MutLβ does not support mismatch repair (Cannavo et al. 2007). Recent evidence shows *FAN1* binds to MLH1 via protein-protein interactions, sequestering MutL complexes that would otherwise promote CAG repeat expansion (Goold et al. 2021; Porro et al. 2021). Other DNA repair associated genes reaching genome-wide significance included *MSH2*, *MSH3*, *MSH6* – all of which are involved in mismatch recognition in MutS complexes (Owen et al. 2005) – and *LIG1*, a mismatch repair enzyme (Lee et al. 2019).

Four independent AMO-altering SNPs have been identified at the locus containing *FAN1*, two of which are associated with earlier AMO and are coding SNPs that specify missense variants. rs150393409 ($p = 1.6x10^{-17}$) specifies R507H and is associated with a 5.2-year hastening of motor onset, while rs151322829 ($p = 4.3x10^{-07}$) specifies R377W and is associated with a 3.8-year hastening of motor onset. The other two SNPs, rs35811129 and rs34017474, correspond with cis-eQTLs (expression quantitative trait loci) that are associated with increased cortical FAN1 expression and delayed AMO by 1.3 and 0.8 years respectively, suggesting FAN1 may be neuroprotective. Knockout of Fan1 in mouse models of HD increases somatic expansion of the CAG repeat (Loupe et al. 2020), a result which has now been

replicated in human induced pluripotent stem cells (iPSCs) (McAllister et al. 2022). Fan1 has also been shown to play a expansion-preventing role in a mouse model of another triplet repeat disorder, fragile X syndrome (Zhao and Usdin 2018), suggesting that there are common mechanisms underlying pathology across trinucleotide repeat expansion diseases.

As well as these trans-acting modifiers, cis-acting modifiers, i.e., variants in the expanded *HTT* CAG repeat sequence, have also been associated with altered AMO in the GeM-HD study and others (Ciosi et al. 2019; Lee et al. 2019; Wright et al. 2019; McAllister et al. 2022). The canonical sequence of the *HTT* CAG repeat adopts the following structure: $(CAG)_nCAACAG$ followed by a variable but repetitive proline-encoding sequence. This sequence structure is observed in around 95% of HD patient expanded repeats, with the two most common alterations observed being the loss of the CAACAG or its duplication (Figure 1.3) (Lee et al. 2019). Lee et al. found that when the length of the uninterrupted CAG is accounted for, CAACAG loss resulted in consistently earlier AMO, while CAACAG duplication resulted consistently later AMO. These findings are replicated by Black et al. and McAllister et al., with the former finding that CAACAG loss was associated with a 9.5 year earlier AMO (Findlay Black et al. 2020). Data from a related study showed CAACAG duplication was associated with a 4.2 year later AMO (Wright et al. 2019). McAllister et al. show that loss of the CAACAG is significantly associated with earlier onset with a mean change of -10.2 years, while duplication of the CAACAG is significantly associated with later onset with a mean change of +10.4 years. Ciosi et al. show CAACAG loss is associated with a 9-year earlier AMO but found no significant difference between carriers of a CAACAG duplication and the canonical sequence. Variation also occurs in the proline-encoding repeat sequence but this has no apparent effect on AMO (Panegyres et al. 2006). While these sequence alterations are rare (< 5% of HD expanded alleles), multiple studies have now found that they drive a small but significant amount of the variation in HD AMO not attributable to uninterrupted CAG length.

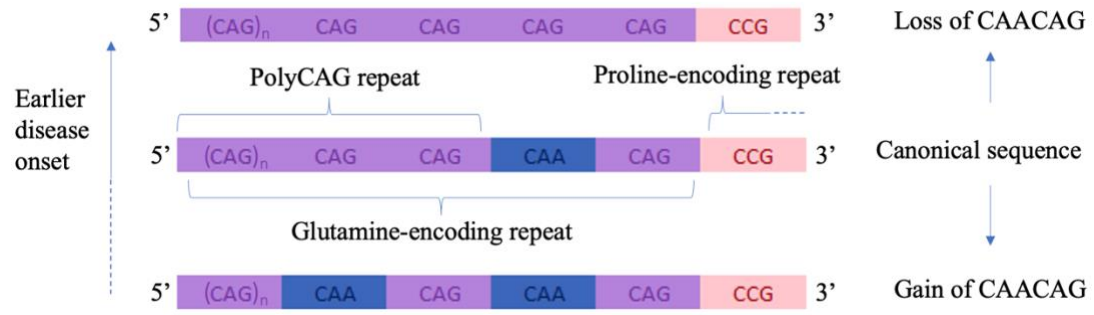


Figure 1.3. Sequence of the canonical *HTT* CAG repeat and common alterations.

1.3.3. CAG repeat instability

In addition to being gametically unstable the *HTT* CAG repeat is somatically unstable, exhibiting progressive increases in length throughout a patient's life (Kennedy et al. 2003). This effect is most pronounced in the liver and brain and especially in regions of the brain which are observed to atrophy during HD, namely the striatum and cortex (Kennedy et al. 2003; Shelbourne et al. 2007). An area of the brain which shows no pathology, the cerebellar cortex, displays the least CAG instability in adult onset HD, providing further evidence that repeat instability is tissue specific (Telenius et al. 1994). Evidence from HD patient brains is mirrored in mouse models of HD, in that somatic expansion is also correlated with brain degeneration, occurs in a tissue specific manner and the greatest expansions are observed in the liver, striatum and cortex (Mangiarini et al. 1997; Lee et al. 2011). It has been further shown in mouse models of HD that blocking repeat expansion pharmacologically slows the progression of the disease. Budworth et al. show that a mitochondrial-targeted scavenger of reactive oxygen species that suppresses motor decline inhibits somatic expansion via an oxidized base repair pathway involving OGG1 (Budworth et al. 2015), while Suelves et al. show that life-long treatment with an inhibitor of histone deacetylase 3 prevents long-term memory impairment and striatal repeat expansions (Suelves et al. 2017).

Longer inherited CAG repeats are more somatically unstable (Veitch et al. 2007) and somatic instability of the repeat in HD patient cortices is inversely correlated with AMO (Swami et al. 2009). Further to this, work by Ciosi et al. showed that the rate of expansion increases with age and that individuals with higher expansion scores in blood DNA have significantly worse HD outcomes (Ciosi et al. 2019).

Mechanisms of CAG repeat instability proposed so far are based on the CAG repeat's tendency to form unusual DNA structures, including stable hairpin loops, slipped-DNA, and G-quadruplexes (Mirkin and Frank-Kamenetskii 1994). Hairpin stability depends on DNA sequence and length and more stable structures are correlated with their propensity to expand in human disease (Gacy et al. 1995). Changes in CAG number may occur via a mismatch repair (MMR) pathway as MSH2, a component of the MutS complex involved in mismatch recognition, binds directly to the slipped-DNA conformation of CAG *in vitro* (Lang et al. 2011) and is required for repeat expansion in some cell-free assays (Stevens et al. 2013). The knockout of another protein involved in DNA MMR, Mlh1, ablates the repeat expansion observed in Fan1 knock-out HD mice (Loupe et al. 2020). Two recent studies show that FAN1 can inhibit repeat expansions by directly binding to and sequestering MLH1 (Goold et al. 2021; Porro et al. 2021).

Evidence from human tissues and mouse models indicates that somatic expansion may play a key role in disease pathogenesis as a driver of neuronal death, however the exact mechanism(s) involved remain unknown and further work must be done to establish the causes and effects of repeat expansion.

Research progress in this area relies on accurate methods of quantifying the CAG repeat length. One important consideration in this regard lies in the distinction between the polyglutamine repeat and the polyCAG repeat. Studies have shown that the length of the uninterrupted CAG tract drives AMO, independent of the number of additional consecutive glutamines (Ciosi et al. 2019; Lee et al. 2019). Current clinical standards for diagnosis of HD employ PCR-electrophoresis approaches to determine CAG repeat length (Losekoot et al. 2013), but sizing based on fragment size alone gives no information about polymorphisms within the CAG/CAA repeat, and is therefore unable to reliably call the length of the uninterrupted CAG repeat tract length in all patients. The accuracy of HD prognosis further relies on knowledge of these polymorphisms as alterations in the "CAACAG" cassette at the 3' end of the CAG repeat are associated with changes in AMO of up to 10 years (Ciosi et al. 2019; Mcallister 2019). Genotyping-by-sequencing provides this information and can be used to quantify somatic expansion, therefore allowing for some of the limitations of PCR-electrophoresis to be overcome (Ciosi et al. 2021).

1.4. Pathological features of Huntington's disease

The *HTT* CAG trinucleotide encodes the amino acid glutamine, meaning the *Huntingtin* protein (HTT) has an expanded glutamine tract, which may be responsible for some of the pathogenic consequences of an expanded repeat. HTT is expressed ubiquitously in both mice and humans, and is essential for normal development and brain function (White et al. 1997; Vonsattel and DiFiglia 1998). Homozygous knockout of the *HTT* gene is embryonic lethal in mice (Duyao et al. 1995; Nasir et al. 1995; Zeitlin et al. 1995). HTT's cellular roles are numerous and wide-ranging, including transcription, cell division, autophagy, vesicular transport and as a scaffold protein in the DNA damage response (Saudou and Humbert 2016; Maiuri et al. 2017).

Expansion of the *HTT* leads to expanded polyglutamine tracts and protein aggregates (Adegbuyiro et al. 2017). N-terminal HTT fragments form inclusions which have been observed in the brains (DiFiglia et al. 1997) and peripheral tissues in HD subjects (Sathasivam et al. 1999), however, these can be generated from both wild-type and mutant HTT (Goffredo et al. 2002) and the level of inclusions is not correlated to neuronal toxicity (Kim et al. 1999). Aggregation of HTT has been reported to have both protective (Gutekunst et al. 1999; Kuemmerle et al. 1999; Arrasate et al. 2004) and toxic (Davies et al. 1997; Liu et al. 2015; Woerner et al. 2016; Bäuerlein et al. 2017) effects on neurons. While it has yet to be conclusively disproven that HTT inclusions are a driver of HD pathology, the link between their appearance in the brains of HD patients and the widespread neuronal degeneration, which is a hallmark of HD, is still poorly understood.

HD results in the bilateral atrophy of the striatum, established by post-mortem examination of human brains (de la Monte et al. 1988; Aylward et al. 1998; Vonsattel and DiFiglia 1998). The basal ganglia, a set of forebrain structures associated with motor control, learning and cognition, receives its primary input from the striatum. (Graybiel 1998). Striatal degeneration has been observed to start over a decade before symptoms do (Tabrizi et al. 2009), progress at a constant rate (Tabrizi et al. 2013; Langbehn et al. 2019) and lead to some of the symptoms observed in HD (Alexander 1994; Kassubek et al. 2004). As the disease progresses, more widespread degeneration in the brain occurs. Progressive atrophy is also observed in the cerebral cortex and cerebellum and the brain as a whole (Ruocco et al. 2008; Johnson et al. 2021). The

rate of striatal, cerebellar and whole-brain atrophy are strongly associated with CAG repeat length (Ruocco et al. 2008; Langbehn et al. 2019).

The striatum is comprised primarily (~90-95%) of medium spiny projection neurons (MSNs) (Kita and Kitai 1988; Waldvogel et al. 2015). MSNs are the most vulnerable cell type to degeneration in HD (de la Monte et al. 1988; Aylward et al. 1998; Vonsattel and DiFiglia 1998), however the widespread atrophy observed in the brain is also likely to play a role in the disease. High instability of the *HTT* CAG repeat in MSNs may implicate somatic expansion as a mechanism underlying HD (Telenius et al. 1994).

1.5. Other repeat expansion disorders

HD is one of a family of disorders that have related neurological phenotypes caused by underlying pathological mechanisms. 48 repeat expansion diseases (REDs) are currently known, of which 16 are caused by CAG (or complementary CTG) repeat expansions, with more being discovered each year (Khristich and Mirkin 2020; Chintalaphani et al. 2021). Some element of neurological dysfunction occurs in all trinucleotide REDs suggesting that cells of the nervous system are particularly sensitive to trinucleotide repeat expansions (Orr and Zoghbi 2007). However, there is considerable phenotypic diversity between REDs, even within the CAG expansion diseases (Massey and Jones 2018). Neurodegeneration occurs in all the CAG/CTG expansion diseases, each with its own signature of neuronal atrophy. The exact number of repeats is important in most REDs, with longer repeats typically associated with earlier onset and disease severity (Khristich and Mirkin 2020). However, the thresholds at which they cause disease are different for each one- with coding repeats tending to have lower pathogenic thresholds than non-coding ones. The genomic location of the causal repeat also varies with each disease: they have so far been discovered in exons, introns and 5’- or 3’-UTRs. Interestingly, the expansion of one disease-causing repeat does not promote expansion of other genomic repeats. However, genetic modifiers which are associated with altering AMO in HD, especially those in DNA damage repair pathways, have also been seen to alter onset in other repeat diseases. This suggests they have a shared underlying pathogenic mechanism at the DNA level (Bettencourt et al. 2016).

Many REDs have pathogenic repeat thresholds of several hundred to several thousand tandem repeats. For example, the *C9orf72* variant of amyotrophic lateral sclerosis has a pathogenic range of 250-1600 GGGGCC repeats, while that of familial adult myoclonic epilepsy 1 is 440-3680 TTTCA/TTTTA repeats (Khristich and Mirkin 2020). Traditional repeat sizing methods cannot accurately size repeats in this range and also give no information about repeat interruptions which are known to play an important role in repeat instability in many REDs of varying repeat sequence, including HD (Ciosi et al. 2019), fragile-X syndrome (Pollard et al. 2004) and myotonic dystrophy type 2 (Liquori et al. 2001). Interestingly, in addition to HD, loss of the CAA repeat interruption is associated with earlier AMO in other CAG REDs such as SCA2 and SCA17 (Choudhry et al. 2001; Gao et al. 2008). Biophysical assays suggest that interrupted CAG repeats form shorter hairpin branches leading to reduced strand slippage and increased repeat stability (Xu et al. 2020). Methods that capture the sequence of pathogenic STRs would further our understanding of these important sources of genetic variation which are likely to be common mechanisms underlying many REDs.

1.6. Ways of measuring repeat loci

1.6.1. Introduction

Quantifying repeats is important in all REDs both as the basis of genetic testing and in gaining insight into disease pathology, however, because repetitive DNA behaves unlike heterogenous DNA and because many repetitive genomic regions have large repeats, they can be difficult to measure accurately using traditional molecular techniques. Furthermore, traditional methods of repeat sizing give no information about the sequence of the repeat, which is known to influence the disease course in many REDs.

The simplest methods of measuring repeat size are based on amplifying the repeat by polymerase chain reaction (PCR), however this can introduce errors including bias towards the amplification of smaller repeats, and "PCR stutter", arising from the polymerisation of repetitive loci (Daunay et al. 2019). Southern blotting avoids the use of PCR by utilizing digestion of large quantities of genomic DNA, but this is very labour intensive, low throughput and not able to provide an accurate repeat length (Massey et al. 2018).

The method currently used most widely for measuring repeat size is PCR-electrophoresis, also known as fluorescence PCR or fragment analysis. This high-throughput method is used clinically in genetic testing of HD and is accurate on repeats shorter than 40 CAGs (Losekoot et al. 2013). While there have been incremental advances in this method (Vnencak-Jones 2003) it is still essentially the same test used in the first genetic test for HD developed in 1993 (Warner et al. 1993). Because it is a bulk method (see following paragraph), the sensitivity is too low to detect rare alleles, i.e., those < 10% of the modal peak. Also, it gives no information about the sequence of the repeat. A modification of fluorescence PCR, triplet-primed PCR, is better suited to assessing longer repeats as the inclusion of a CAG-binding reverse primer generates a continuous ladder of products with which to size alleles, however this has the same limitations as the former method (Warner et al. 1996; Jama et al. 2013).

Most of the methods used to quantify repeats are referred to as 'bulk' methods, in which thousands of DNA molecules are assessed or used as PCR templates. Bulk methods allow for large changes in repeat number to be observed and can be reliably used to measure the modal CAG and somatic mosaicism in the major alleles in a sample, but they miss rare alleles including large expansions. One non-bulk method, small-pool PCR (SP-PCR), involves diluting the input DNA to a few molecules per reaction. This reduces the effect of amplification bias and enables detection of rare large expansions but is labour-intensive and contamination-prone (Massey et al. 2018; Ciosi et al. 2021).

Next-generation sequencing (NGS) technologies that generate short reads (~ 100-150 bp) can be used to assess short repeats including the wild type (WT) allele in HD but are not long enough to quantify expanded repeat sizes as they do not span the entire repeat tract. Massively parallel sequencing (MPS) approaches adapted to generate longer reads (up to 400 bp on Illumina's MiSeq) have been used to accurately quantify modal CAG and somatic mosaicism on alleles with 55 repeats though it is predicted this could be done on repeats of up to 90 CAGs, with an maximum allele detection limit of 123 CAGs (Ciosi et al. 2021).

There are methods for analysing repeats across the genome based on short-read whole-genome NGS paired with new bioinformatics tools (e.g. GangSTR) (Mousavi et al. 2018), which allow for all repeats across the whole genome to be genotyped and have recently led to the recent discovery of several REDs (Chintalaphani et al. 2021).

However, the limitations of PCR and short reads mean that not all STRs have sufficient coverage and it is not possible to determine the length of those repeats which exceed the short read length.

Repeat sequences can also be read using the original sequencing method, Sanger sequencing (Sanger et al. 1977), and is still routinely used as a reliable benchmark to validate data from newer technologies (Turner 2011). Machines such as ThermoFisher's 3730XL allow for the simultaneous loading and sequencing of up to 96 samples, routinely generating sequences, or reads, of up to 1000 bases with "per-base 'raw' accuracies as high as 99.999%"(Shendure and Ji 2008). Despite this, sample preparation is labour-intensive and time consuming, and the method relies on amplification of the DNA, often requiring multiple rounds. Read lengths and basecall qualities on repeats are often far lower than that of heterogeneous DNA, meaning repeats are often not read all the way through.

Long-read NGS technologies can read through repeat lengths far longer than their short read counterparts. Paired with accurate repeat sizing, high throughput and multi-locus protocols, these technologies have the potential to overcome the limitations of short read NGS and revolutionise the research, discovery, and diagnosis of REDs.

Measuring the repeat length and stability of STRs associated with REDs allows us to gain insight into the role of repeat expansion in these diseases. Doing so accurately is difficult, especially for longer repeat sizes, however technologies that have emerged over the last decade promised to deliver both accurate sizing and the sequence of the repeat.

1.6.2. Long-read sequencing

Long-read NGS technologies, so called 'third generation' platforms, are based on fundamentally different technologies to the previous two generations. Two truly long-read sequencing (LRS) technologies have been established to date: that of Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio).

PacBio's single-molecule real-time (SMRT) sequencing was the first long-read technology to achieve wide deployment (Pollard et al. 2018). SMRT sequencing involves Circular Consensus Sequencing (CCS), whereby double-stranded DNA samples are circularised by ligating hairpin adapters which have a polymerase primer binding site (Figure 1.4A). A polymerase binds to the primer and a microscopic

camera, or Zero-Mode Waveguide (ZMW), detects the incorporation of fluorescent bases opposite the template strand (Figure 1.4B). The time and wavelength signature of the resulting video is then converted into basecalls (Figure 1.4 C). PacBio's SMRT cells contain 1 or 8 million ZMWs per chip and enable highly parallel sequencing of DNA up to 100 kb long. The raw error rate of the technology is around 15% (Ardui et al. 2018), which is too high for most research purposes. This is overcome by passing over the target DNA multiple times, which is enabled by the circularisation. Raw data from each pass, or subread, is polished by building a consensus using a Hidden Markov Model algorithm that computes a log-likelihood for the most likely draft sequence. This log-likelihood is used to calculate a quality score for each base in the final consensus. The average of the per-base qualities is the "predicted" read accuracy. Phred quality scores, Q, are logarithmically related to the error probabilities assigned to all base calls. For example, a Q of 10 means the probability of an incorrect base call, *P*, is 0.1, while a Q of 60 is equivalent to $P = 0.000001$ (see Table 1.1). Reads with 7 passes typically have an error rate of around 1%, equivalent to Q20, with more passes resulting in lower error rates (Wenger et al. 2019). Reads with a quality score of Q20 or higher are defined as "HiFi" reads. PacBio's website states that a single SMRT Cell 1M run on their Sequel machine will generate up to 20Gb of raw data (https://www.pacb.com/products-and-services/sequel-system/previous-system-releases/ accessed 01/02/2022). Numbers from Ardui et al. suggest typical CCS data yields are between 3.65 and 5.11Gb per SMRT cell, based on average read length of 10-14,000 bp and number of CCS reads ~365,000, run on PacBio Sequel (Ardui et al. 2018). Read length is equal to insert size, although the number of passes achieved(and therefore the quality of base calls) is dependent on run time. Recent data from PacBio's website suggests HiFi reads averaging ~15 kb are routinely generated on the Sequel IIe System (https://www.pacb.com/smrt-science/smrt-sequencing/ accessed 25/01/2022).

Phred quality score (Q-score)	Probability of incorrect base call (*P*)	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Table 1.1 Phred quality scores and base calling accuracy.

The high accuracy of PacBio SMRT sequencing means STR length and sequence can be resolved, and detecting variants in the surrounding regions allows for the allelic phasing of STRs, such as that in *HTT* (Svrzikapa et al. 2020). Furthermore, data from Glasgow University was recently published demonstrating that SMRT sequencing can be used to determine the length and mosaicism in *HTT* repeats up to ~250 CAGs (Ciosi et al. 2021).

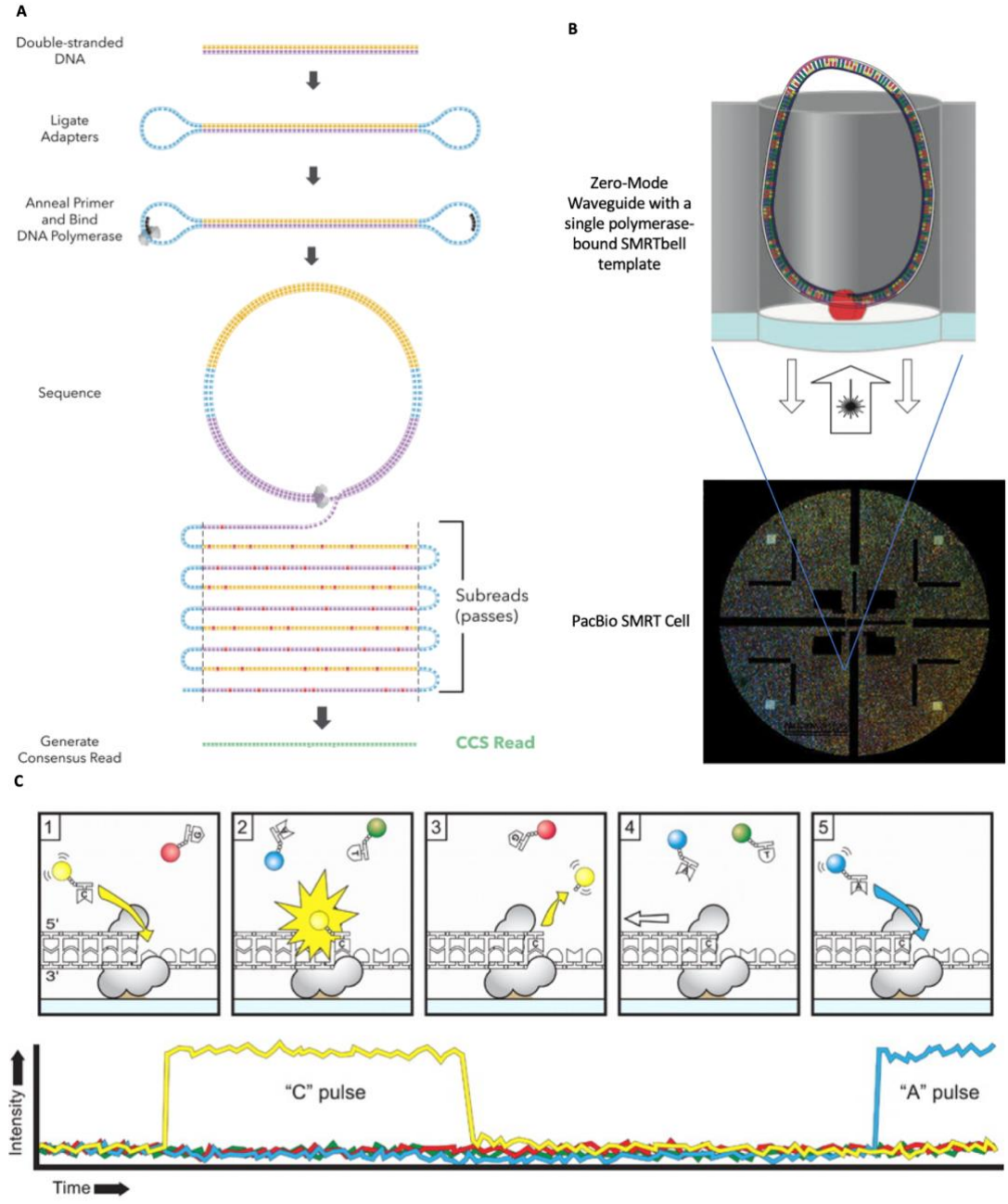


Figure 1.4. Long-read PacBio SMRT Sequencing. (A) PCR amplicons are circularised by ligating hairpin adapters to both ends to produce SMRTbell templates. Primers are annealed to the hairpins, enabling binding of a DNA polymerase. Polymerisation proceeds around the circularised DNA, passing repeatedly over the original template to produce subreads. In the downstream analysis, subreads are combined to build polished Circularised Consensus Sequencing (CCS) reads. (B) Polymerase-bound SMRTbell templates are captured in the wells of an array of Zero-Mode Waveguides (ZMWs), into which excitatory light is shone from below. (C) The incorporation of fluorescently labelled dNTPs by polymerisation results in changes in the emitted light. The changes are captured by a camera and ultimately converted into basecalls. Figure adapted with permission from Pacific Biosciences website (https://www.pacb.com).

The second long-read NGS technology to reach the market was Oxford Nanopore Technologies'. Chips in ONT's devices contain an array of protein nanopores embedded in a synthetic surface membrane, forming channels across which an electrical current is passed. Single strands of DNA loaded with a DNA helicase are

pulled through the nanopores by the ionic current with the helicase acting as a molecular brake. Changes in the current corresponding to the changing bases occupying the pore are amplified and converted into a signal. A recurrent neural network then converts the signal into base calls. The error rate of raw reads is around 15%, like PacBio, but this is mitigated substantially by high coverage, which allows a consensus sequence to be determined. ONT read lengths have no theoretical upper limit with reports of reads over 1Mb (Jain et al. 2018). This allows for the measurement of very long expanded repeat lengths: a study of the CCCCGG repeat in *C9orf72* showed that 80-99.5% of ONT reads completely spanned the repeat with a median of 406 repeats (Ebbert et al. 2018). Another study showed that the methylation status of repeats can be determined at the same time as DNA or RNA sequences, which is a marker for diagnosis in some REDs, such as Fragile X syndrome (Giesselmann et al. 2019).

Both ONT and PacBio's technologies sequence amplicons but can also be used with amplification-free methods based on targeting by CRISPR-Cas9 guide-ribonucleoproteins (Höijer et al. 2018; Gilpatrick et al. 2020). In PacBio's method, target loci are selectively cleaved before ligation to a magnetic adapter enabling subsequent isolation (Tsai et al. 2017). In ONT's method, sequencing adapters are ligated to cas9 cleavage sites allowing them to be selectively sequenced (Giesselmann et al. 2019). Because these technologies produce single molecule data, they can be used to assay the distribution of individual allele repeat lengths. Doing so will allow researchers to assess somatic instability and, given sufficient read depth, capture rare expansion and contraction events.

PacBio may be preferable for assaying rare large expansions in repeats of under 15 kb due to the high single molecule consensus read accuracies, whereas ONT may be preferrable for detecting repeats which are longer than 15 kb or applications where high single molecule accuracies are not needed, or methylation status is important. Both allow for the assay of multiple loci in a single run. Attributes of the two true long-read NGS technologies, as well as short-read NGS technologies and modern Sanger sequencing are shown in Table 1.2.

An alternative to these technologies, 10x Genomics, is a pseudo-long read technology that splits genome assembly into a handful of smaller assemblies before combining them into a full assembly. HMW double-stranded DNA molecules are separated into

droplets, or “GEMs”, each containing ~10 DNA molecules and an enzyme before fragmenting and barcoding them by emulsion PCR. The resulting short fragments are then sequenced by Illumina HiSeq technology. Barcodes allow orientation and position of fragments to be determined. Once these local assemblies have been performed, the pseudo-long reads are then assembled to make full-length contigs. While this technique is associated with the lowest cost and highest accuracy, at the time of writing all library preparation requires PCR and, because the method utilizes Illumina sequencing, it is hampered by GC-rich read dropout, and under-representation of STR sequences (Chintalaphani et al. 2021). More critical still is that it is not possible to map reads that start or end in a repeat, meaning long read technologies will still be needed to span the entire repeat tract.

Despite the advantages of long-read NGS, costs remain high, and the volume of data generated presents a challenge for storage and analysis. In addition, the length and complexity of some repeat sequences make it difficult to count repeats and determine interrupting sequences accurately. This is an active area of development and multiple software packages have been written for this purpose during the course of this project (Giesselmann et al. 2019; Mitsuhashi et al. 2019; De Roeck et al. 2019; Liu et al. 2020). Critically, because they can read through large, expanded repeats, long read sequencing is essential to obtain the size and sequence variation of repeats in some repeat diseases and HD cell and animal models.

Provider	Instrument	Sequencing chemistry	Raw accuracy/ %	Single-molecule consensus accuracy/ %	Max. reads per run	Max read length/ kbp	Max yield per run/ Gb	Sequencing run time/ hours	Cost per Gb/ £	Maximum CAG sizing	References
First generation											
Applied Biosystems	3730xl DNA Analyzer	Dideoxy chain termination	99.999	N/A	96	0.9	0.086	Up to 3	5 [c]	~50	(Shendure and Ji 2008)
Second generation											
Illumina	HiSeq 4000	Synthesis	99.6-99.8	N/A	1 x 10^{10}	0.15	1,500	24- 84	~4	~30	(Schirmer et al. 2016; Ciosi et al. 2021)
	MiSeq				5 x 10^{7}	0.4	15	56	~4	~90	
Third generation											
Pacific Biosciences	Sequel[a]	Synthesis	85-87	>99.999, dependent on insert size/ run time	5 x 10^{5} [b]	>100 (median 13.5)	20	Up to 20	30	~250	(Ardui et al. 2018; Wenger et al. 2019)
	Sequel II[a]				4 x 10^{6} [b]		160	Up to 30	10		
Oxford Nanopore Technologies	MinION	Nanopore	75-95	99.995	5 x 10^{5}	DNA length	30	Up to 72	~30	>> 250	(Wick et al. 2018; Kumar et al. 2019)
	GridION				2.5 x 10^{6}		150		~30		
	Prometh-ION				3.75 x 10^{8}		8,000	Up to 64	~6		

Table 1.2. Main attributes of major sequencing technologies. Metrics given are for bulk-PCR sequencing library preparations. Kbp: kilobase pairs. Gb: Gigabases. ng: nanograms. Note: except raw accuracy and unless stated otherwise, all information was accessed from references given in table or company websites on 29th September, 2020: https://www.thermofisher.com/order/catalog/product/3730XL#/3730XL , https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html, https://emea.illumina.com/systems/sequencing-platforms/hiseq-3000-4000/specifications.html, https://www.pacb.com/products-and-services/sequel-system/, https://nanoporetech.com/products/comparison. [a] : Parameters quoted are for one SMRT Cell. Up to 4 cells can be run in parallel. [b] : number of reads with accuracy of at least Q30 (99.9%) based on an insert size of 1kb with a 10-hour run time. Longer inserts yield fewer Q30 reads. [c] cost per sample.

1.7. HD iPSCs

To date, much of the modelling of HD has been conducted in mice. While this has yielded many useful insights into the nature of the repeat such as how it expands over time and in different tissues, as well as information about the role of the mutant huntingtin protein and genetic modifiers of the disease, there are major drawbacks to using these systems (Fisher and Bannerman 2019). Animal studies are time-consuming, labour-intensive, ethically contentious, costly and have limited applicability to human disease. By contrast, using human somatic cells which have been reprogrammed to pathologically relevant cell types can be cheaper, quicker and represent a more applicable genetic context for human diseases, and as such have great potential for disease modelling and clinical research (Hu et al. 2016). Since the method of generating induced pluripotent stem cells (iPSCs) was discovered in 2006 (Takahashi and Yamanaka 2006; Takahashi et al. 2007), they have been generated from patients with many different diseases including Parkinson's (Soldner et al. 2009), ALS (Dimos et al. 2008) and HD (The HD iPSC Consortium et al. 2012).

iPSC derived from patients with HD have now been differentiated to medium spiny neurons, the cell type most vulnerable in HD. These have been used to model the disease by numerous groups and show comparable phenotypes to those seen in both mouse models and human tissues (Camnasio et al. 2012; The HD iPSC Consortium et al. 2012; Mattis et al. 2014; The HD iPSC Consortium 2017; Goold et al. 2019). Despite a large range in the *HTT* CAG repeat lengths of HD-iPSCs studied so far, somatic expansion is commonly not observed (Zhang et al. 2010; Camnasio et al. 2012). In one line, however, 109-HD, derived from an individual with juvenile onset HD with 109 *HTT* CAG repeats, the expanded allele modal CAG increased from 110 to 118 over several passages (The HD iPSC Consortium et al. 2012).

In her work with 109-HD, Jasmine Donaldson generated a family of cell lines incorporating genetic manipulation by CRISPR-Cas9 to introduce a homozygous knock-out mutation to *FAN1*, the gene most strongly associated with modifying HD AMO (Donaldson 2019; GeM-HD Consortium 2019). Donaldson showed that lines with the knockout mutation exhibited significantly faster expansion rates than those with the wild-type genotype and that this is reproducible across several subclones. Repeats of this line have expanded over time, so the original 109 CAG tract is now ~130 CAGs.

1.8. Project aims

Long-read NGS has the potential to read through the large, expanded repeats found in animal and cell models of HD and provide the repeat size and sequence of single alleles. If shown to be the case, this would allow for novel experiments to be conducted on these models which investigate somatic expansion at the level of the DNA sequence. iPSCs are a renewable source of cells which allow for the assay of repeat expansion over multiple time points and can be engineered to conduct precise experiments relating to genetic modifiers of disease. The 109-HD line affords exciting opportunities to model repeat expansions but has so far been assayed using Sanger sequencing and fluorescence PCR only meaning no high-depth repeat-spanning sequencing data exists for this model. In this project I propose to use long-read PacBio sequencing to investigate the dynamics of repeat expansion and sequence variation in this model in the context of genetic modifiers of HD.

Primary objectives:

- Evaluate the accuracy of measures of the *HTT* CAG repeat derived from long-read PacBio sequencing compared to existing methods in HD patient samples.
- Evaluate the accuracy of measures of the *HTT* CAG repeat derived from long-read PacBio sequencing compared to existing methods in 109-HD iPSCs.
- Conduct experiments that examine the effect of *FAN1* genotype and cell maturity on repeat length, instability, and sequence variation in 109-HD iPSC samples.

Chapter 2 : Materials and methods

2.1 *FAN1* knock-out by CRISPR-Cas9

Genetic manipulation of iPSC clones was conducted by Jasmine Donaldson at Cardiff University. Briefly, two guide RNAs (gRNAs) targeting exon 2 of *FAN1*; 5'-CTGATTGATAAGCTTCTACG**AGG**-3' and 5'-GCACCATTTTACTGCAAACG**GGG**-3' were designed on DESKGEN cloud (www.deskgen.com) to produce a 95 bp deletion. crRNA and tracrRNA-ATTO-550 (IDT) were combined in nuclease-free duplex buffer (IDT), annealed (95°C, 2 minutes), combined with Cas9 (IDT) and incubated (RT, 20 minutes) to form a ribonucleoprotein (RNP). iPSCs were nucleofected with both RNPs using the 4D-Nucleofector and P3 Primary Cell 4D- Nucleofector X Kit, and program CA137 (Lonza). After 24 hours, iPSCs were sorted on the FACS ARIA Fusion to obtain the top 10% of cells, which were plated as single cells. After 7 days, individual colonies were manually dislodged and plated into single wells of a 96-well plate, which, after 7 days, were passaged into replicate plates using Gentle Cell Dissociation Reagent (STEMCELL Technologies). For screening DNA was extracted using QuickExtract (Cambio) (RT 10 minutes, 65°C 6 minutes, 95°C 2 minutes) and PCR amplified using two primer pairs amplifying exon 2 of *FAN1*; FAN-KO, 5'-CCTGTGTTTTATTGCTCAGAACA-3' and 5'-CATTTCATCAAGGTGCCGGT-3' and *FAN1*-T7, 5'-TCAGAGTTCGCTTTTCCCCT-3' and 5'-GATGCTAGGCTTCCCAAACA-3'. Amplicons were visualised on a 1.5% agarose gel on the Geldoc XR system (Bio-Rad). Sanger sequencing was used to confirm successful editing.

2.2 Cell culture

2.2.1 Lymphoblastoid cells

As described elsewhere (McAllister et al. 2022), lymphoblastoid cell lines from individuals with HD were cultured under standard conditions (www.coriell.org) in RPMI-1640 Glutamax (Thermofisher) supplemented with 15% fetal bovine serum and 1% penicillin/streptomycin, and passaged three times per week.

2.2.2 iPSCs

Human Q109 induced pluripotent stem cells (iPSCs) were generated by the HD-iPSC consortium from a human fibroblast line with an expanded CAG *HTT* allele of 109 repeats (The HD iPSC Consortium et al. 2012). Three clonal lines were generated from the Q109 line; Q109 N1, N4 and N5. Only the N1 line was used in this project. Figure 3.13 summarises all the available lines in a family tree, including all those sequenced in this project.

2.2.2.1 iPSC maintenance

The iPSCs were cultured on vitronectin-coated plates (0.5 μg/cm^2) (Life Technologies) in Essential 8 Flex medium (Life Technologies) under standard culturing conditions (37 ℃, 5% CO_2). Cells were passaged every 3-4 days, when reaching confluency of ~ 70%. For passaging, cells were incubated with ReLeSR (Stem Cell Technologies) for 1 minute at 37 ℃. After aspirating the ReLeSR, cells were dissociated into small clumps in fresh warmed medium and were seeded into a new plate at a density of 1:12.

For freezing, cells were dissociated with ReLeSR as described above, centrifuged at 1000 rpm for 3 minutes and resuspended in CryoStor CS10 (Stem Cell Technologies) with approximately 1x10^6 cells/ 0.5 mL CryoStor CS10. Cryovials containing the cell suspension were then transferred to a CoolCell Freezing Container (Corning) and placed at -80°C where cells were frozen at a rate of -1°C/minute.

For thawing iPSCs, cryovials were warmed to 37 ℃ in a water bath for 1-2 minutes until partially thawed. 1 mL of warm Advanced DMEM/F-12 (ADF) (Life Technologies) was then added to the cells dropwise. The cell suspension was then transferred to an Eppendorf, centrifuged at 1000 rpm for 3 minutes, and resuspended in warmed E8 Flex Medium containing 10μM Y-27632 dihydrochloride (Rock Inhibitor).

2.2.2.2 iPSC differentiation to striatal neurons

iPSCs were differentiated from pluripotent cells to striatal neurons by me as described elsewhere (McAllister et al. 2022). Briefly, iPSC colonies were dissociated into a single cell suspension using Accutase (Life Technologies), seeded into 12 well plates coated with Growth Factor Reduced Matrigel (0.5 ng/mL) (BD Biosciences) and cultured in Essential 8 Flex medium until the cells reached ~80% confluency. iPSCs

were differentiated to forebrain neurons using adaptations of published protocols (Telezhkin et al. 2016; Smith-Geater et al. 2020), as follows. iPSCs were induced into Neuronal Precursor Cells (NPCs) using Advanced DMEM/F-12 (ADF) (Life Technologies) supplemented with 1% Glutamax (Thermo Fisher), 1% Penicillin/ Streptomycin (5000U/5000 µg) (Gibco), 2% MACS neurobrew without retinoic acid (Miltenyi), 10 µM SB431542 (Miltenyi), 1 µM LDN-193189 (StemGent) and 1.5 µM IWR-1-endo (Miltenyi) up until day 8, upon which SB was omitted from the medium and 25 ng/ml Activin A (PeproTech) was added. Full media changes were performed daily up until day 16. Day 16 NPCs were passaged into plates coated with Poly-D-Lysine (Thermo Fisher) and Growth Factor Reduced Matrigel. Cells were fed with SJA medium consisting of ADF with 1% Glutamax, 1% Penicillin/ Streptomycin, 2% MACS neurobrew with retinoic acid, 2 µM PDO332991 (Bio-Techne), 10 µM DAPT (Bio-Techne), 10 ng/ml BDNF (Miltenyi), 0.5 µM LM22A4 (Bio-Techne), 10 µM Forskolin (Bio-Techne), 3 µM CHIR 99021 (Bio-Techne), 0.3 mM GABA, 1.8 mM $CaCl_2$ (Sigma-Aldrich) and 0.2 mM Ascorbic acid (Ascorbic Acid). After 7 days in SJA medium, cells were fed with SJB medium consisting of equal amounts of ADF and Neurobasal A (Life Technologies) with 1% Glutamax, 1% Penicillin/ Streptomycin, 2% MACS neurobrew with retinoic acid, 2 µM PDO332991, 10 ng/ml BDNF, 3 µM CHIR 99021, 1.8 mM $CaCl_2$ and 0.2 mM Ascorbic acid. After 14 days in SJB medium, cells received half media changes every 3-4 day with medium consisting of equal parts SJB medium and BrainPhys Neuronal Medium (STEMCELL Technologies) supplemented with 1% Penicillin/ Streptomycin, 2% MACS neurobrew with Vitamin A and 10 ng/ml BDNF.

Samples were pelleted by centrifuging at 1000 rpm for 3 minutes at day 16, 35, 50, 71 and frozen at -20℃ for downstream CAG repeat sizing.

2.2.2.3 SNP array genotyping of iPSC lines

To ensure no gross genomic rearrangements in the cell lines, SNP array genotyping (virtual karyotyping) was carried out in-house at Cardiff University by the core team technician Alexandra Evans as described elsewhere (McAllister et al. 2022). Briefly, genomic DNA was extracted using QIAamp DNA Mini Kit (Qiagen) and 200 ng (50 ng/ µL) used for genotyping. Samples were genotyped on the Infinium PsychArray-24 Kit (Illumina) or the Infinium Global Screening Array-24 (Illumina) and scanned using the iScan System (Illumina). Data were exported from Genome Studio and

analysed using PennCNV(Fang and Wang 2018). Sample level quality control was applied based on the standard deviation of Log R ratio set at 0.3, minimum SNP number of 10 and minimum region size of 100,000 bp.

2.2.3 Cell imaging

Cell images were taken using an LRS brightfield microscope at 10x magnification using an EVOS FL microscope (Life Technologies).

2.3 Nucleic acid extraction and quantification

2.3.1 DNA extraction

All DNA extraction was conducted with QIAGEN QIAamp DNA Mini Kits.

2.3.2 Nucleic acid quantification

All DNA quantified was conducted by PicoGreen (Invitrogen) or Qubit Fluorometers (Invitrogen).

2.4 CAG repeat sizing by fragment analysis

Sequencing data was validated against repeat counts generated by fragment analysis. Genomic DNA was isolated from cells using a QIAamp DNA mini kit (Qiagen) before being amplified by fluorescently labelled primers (forward: 5'-6-FAM-ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC-3', reverse: 5'-GGCGGCTGAGGAAGCTGAGGA-3) targeting the *HTT* repeat locus using the PCR recipe shown in Table 2.1.

Reagent	Volume / µL
TaKaRa LA Taq (5 U / µL)	0.1
GC Buffer II (2X)	5.0
dNTP Mixture (2.5 mM each)	1.6
Forward primer (10 µM)	0.5
Reverse primer (10 µM)	0.5
Nuclease-free water	1.3
DNA (25 ng)	1.0
Total	10

Table 2.1. PCR recipe used to prepare samples for fragment analysis.

Reactions were mixed by aspiration and centrifuged before being loaded in a thermal cycler and run with method shown in Table 2.2.

Cycles	Temperature / ˚C	Time / s
1	94	180
35	94	30
	65	30
	72	90
1	72	300
	4	Forever

Table 2.2. Thermal cycler method used to amplify samples for fragment analysis.

Once checked by gel electrophoresis (see 2.5.1.4), PCR products were mixed with Hi-Di™ Formamide and a sizing standard before a heat denaturation step. Samples were sized with the GeneScan LIZ600 dye Size Standard (Applied Biosystems) on a GA3130xL Genetic Analyser (Applied Biosystems) at The All Wales Medical Genomics Service (Institute of Medical Genetics, University Hospital of Wales, Cardiff) where the machine we use is validated for clinical use and therefore reliable. The machine separates fragments by capillary electrophoresis and detects the resulting fluorescent signal. Files were analysed by GeneMapper (Applied Biosystems) and Autogenescan (https://github.com/BranduffMcli/AutoGenescan), an algorithm adapted by Branduff McAllister from the R package Fragman (Covarrubias-Pazaran et al. 2016). Modal peak repeat sizes, expansion and instability indices were calculated using a 10% peak heigh threshold for all samples (Lee et al. 2010). All modal CAG values reported here represent the number of CAG triplets in the pure, uninterrupted CAG tract.

2.5 Sequencing library preparation

2.5.1 PacBio

2.5.1.1 Overview

A 3 kb locus around the *HTT* CAG repeat was amplified by PCR in genomic DNA of 48 HD patient samples for library 3000-LBC-PBMC. Library preparation and quality control was conducted by me in accordance with PacBio's SMRTbell™ library protocol (Pacific Biosciences 2018; Pacific Biosciences 2019), which is summarised in Figure 3.4B.

Briefly, 5' blocked primers and were used in the first of two rounds of amplification. Amplification size was verified by gel electrophoresis (see 2.5.1.3) before samples

were purified using Ampure PB beads (see 2.5.1.5). DNA concentrations were quantified using a Qubit fluorometer. The samples were normalised to 1 $ng.ul^{-1}$ and amplified again, this time using barcoded universal primers (available from PacBio, part number 100-466-100), with each sample being amplified by primers with a unique barcode. The resulting PCR products were verified by gel electrophoresis and purified as above. DNA concentrations were measured by Qubit and ~30 ng of each sample was pooled. SMRTbell Express Template Prep Kit 2.0 (PacBio part 100-938-900) was used to prepare up to 750 ng of the pooled library as per the amplicon sequencing protocol available on the PacBio website (Pacific Biosciences 2018; Pacific Biosciences 2019). All libraries were verified by capillary electrophoresis (see 2.5.1.6).

A high depth 3 kb amplicon library (library name: 3000-iPSC) of 6 iPSC samples was prepared in the same way. Later, the same 6 gDNA samples were used to generate an equivalent 600 bp amplicon library (600-iPSC-1). All subsequent libraries were comprised of 600 bp amplicons. 600 bp libraries were prepared in the same way as 3000 bp libraries, except 0.6x (rather than 1.0x) volumes of Ampure PB beads were in the two bead purification steps (see 2.5.1.5).

2.5.1.2 Primer design

First round primers TOM48B/49B, designed by Tom Massey, were used to generate all 3 kb libraries, while ANT1/2, designed by me, were used to generate all 600 bp iPSC libraries (Table 2.3, Figure 3.4A). First round primers included 5' amino methyl C6 blocker groups to prevent ligation of first-round PCR products (i.e., non-barcoded amplicons) to sequencing adapters. All primers were supplied by ThermoFisher Scientific, Kent, UK.

Primer name	Direction	Sequence
TOM48	FW	5'-GCAGTCGAACATGTAGCTGACTCAGGTCAC **CTGACACAGTGGACAAAGGC**
TOM49	REV	5'-TGGATCACTTGTGCAAGCATCACATCGTAG **AAACAAGTTCTCGCCCCAAC**
TOM50	FW	5'-GCAGTCGAACATGTAGCTGACTCAGGTCAC **TTTACTGGGCTCCTCTCTGC**
TOM51	REV	5'-TGGATCACTTGTGCAAGCATCACATCGTAG **AGCAACAGAAACCCCTAGCT**
TOM52	FW	5'-GCAGTCGAACATGTAGCTGACTCAGGTCAC **CTCCCATAAAGAAACGCCCC**
TOM53	REV	5'-TGGATCACTTGTGCAAGCATCACATCGTAG **GACACACAGACTTCCAGGGA**
TOM48B	FW	/5AmMC6/GCAGTCGAACATGTAGCTGACTCAGGTCAC **CTGACACAGTGGACAAAGGC**
TOM49B	REV	/5AmMC6/TGGATCACTTGTGCAAGCATCACATCGTAG **AAACAAGTTCTCGCCCCAAC**
ANT1	FW	/5AmMC6/GCAGTCGAACATGTAGCTGACTCAGGTCAC **GCGACCCTGGAAAAGCTGATGA**
ANT2	REV	/5AmMC6/TGGATCACTTGTGCAAGCATCACATCGTAG **AGCAGCGGCTGTGCCTGC**
ANT3	FW	/5AmMC6/GCAGTCGAACATGTAGCTGACTCAGGTCAC **GCCTTCGAGTCCCTCAAGTCC**
ANT4	REV	/5AmMC6/TGGATCACTTGTGCAAGCATCACATCGTAG **GGCTGAGGAAGCTGAGGAGG**
ANT5	FW	/5AmMC6/GCAGTCGAACATGTAGCTGACTCAGGTCAC **GCCGCTCAGGTTCTGCTTTTACC**
ANT6	REV	/5AmMC6/TGGATCACTTGTGCAAGCATCACATCGTAG **GCTCCTCAGCCACAGCCG**
ANT7	FW	/5AmMC6/GCAGTCGAACATGTAGCTGACTCAGGTCAC **CCAGAGCCCCATTCATTGCC**
ANT8	REV	/5AmMC6/TGGATCACTTGTGCAAGCATCACATCGTAG **CCAAACTCACGGTCGGTGCAG**

Table 2.3. Round 1 PCR amplification primers. FW: forward primer; REV: reverse primer. 5': 5-prime phosphate. /5AmMC6/: 5'amino methyl C6 blocker group. Non-bold sequences are tails comprising barcoded universal primer binding sites. Bold sequences are the *HTT*-specific annealing regions.

In silico product sizes for primer pairs in Table 2.3 were generated using the USCS In-Silico PCR tool (Kent et al. 2002) (Table 2.4). Primer annealing regions only were used with the default parameters: genome – human; assembly – hg38; max product size – 4000; min perfect match – 15; min good match – 15. All primer pairs were required to generate unique *in silico* products. To calculate the predicted size of first

round PCR products in Table 2.4 (column 2), 60 bp was added to the length of each *in silico* product, representing the 30 bp 5-prime universal flanking sequences in each first round primer pair (Table 2.3). Because the *HTT* gene of hg38 has a pure CAG tract of 19 repeats, 63 bp was added to all WT sizes to represent typical human HD expanded alleles with 40 repeats (Table 2.4, column 3). To calculate the size of a typical 109NI iPSC expanded allele with a pure CAG tract of 130 repeats, 333 bp was added to WT product sizes (Table 2.4, column 4).

Primers	WT allele (19 CAGs)	HD expanded allele (40 CAGs)	iPSC expanded allele (130 CAGs)
TOM48/49	2,904	2,967	3,237
TOM50/51	2,984	3,047	3,317
TOM52/53	2,840	2,903	3,173
ANT1/2	242	305	575
ANT3/4	194	257	527
ANT5/6	428	491	761
ANT7/8	422	485	755

Table 2.4. First round PCR product sizes based on *in silico* PCR for different *HTT* allele lengths. Product sizes are in base pairs. CAG sizes given are for the pure CAG repeat.

The method for the 600 bp iPSC library preparation was modified to introduce a size selection step designed to remove most of the WT allele amplicons and thus enrich for the expanded allele. Instead of using 1x Ampure PB beads as in the 3 kb library prep, amplified DNA was purified with 0.6x beads immediately after both PCR reactions.

2.5.1.3 Amplification of *HTT* Locus by PCR

All PCR reactions were conducted in 10 µL reactions using reagents from the TaKaRa Bio LA Taq with GC buffers kit (TaKaRa Bio Europe cat #RR02AG).

2.5.1.3.1 First round PCR

First round PCR primers used are listed in Table 2.4. Table 2.5 shows the general recipe used for all first-round amplification reactions. Wherever possible, reactions of samples to be pooled in the same library were prepared from the same PCR master mix comprising all reagents minus the template DNA.

Reagent	Volume (µL)
gDNA (10 ng/ µL)	2.5
GC Buffer 1 (2x)	5
dNTPs (2.5 mM each)	1.6
F primer (100 µM)	0.05
R primer (100 µM	0.05
LA Taq (5 U / µL)	0.1
Nuclease-free water	Up to 10

Table 2.5. Recipe used for First round PCR in PacBio sequencing libraries. F: forward. R: reverse.

After mixing by inversion and flicking, PCR tubes were briefly centrifuged before being loaded into the same T100 Thermal Cycler (Biorad). Table 2.6 shows the heating programme used.

Cycles	Temperature / ˚C	Time / s
1	94	90
30	94	30
	61.7	30
	72	150
1	72	600
1	20	Forever

Table 2.6. Programme used in first round PCR in PacBio libraries.

2.5.1.3.2 Second round

Second round PCR primers used were from PacBio's barcoded universal primers plate (part number 100-466-100). Table 2.7 shows the general recipe used for all second-round amplification reactions. Wherever possible, reactions of samples to be pooled in the same library were prepared from the same PCR master mix of all reagents minus the template DNA and primers.

Reagent	Volume (µL)
Template DNA (1 ng/ µL)	1
GC Buffer 1 (2x)	5
dNTPs (2.5 mM each)	1.6
F + R primers (10 x)	1
LA Taq (5 U / µL)	0.1
Nuclease-free water	Up to 10

Table 2.7. Recipe used for First round PCR in PacBio sequencing libraries. F: forward. R: reverse. Primers are pre-mixed in plate format. GC buffer, dNTPs and LA Taq supplied in kit (TaKaRa Bio Europe #RR02AG).

After mixing by inversion and flicking, PCR tubes were briefly centrifuged before being loaded into the same T100 Thermal Cycler (Biorad). Table 2.8 shows the heating programme used.

Cycles	Temperature / ˚C	Time / s
1	94	90
20	94	30
	64	30
	72	150
1	72	600
1	20	Forever

Table 2.8. Programme used in first round PCR in PacBio libraries.

2.5.1.4 Agarose gel electrophoresis

PCR products were run on 1% agarose gels made with 0.5 X Tris-Borate-EDTA (TBE) buffer and SYBR safe DNA stain used at 1 µL per 50 ml of gel. 1 µL of DNA Gel Loading Dye (6X) (Thermo Scientific) was added to 1 µL of each PCR product before loading. HyperLadder™ 1 (Bioline) was used to size 3 kb products, HyperLadder 100 bp for 600 bp products. 0.5 µL of ladder was loaded into the first and last lane of each gel. Gels electrophoresis was run in 0.5 X TBE at 66 V for 60 - 120 minutes.

2.5.1.5 Paramagnetic bead purification

DNA purification for PacBio libraries followed PacBio's Ampure PB purification protocol (Pacific Biosciences 2018), except 1.0x volumes of beads were used for 3000 bp libraries and 0.6x volumes of beads for 600 bp libraries. Briefly, beads were added to directly to PCR products and mixed thoroughly before binding at room temperature

on a rotating mixer for 10 minutes. Beads were placed on a magnetic rack and washed twice with freshly prepared 70% ethanol before being eluted in 10 ul of PacBio Elution Buffer at room temperature on a rotating mixer for 10 minutes.

2.5.1.6 Capillary electrophoresis

Pooled libraries were checked by capillary electrophoresis to ensure they conformed to the expected size and that they contained no non-specific products. Libraries were run on Agilent's Bioanalyser 2100 (High Sensitivity DNA kits 7000 and 12,000), a fluorescence-based capillary electrophoresis platform that measures the size and mass of DNA in a sample. Agilent's 5400 Fragment Analyser System, comparable to the Bioanalyser but with higher throughput and DNA size detection limit, was also used for library quality control purposes.

2.5.2 MiSeq

A targeted MiSeq NGS sequencing methodology was carried out as described elsewhere (McAllister et al. 2022). Briefly, DNA from low-passage lymphoblastoid cell lines was normalised in concentration using PicoGreen™ to 4 ng.µL^{-1}. Libraries were prepared in 384-plate format using MiSeq-compatible primers as described (Ciosi et al. 2018). PCR reactions used TaKaRa LA Taq® polymerase (RR02AG, TaKaRa) in TaKaRa GC Buffer II. Library clean-up consisted of two AMPure XP SPRI bead (Beckman Coulter, A63881) steps, the first at 0.6X and the second at 1.4X bead concentrations. Libraries were checked using a Bioanalyser (Agilent) with a high sensitivity DNA chip (Agilent, 5067-4626).

2.6 Sequencing

2.6.1 PacBio

All libraries except 600-iPSC-4 were run with an on-plate loading concentration of 8 pM. Library 600-iPSC-4 was run with on-plate loading concentration 8 pM in first SMRTcell and 24 pM in the second SMRTcell. All PacBio sequencing except library 600-iPSC-4 was run on a Sequel machine at Cardiff MRC Centre for Neuropsychiatric Genetics and Genomics (Cardiff University). Library 600-iPSC-4 was sequenced on a Sequel machine at the College of Life and Environmental Sciences (Exeter University). Loading and sequencing at Cardiff University was performed by Jo Morgan. Loading and sequencing at Exeter University was performed by Aaron

Jeffries. All PacBio sequencing was conducted on SMRT Cell 1M chips with a 10-hour run time.

CCS with demultiplexing analysis was run on supercomputing clusters (see 2.8.4) using PacBio's SMRTlink software using barcode set 'RSII_96_barcodes' (available at https://github.com/AntWarland/doctoral_thesis/blob/main/PacBio/RSII_96_barcodes.fasta). Parameters used were chosen to capture the maximum number of CCS reads: maximum CCS read length: 50,000, minimum CCS read length: 10, minimum number of passes: 0, minimum predicted accuracy: 0, process all reads: true, minimum CCS predicted accuracy: 0, minimum barcode score: 0, same barcodes on both ends of sequence: true, write unbarcoded reads: true, minimum barcode quality: 26. Polished sequencing data was output to FASTA and FASTQ files (1 file per sample).

2.6.2 MiSeq

Libraries were sequenced by Branduff McAllister on a MiSeq at Glasgow Polyomics (Glasgow University) using a 600-cycle MiSeq v3 reagent kit (Illumina, MS-102-3003), running with 400 bp forward and 200 bp reverse sequencing. The sequencing parameters used were as previously described (Ciosi et al. 2019). MiSeq Reporter software was then used to demultiplex the reads using default parameters, which outputs the sequencing reads in FASTQ files.

2.6.3 Sanger sequencing

Library 3000-LBC-PBMC was Sanger sequenced using the primers in Table 2.9 to confirm the correct *HTT* locus had been amplified. Forward and reverse reactions were prepared in TE buffer with 10 µL of DNA and 4 µL of primer at 10 $ng.\mu L^{-1}$ and 1 µM respectively. Samples were sent to LGC GmbH, Berlin, Germany, where they were sequenced on a 3730xl DNA Analyzer (Applied Biosystems).

Primer	Direction	Sequence
TOM54	FW	5'-CTGACACAGTGGACAAAGGC-3'
TOM55	REV	5'-AAACAAGTTCTCGCCCCAAC-3'

Table 2.9. Sanger sequencing primers. FW: forward primer; REV: reverse primer. 5': 5-prime phosphate.

2.7 *HTT* CAG counting and flanking sequence determination

2.7.1 RepeatDecoder

FASTA files were tidied using UNIX shell BASH (https://www.gnu.org/software/bash/) commands invoking the utilities Sed and AWK (commands used viewable at https://github.com/AntWarland/doctoral_thesis/blob/main/RD/RD_bash_commands.txt). Briefly, line breaks within reads were removed and read IDs were trimmed leaving just the read ID preceded by a '>'. Tidied FASTA files were then analysed by RepeatDecoder v1.0.15 (RD) (Vincent Dion, Thierry Scheupbach, unpublished) using restrictive and permissive profiles configured to count CAG repeats (available at https://github.com/AntWarland/doctoral_thesis/tree/main/RD/profiles). Arguments used were: --with-revcomp, -t 8, --optimal, -o TSV, --source. Restrictive and permissive counting metrics for each read were output to '.txt' files (1 file per profile per sample, full commands viewable at https://github.com/AntWarland/doctoral_thesis/blob/main/RD/RD_bash_commands.txt). Columns 1, 7 and 8 were trimmed from the resulting files before being sorted into 'permissive' and 'restrictive' directories. Flanking sequences were determined in the downstream analysis (see 2.8.1).

2.7.2 ScaleHD

3 kbp PacBio reads were too long for ScaleHD so had to be trimmed and formatted using cutadapt (Martin 2011) before being analysed. Except for polishing and demultiplexing (see 2.6.1), and trimming, PacBio reads were analysed in the same way as MiSeq data, as described elsewhere (Mcallister 2019). Bioinformatic processing of MiSeq data used the Scale-HD pipeline (v0.322) written by Alastair Maxwell. ScaleHD's installation and usage are described in detail in its documentation (https://scalehd.readthedocs.io/en/latest/). Briefly, ScaleHD works by aligning reads to a list of 4000 canonical *HTT* reference sequences with varying uninterrupted CAG lengths and CCG repeat structures using the BWA-MEM alignment algorithm (Li and Durbin 2009). Any atypical structures detected undergo further alignment to a separate database of non-canonical *HTT* structures (8000). This process is repeated for the second allele in the sample, and the alignments are used to create a BAM file for each

sample. Allele-specific attributes are then calculated, such as modal CAG and Somatic Expansion (see Figure 3.7).

Scripts and reference files used, as well as instructions for running ScaleHD on PacBio reads are available at https://github.com/AntWarland/doctoral_thesis/tree/main/ScaleHD

2.8 Analysis

2.8.1 PacBio analysis pipeline and data quality control

Data from RD and FASTQ files were imported into custom analysis pipeline written by me, with early contributions by Ellis Pires, using Jupyter notebooks (https://jupyter.org) running Python 3 (https://www.python.org) in Anaconda Navigator (https://docs.anaconda.com/anaconda/navigator/). Data manipulation was conducted using the following packages: Biopython, FuzzyWuzzy, Levenshtein, Regex, DocX; python libriaries: numpy, pandas, scipy.stats; and Biopython labriaries: Seq, SeqIO and AlignIO. Python plotting was conducted using libraries matplotlib and seaborn.

Briefly, FASTQ files are converted into SeqIO objects, and packed in a pandas dataframe, where they are associated with a barcode ID, with each read represented by a single row. These are then merged with CAG count data from RD text files before additional fields relating to CAG length and sequence are calculated for downstream processing, including the identification of the 3' flanking sequence using the positions of the 3' end of the polyCAG and polyglutamine from restrictive and permissive count data respectively. 3 filters, shown in Figure 2.1, are then applied to the data for quality control purposes. The number of reads present at each stage of filtering is shown for library 600-iPSC-4 in Table 4.6.

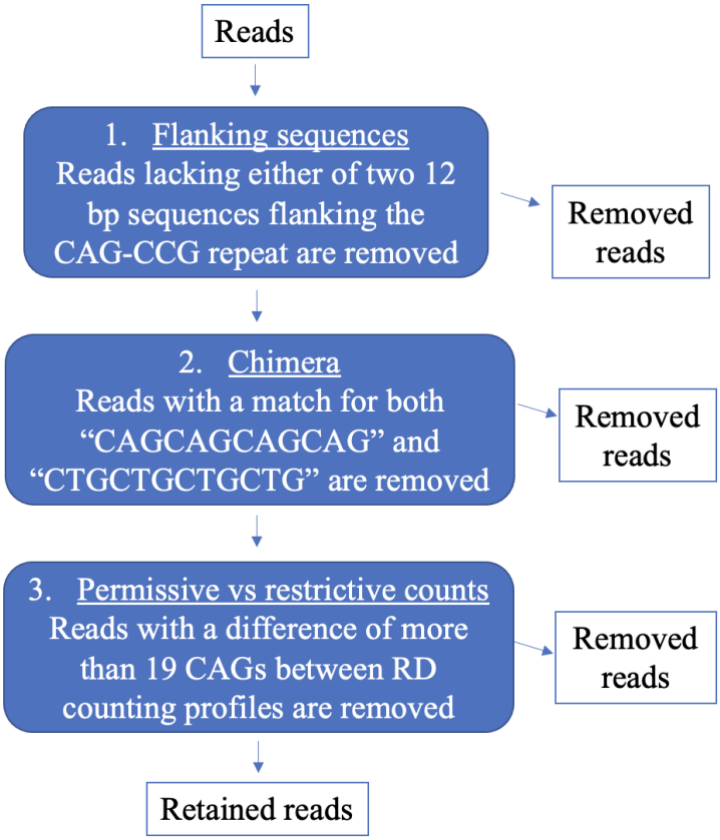


Figure 2.1. Filtering steps applied to PacBio-RepeatDecoder reads in my python analysis pipeline.

Retained reads are then annotated by merging the read dataframe with a sample information dataframe. The resulting annotated dataframe formed the basis of all subsequent PacBio-RD analysis in this project, including determining the modal CAGs for WT and expanded alleles, calculating expansion indices and sequence-based analyses.

2.8.2 Flanking sequence-CAG length association testing

Flanking sequence-CAG length association testing shown in Figure 4.20 was conducted using the stats package in R (version 3.6.0; R Core Team 2019, https://www.r-project.org). A dataframe containing the frequency of reads at each CAG length for each flanking sequence category was generated in Python and imported into R, where binomial regression was performed using the following model:

```
model <- glm(cbind(x,y)~n, family="binomial")
```

4 binomial models were generated where x was the number of non-canonical, loss, gain or other reads, y was the number of canonical reads and n was the change in CAGs from the modal CAG.

2.8.3 Other analysis

Normality and correlation testing was conducted using the scipy.stats package in python. Analysis of *FAN1* genotypes including the plotting of Figures 4.2, 4.7 and 4.14, and 2-way ANOVA was conducted in GraphPad Prism version 9.2.0 for MacOS, (GraphPad Software, San Diego, California USA, www.graphpad.com). See section 2.4 for details of capillary electrophoresis analysis.

Figures 3.1 and 4.1 were generated using Python package matplotlib. Read quality score and read length distribution plots were generated by PacBio's SMRTlink analysis. Other plots were generated using Microsoft Excel. Other figures were prepared using Microsoft Powerpoint.

2.8.4 Computing facilities

SMRTlink analysis used the Raven supercomputing cluster of the Advanced Research Computing Division (ARCCA) at Cardiff University and the Hawk supercomputer housed at Cardiff University, part of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) *via* the Welsh government. Long-term storage of *HTT* sequencing data is in a dedicated storage cluster on Hawk.

2.8.5 Data and code accessibility

PacBio sequencing data in the form of CCS reads in FASTQ and FASTA files will be made available at https://github.com/AntWarland/doctoral_thesis/tree/main/PacBio/sequencing_data. RD count data will be available at https://github.com/AntWarland/doctoral_thesis/tree/main/RD/count_data. Python analysis pipeline code will be made available at https://github.com/AntWarland/doctoral_thesis/tree/main/Analysis.

2.9 Small-pool PCR

2.9.1 PCR

To find the appropriate concentration for small pool PCR (SP-PCR), an initial membrane was run with a 10-fold dilution series of gDNA. DNA was diluted to 10 $ng.ul^{-1}$, 1 $ng.ul^{-1}$, 100 $pg.ul^{-1}$ and 10 $pg.ul^{-1}$. Table 2.10 shows the PCR recipe used. Primers oVIN1333 (forward) (5'-CCGCTCAGGTTCTGCTTTTA-3') and oVIN1334

(reverse) (5'-CAGGCTGCAGGGTTACCG-3') were used for all small pool PCR (supplied by Merck Life Science UK Ltd., Dorset, UK).

Reagent	Volume (µL)
gDNA (10 ng/ µL)	1
GC Buffer 1 (2x)	5
dNTPs (2.5 mM each)	1.6
F primer (10 µM)	0.5
R primer (10 µM)	0.5
LA Taq (5 U / µL)	0.1
Nuclease-free water	Up to 10

Table 2.10. Recipe used for small pool PCR. F: forward (oVIN1333). R: reverse (oVIN1334). GC buffer, dNTPs and LA Taq supplied in kit (TaKaRa Bio Europe #RR02AG)

One membrane per genotype was run. Each membrane would include 7 replicates and one negative control at each of the 4 template DNA concentrations and one positive control (50 ng DNA). Therefore a master mix for 70 reactions was prepared minus the template DNA/water and kept on ice until needed. 9 µL of master mix was added to 1 µL of template DNA and mixed by vortex. PCR tubes were briefly spun down before being loaded in a T100 Thermal Cycler (Biorad). Table 2.11 shows the heating programme used.

Cycles	Temperature / ˚C	Time / s
1	94	90
4	94	20
	52	20
	72	150
24	94	30
	55	30
	72	150
1	72	600
1	20	Forever

Table 2.11. Small pool PCR programme.

2.9.2 Agarose gel electrophoresis

10 µL of positive control products was run on a 30 ml 1.5% agarose gel made with 1 X Tris-Acetate-EDTA (TAE) buffer and SYBR safe DNA stain to check PCR worked

and products run to expected place. 10 µL each of the remaining products were run alongside DNA ladder on a 300 ml, 36-well 1.5% agarose gel made with 1x TAE and SYBR safe DNA stain. Gels were run in 1X TAE at 140 V for 30 minutes, followed by 50 V for 14.5 hours. Gels were imaged by UV transilluminator with and without fluorescent rulers for later alignment.

2.9.3 Southern blot

Gels were bathed in fresh alkaline transfer buffer (0.4M NaOH, 1M NaCl) for 20 minutes twice before being transferred to cellulose membranes in a tank containing alkaline transfer buffer overnight. Membranes were washed in neuralisation buffer (1.5M NaCl, 0.5M Tris base, pH 7.4) for 5 minutes before being transferred to pre-heated (48 ºC) cylinders containing 15 ml of UltraHyb hybridization buffer (Thermo Fisher Scientific #AM8670) to pre-hybridize at 48 ºC for 1 hour. A CAG repeat probe was prepared according to the recipe show in Table 2.12.

Reagent	Volume (µL)
Buffer T4 PNK	2.5
oVIN-100 (1uM)	5
Nuclease-free water	11.5
(In perspex cabinet/radiation area)	
α-P32-dATP	5
T4 PNK	1

Table 2.12. CAG repeat probe recipe. Buffer T4 PNK and T4 PNK supplied by New England Biolabs UK. oVIN-100 sequence 5'-AGCAGCAGCAGCAGCAGCAGCAGCAGC-3', supplied by Merck Life Science UK Ltd. α-P32-dATP: ATP[γ-32P]-3000 Ci/mmol (Perkin Elmer #NEG002 A250UC).

The reaction was mixed by vortex before being spun down in a microcentrifuge. It was then transferred to a heat block and heated at 37 ºC for 30 minutes, followed by 65 ºC for 20 minutes. Once ready, 8 µL of probe was added directly to cylinder containing the prehybridized membrane and hybridized at 48 ºC for 2 hours. Wash buffer (0.5X SSC, 0.1% SDS) was preheated to 48 ºC. Membranes were washed twice in 15 mls of pre-heated wash buffer for 30 minutes before transferring to a phosphoscreen for overnight exposure. The exposed screen was imaged using a PharosFX Plus Molecular Imager (Bio-Rad). Images were aligned and annotated with images of the agarose gels with rulers using Adobe Photoshop 7.0.

Chapter 3 : Long-read sequencing the *HTT* CAG repeat

3.1. Introduction

HD is one of a growing number of diseases caused by the expansion of genomic short tandem repeats (Chintalaphani et al. 2021). The inheritance of a single expanded *HTT* CAG repeat longer than 39 CAGs is completely penetrant for HD, however many repeat diseases have pathogenic thresholds of hundreds or thousands of repeats (Khristich and Mirkin 2020). Furthermore, there is growing evidence that variation within the DNA sequence in the repeat itself is a determining factor in the timing of disease onset in HD (Ciosi et al. 2019; Lee et al. 2019; Wright et al. 2019) and other repeat diseases (Choudhry et al. 2001; Gao et al. 2008). These findings highlight the pressing need for investigation of repeat expansion diseases with methods which can both size and sequence repetitive DNA of pathogenic length.

Ciosi et al. show that a short-read NGS method adapted for longer read lengths (Illumina MiSeq) is capable of spanning *HTT* repeats with up to 120 CAGs and predict that alleles up to 90 CAGs could be reliably sized and sequenced (Ciosi et al. 2021). However, for repeats which are longer than this, such as those found in some juvenile cases of HD, Fragile-X syndrome and myotonic dystrophy (Khristich and Mirkin 2020), long-read NGS is necessary. To date, two truly long-read sequencing technologies have been developed: that of PacBio and Oxford Nanopore Technologies (ONT). Both technologies have been applied to sequencing pathogenic repeat loci in numerous studies and several reviews published recently explore their use to date in human disease (Mantere et al. 2019) and specifically in neurological disease (Chintalaphani et al. 2021; Su et al. 2021).

PacBio SMRT sequencing has been applied to bulk-PCR library preparations demonstrating *HTT* CAG repeats of up to 550 CAGs can be sequenced (Ciosi et al. 2021), however preferential amplification of shorter alleles in PCR is noted here and elsewhere (Massey et al. 2018). Hoijer et al. demonstrated that the *HTT* CAG repeat can be sequenced and studied without the need for PCR amplification using a CRISPR/Cas9 targeting approach (Höijer et al. 2018). Other disease-causing loci have been studied using PacBio long reads, including the Fuchs endothelial corneal

dystrophy-associated TCF4 repeat, in which repeats with more than 1500 CTG were sequenced (Hafford-Tear et al. 2019). A study by Giesselmann et al. shows that Oxford Nanopore Technologies sequencing can be used to assay multiple disease-causing repeats with over 400 repeat counts, including those in the *C9orf72* and the *FMR1* genes, using an amplification-free approach which allowed DNA methylation status to be determined (Giesselmann et al. 2019). Amplification-free target enrichment approaches have also been shown enable high-coverage (>100x) sequencing of multiple repeat loci simultaneously on both PacBio (Tsai et al. 2017) and ONT (Miller et al. 2020) platforms. These advances have the potential to dramatically improve the efficiency of repeat expansion disease diagnosis.

The fact that interruptions have been shown to modify onset in multiple diseases, including HD, mean sequencing of alleles will be needed to predict uninterrupted CAG length accurately. For any allele-specific silencing therapeutic approaches, long-read sequencing to establish phase of the repeat with any SNPs targeted by such therapies will be essential (Svrzikapa et al. 2020).

Long-read NGS is well-placed to address pressing research questions related to repeat expansion diseases, such as the effect of repeat interruptions on expansion and onset, the determination of the pathogenic threshold of disease-causing repeat expansions, the role of epigenetic modifications in repeat expansion and the possible therapeutic utility of allele-specific silencing of expanded repeats. The ability to sequence through very long repeats will enable advances in our understanding of the genetic causes neurodegenerative diseases. To date these sequencing platforms have not been extensively validated. Efforts made here aim to contribute to our understanding of the capabilities and limitations of these emergent techniques.

3.2. Chapter aims

In this chapter I aim to answer the following research questions:

1. Can PacBio sequencing data be used to reliably quantify expanded *HTT* CAG repeats in HD patient DNA?
2. How do the *HTT* CAG repeat counts of lymphoblastoid and peripheral blood mononuclear cells compare?
3. Can PacBio sequencing data be used to reliably quantify the expanded *HTT* CAG repeats in 109NI iPSC model DNA?

3.3. Results

3.3.1. Developing a Method for Long-Read Sequencing of the *HTT* CAG repeat

3.3.1.1. 3 kbp Amplicon Library Preparation

To generate long read sequence data using the PacBio Sequel System, I designed an amplicon that spanned the *HTT* CAG repeat. Three pairs of primers were designed to amplify a 3 kbp genomic region containing the CAG tract in exon 1 of human *HTT* based on genome assembly hg38 (Figure 3.4A, Table 2.3). An amplicon length of 3 kbp was chosen as I was initially interested in whether long-read sequencing could be used to phase the alleles using SNPs in linkage disequilibrium with the CAG repeat. Ultimately, I decided not to pursue phasing due to the lack of polymorphic SNPs in the 3 kb region in our samples, which meant it was not possible to generate a unique haplotype for each patient. Primer pairs were designed to have unique PCR products *in silico* and included 5' flaps to allow PCR to occur with PacBio's barcoded universal primers in a second round of amplification (Figure 3.4B). PCR products between 2.5-3 kbp were consistent with *in silico* PCR products for all primer pairs (Figure 3.1, Table 2.4). An experiment to determine optimal primers and annealing temperature (Figure 3.1) found TOM48/49 and annealing at 61.7°C produced the strongest bands in the expected size range relative to the amount of non-specific DNA products (outside 2.5-3kbp). Most of the non-specific PCR products visible in this condition – which could be aborted PCR products or off-target amplification – were below 0.7 kbp, which will be removed during purification.

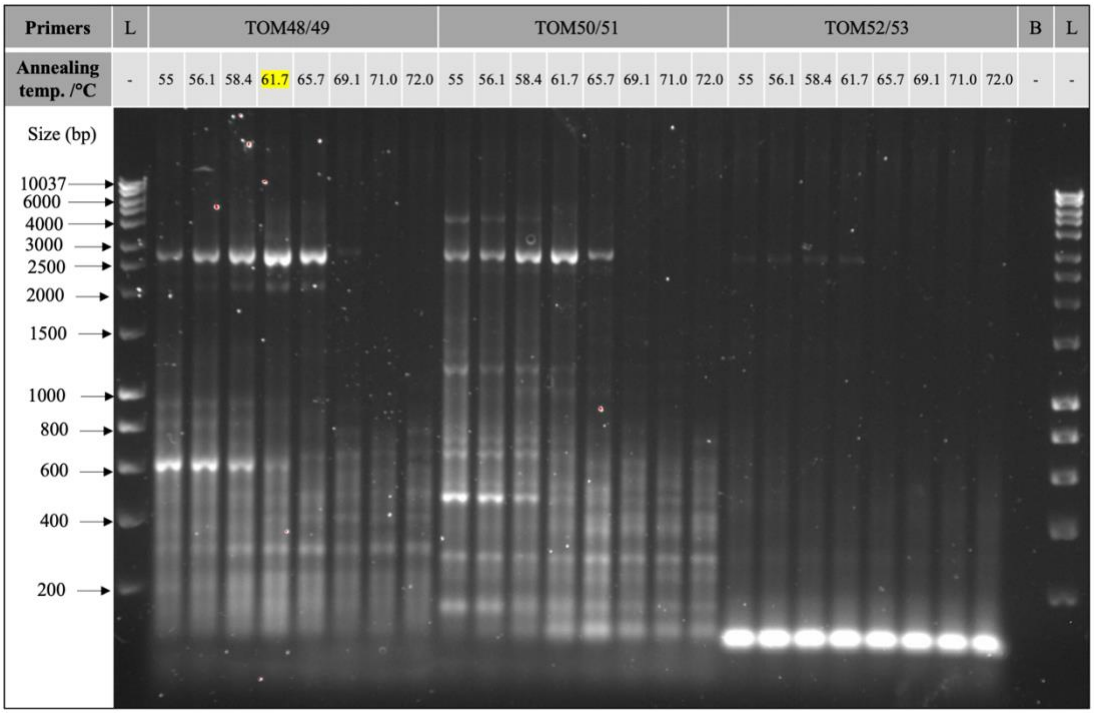


Figure 3.1. Gel electropherogram showing 3 kbp PCR amplicons of the *HTT* locus. Three primer pairs were tested on human gDNA at a range of annealing temperatures. In silico product sizes for wild-type and expanded allele respectively: TOM48/49: 2904, 2967; TOM50/51: 2984, 3,047; TOM52/53: 2840, 2903. The chosen condition, TOM48/49 at an annealing temperature of 61.7□C, is highlighted in yellow. L indicates 1 kbp ladder (Bioline). B indicates a blank lane.

A further experiment, shown in Figure 3.2, found that 30 cycles and a template DNA input mass of 25 ng were optimal for PCR amplification by TOM48/49. Amplification of the intended locus was verified by Sanger sequencing using primers TOM54 and TOM55, which contain the annealing portions of primers TOM48 and TOM49 respectively (see 2.6.3). Alignment of Sanger sequencing was conducted using EMBOSS Water (Madeira et al. 2019). Appendix 2 shows the complete output from the alignment. Sanger sequencing has a typical max read length of 900 bp (Table 1.2). The forward FASTA sequence was 1,054 bp long and shared 99.4% identity with the *in silico* product. The reverse FASTA sequence was 508 bp long and shared 99.8% identity with the *in silico* product. While the forward sequence terminated before entering the CAG repeat (but beyond the typical max read length), the reverse sequence terminated 56 bases after entering the CTG repeat, and well short of the typical max read length, suggesting it was a problematic template.

Figure 3.2. Gel electropherogram showing further optimisation of PCR of the *HTT* locus. Amplification of human gDNA was tested with two primer pairs on several template DNA masses and PCR cycle combinations. *In silico* product sizes for wild-type and expanded allele respectively: TOM48/49: 2904, 2967; TOM50/51: 2984, 3,047. The chosen conditions are highlighted in yellow. L indicates 1 kbp ladder. B indicates a blank lane.

5'-blocked (amino-C6) versions of TOM48/49 performed identically to the unblocked equivalents. Blocked primers prevent amplicons from the first round of PCR being used to form SMRT bell templates. This in turn eliminates the possibility of amplicons without barcodes being sequenced, resulting in a higher yield of usable sequence data.

The templates for initial library, 3000-LBC-PBMC (Table 3.3), consisted of 48 samples of HD patient DNA extracted from peripheral blood mononuclear cells (PBMCs) and lymphoblastoid cell lines (LBCs) by Branduff McAllister. I wanted to establish the accuracy and depth of long-read sequencing of 3kb amplicons of patient-length HD repeats. I chose to run 48 samples initially to obtain approximately 10,000 reads per sample based on an estimate of 500,000 reads per chip. These samples had the advantage of being previously sequenced on Illumina's MiSeq and quantified by fragment analysis, which would form the basis of my validation of PacBio sequencing as a method to quantify the *HTT* CAG repeat.

An overview of library preparation can be seen in Figure 3.4B. Briefly, optimal PCR conditions were used in conjunction with blocked primers to amplify DNA extracted from 48 PBMC and LBC pellets from HD patients (section 2.5.1.3.1). PCR products were verified by gel electrophoresis before purification with paramagnetic beads (section 2.5.1.5) and fluoroscopic quantification (see 2.3.2). Barcoded universal primers were used in a second round of PCR before a further round of bead purification and quantification. Samples were pooled to equimolar concentrations and 750 ng of the pooled library converted to SMRTbell template (Figure 3.4B). The final library was checked by capillary electrophoresis (Figure 3.3, section 2.5.1.6). Based on the *in silico* PCR product sizes of 2,904 and 2,967 bp for the WT and expanded allele respectively, with an additional 32 bp for barcodes and 88 bp for adapters, I was expecting the library generated using TOM48/49 primers to contain DNA with 3,036 and 3,099 bp respectively. The trace peak at 3,077 bp (Figure 3.3) was in the expected size range and has a small shoulder extending to 5000 bp at its base. No other non-specific DNA appears in the trace. After a final DNA quantification, the sequencing primer and enzyme were added to the library before loading on a SMRT cell and sequencing according to the technical guidance of PacBio.

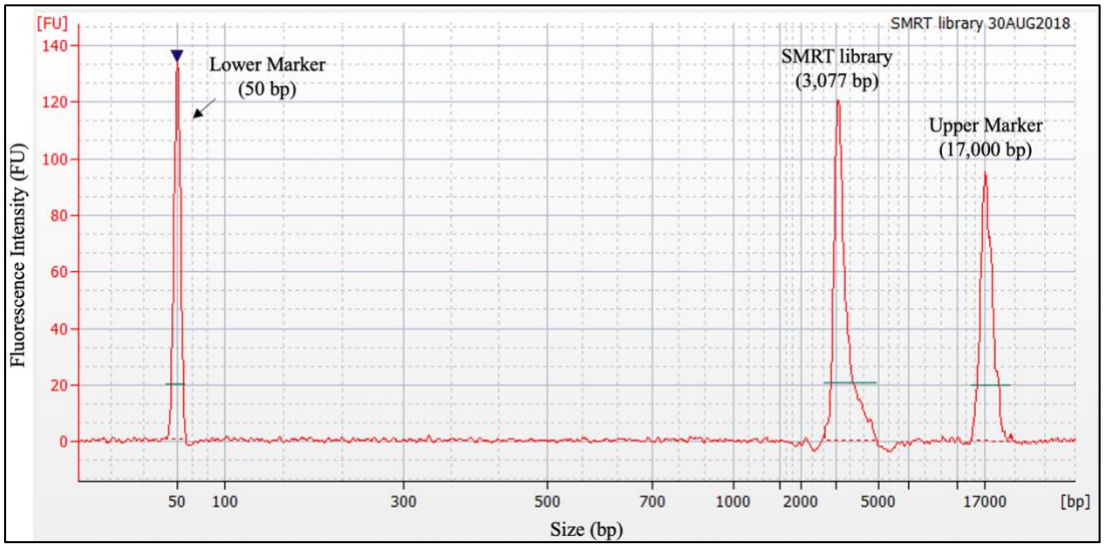


Figure 3.3. Capillary electrophoresis trace of the pooled SMRTbell library. The peak at 3,077 corresponds with gel electrophoresis of the PCR amplicons that were used to make the library. No non-specific DNA products are visible. Agilent Bioanalyzer 2100, DNA 12000 kit.

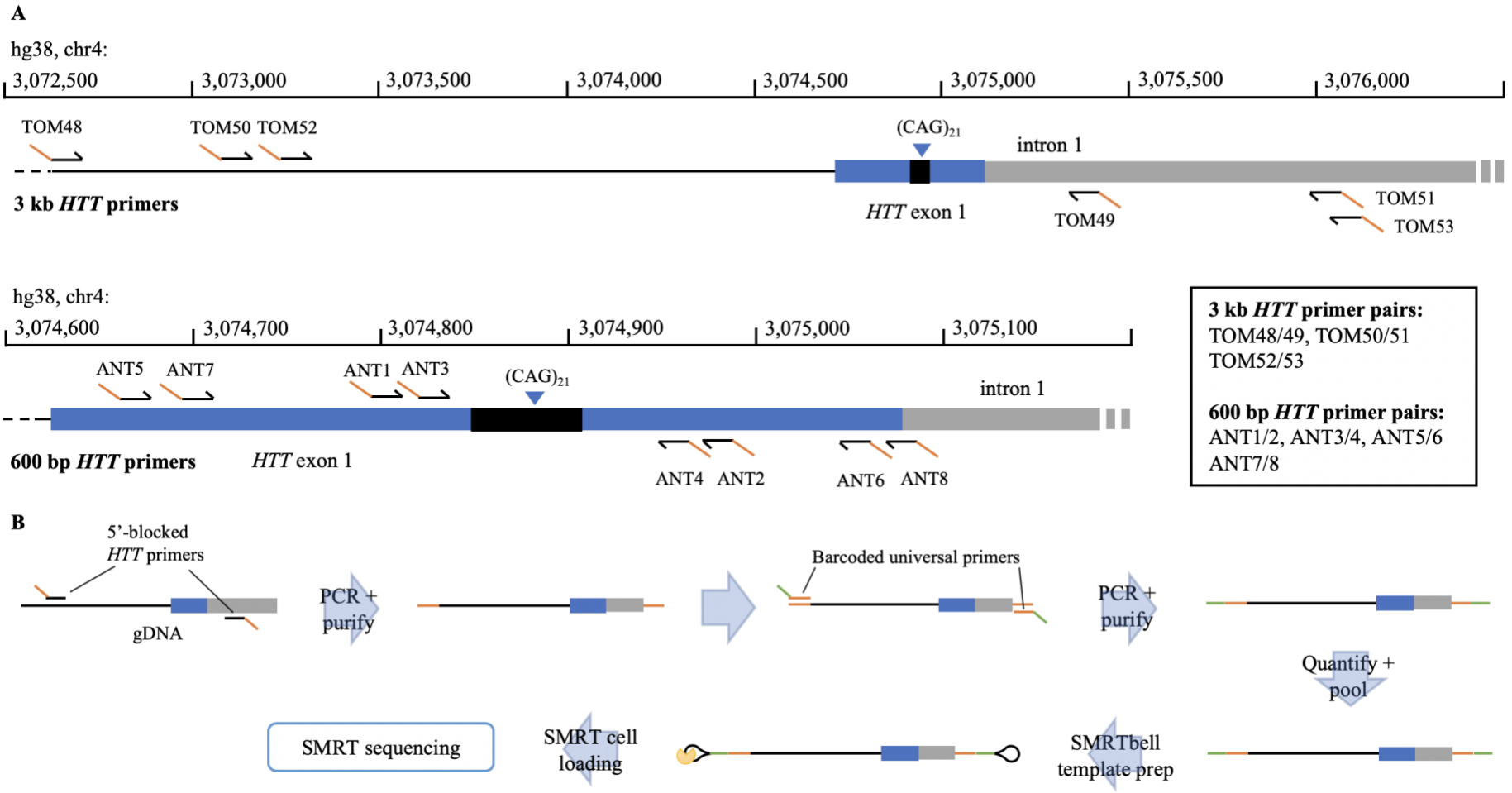


Figure 3.4. Primer design and sequencing library preparation method. (A) Multiple pairs of primers were designed to amplify the CAG repeat on exon 1 of *HTT*. (B) Schematic showing an overview of the sequencing library preparation method used.

3.3.1.2. Sequencing Data, Quality Control and Filtering

Library 3000-LBC-PBMC, the preparation of which is described in section 3.3.1.1, produced 7.6 Gb of raw data from a single SMRT Cell 1M (typical yields 3.65-5.11 Gb per cell. See methods 1.6.2. for more info), comprising 3 million subreads (see section 1.6.2 and Figure 1.4 for an explanation of subreads). Library loading was slightly above target with 65% of the 1 million ZMWs occupied by a single template-polymerase complex. The higher this proportion the higher the throughput, with PacBio's technical documentation recommending a target of >50% for most applications (https://github.com/AntWarland/doctoral_thesis/blob/main/PacBio/Loading-Recommendations.pdf accessed 11/04/2022). Demultiplexing identified 48 out of 48 barcodes and 88% of the raw data was associated with a barcode. CCS analysis, which polishes subreads by building a consensus from each polymerase read (Figure 1.4A), was conducted using default parameters. The total number of CCS reads was 146,197, giving a mean number of reads per sample of 3,046.

Library name	Samples	Amplicon size (primers)	Size selection (Ampure x)	CCS reads
3000-LBC-PBMC	48 human LBC/PBMC	3 kbp (TOM48/49)	No (1.0x)	146,197
3000-LBC-A	47 human LBC	3 kbp (TOM48/49)	No (1.0x)	164,177
3000-LBC-B	47 human LBC	3 kbp (TOM48/49)	No (1.0x)	167,052
3000-LBC-PBMC-iPSC	28 human LBC/PBMC, 12 iPSC	3 kbp (TOM48/49)	No (1.0x)	183,991
3000-iPSC	6 109NI iPSC	3 kbp (TOM48/49)	No (1.0x)	165,778
600-iPSC-1*	6 109NI iPSC	600 bp (ANT1/2)	Yes (0.6x)	160,198
600-iPSC-2	12 109NI iPSC	600 bp (ANT1/2)	Yes (0.6x)	139,065
600-iPSC-3	12 109NI iPSC	600 bp (ANT1/2)	Yes (0.6x)	145,181
600-iPSC-4	48 109NI iPSC	600 bp (ANT1/2)	Yes (0.6x)	361,347

Table 3.1. Details of all PacBio sequencing libraries run. *: Amplified from the same 6 samples as 3000-iPSC. LBC: lymphoblastoid cells. PBMC: peripheral blood mononuclear cells. iPSC: induced pluripotent stem cells.

The quality of CCS reads varies and can be quantified by Q score, or Phred score, a measure of the probability of correct basecalls. Q scores are calculated for all bases from log-likelihood values computed by a Hidden Markov Model algorithm during

CCS polishing. The average of these per-base qualities is taken as the predicted read accuracy (Q score). Q scores are logarithmically related to error probabilities so that with each increase of 10 Q, the probability of an incorrect basecall reduces 10-fold. See section 1.6.2 and Table 1.1 for more about Q-scores. A Q score of more than 20 (> 99% accuracy) has been used by PacBio to define High Fidelity (HiFi) reads (Wenger et al. 2020).

PacBio's website states that up to 500,000 HiFi reads can be obtained on a single SMRT Cell 1M, but that the number depends on sample quality, sample type and amplicon size (PacBio's website: https://www.pacb.com/products-and-services/sequel-system/previous-system-releases/ accessed 07 Dec. 2021). It should also be noted that manufacturer websites often quote performance values obtained with in-house optimized conditions. In the 3000-LBC-PBMC library run, 103,386 CCS reads (71%) were HiFi reads, the median quality of which was Q26. PacBio's library preparation is known to generate chimeric products and the repetitive and GC-rich DNA in my libraries may increase the number of these.

Figure 3.5A shows the distribution of the read scores, with the vast majority of reads between Q10 and Q40 and a secondary peak at Q60, the maximum score assigned for a read. Only HiFi reads (> Q20) were taken forward for analysis.

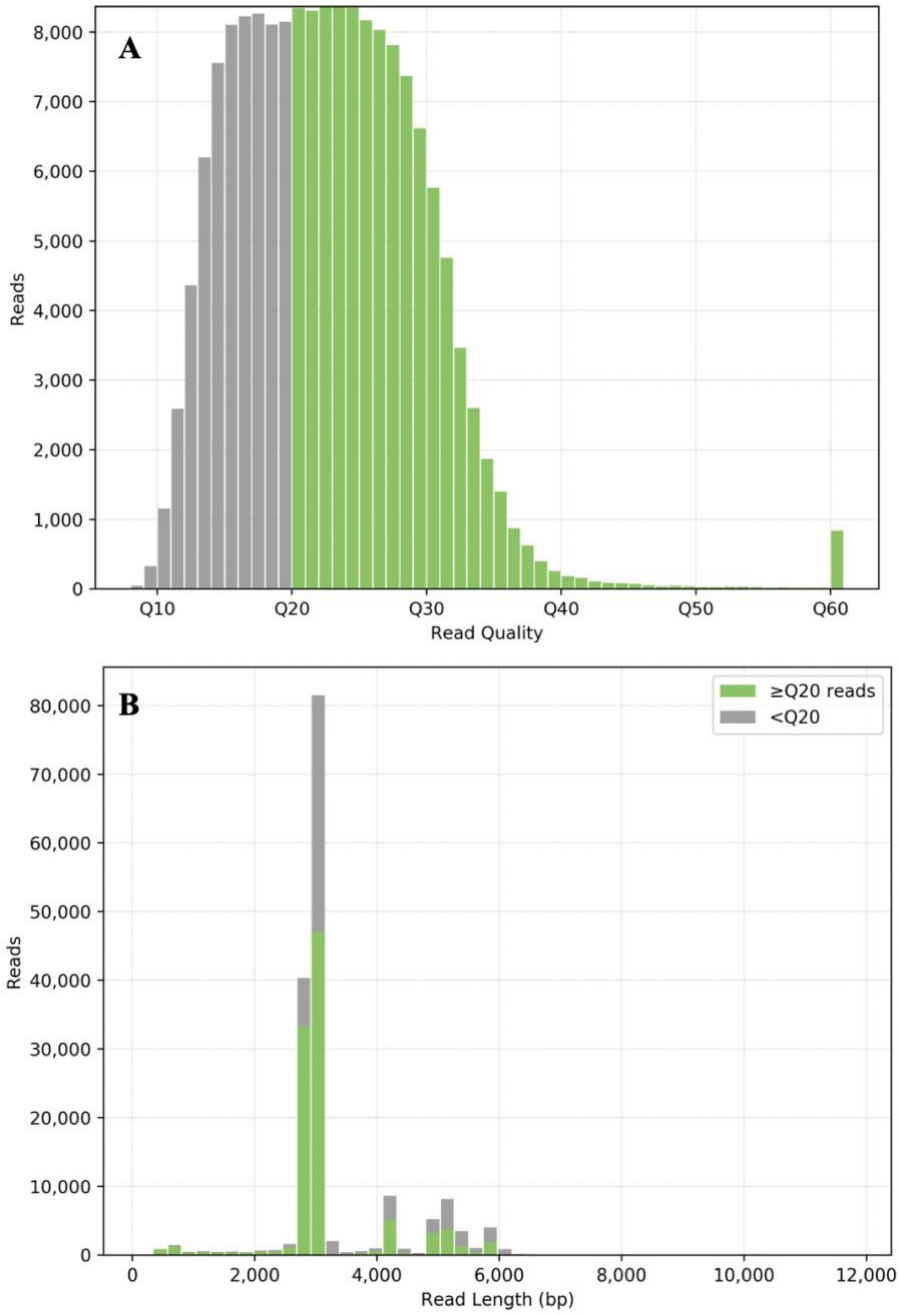


Figure 3.5. Read quality score and read length distributions for 3 kbp *HTT* amplicons sequenced on PacBio. Library 3000-LBC-PBMC. (A) Histogram showing the distribution of CCS read qualities (B) CCS read length distribution. Bp: base pairs. HiFi reads (Q20 or higher) shown in green, non-HiFi reads in grey.

Mean read length was 3,301 bp and the mean number of passes (subreads) per CCS read was 9. While there is no standard number of passes, a higher number is associated with higher accuracy, with 4 passes achieving a median read quality of Q20 (Wenger et al. 2019). Figure 3.5B shows the majority of reads in a peak at approximately 3 kbp, which is consistent with the Bioanalyzer trace (Figure 3.3). There are also several smaller peaks between 4-6 kbp, which may be represented by the shoulder seen in the Bioanalyzer trace (Figure 3.3). These will be filtered out in the analysis pipeline.

To verify that the expected region of *HTT* had been amplified, a random selection of 5 samples were mapped to human genome reference sequence hg38 using minimap2 (Li 2018). Appendix 3 shows an illustration of the alignments visualised using Integrated Genome Viewer (IGV) (Robinson et al. 2011). Reads mapped within the expected 3 kbp region at the *HTT* exon 1 CAG repeat locus.

A custom data analysis pipeline was written in Python to assess CAG repeat lengths and structures in *HTT* (see 2.8). FASTQ files containing HiFi reads were downloaded from SMRTlink and imported into the analysis pipeline. Reads were then filtered (summarised in Figure 2.1). Briefly, for inclusion in downstream analyses, reads had to possess both of two 12 bp sequences immediately flanking the CAG-CCG repeat. Those reads lacking one flanking sequence were filtered out. A visual inspection of FASTQ files showed many reads with a non-typical or chimeric structure. An example of such read had both a CAG repeat tract and a CTG repeat tract, indicating that the target amplicon and its reverse complement were given in the final sequence. PacBio acknowledge that recombinant molecules and chimeras are as common as 20-30% in their data and may be the result of PCR amplification (Oh et al. 2016). While it was not possible to remove all PCR artefacts from this analysis, I was able to remove most of the chimeric reads, defined here as a read with a sequence containing an exact match for both "CAGCAGCAGCAG" and "CTGCTGCTGCTG" strings. All reads satisfying this criterion were filtered out in my analysis pipeline.

Initially, reads were also filtered out if they fell outside the read length range of 2.5-4 kbp, however this filter was dropped when analysing longer repeats due to the need to preserve extremely large repeat expansions and was replaced with a filter that removed reads with unusual CAG repeat structures, e.g., duplications and recombinations based on the number of CAGs from lenient and stringent counting profiles (see section 3.3.1.3.3).

While most 'short' reads (< 7 CAGs) were removed by the flanking sequence filter, 1,782 'short' reads remained when applying the second iteration of filtering. 93,069 reads were classified as 'WT' or 'expanded'. 'WT' represents the wild type allele and is defined here as a read with 7-29 CAGs, while 'expanded' represents the expanded allele and is defined here as a read with >29 CAGs. This cut off was chosen at it captured almost all WT alleles in patient data, while minimising the number of reads

forming expanded allele peaks being falsely categorised. A breakdown of the number of reads surviving filtering is shown in Table 3.2.

Library name	Short reads	WT allele reads	Expanded allele reads	Total
3000-LBC-PBMC	1,782	43,917	50,152	95,851
3000-LBC-A	2,411	54,546	41,770	98,727
3000-LBC-B	2,593	56,028	41,831	100,452
3000-LBC-PBMC-iPSC	2,303	78,081	30,116	110,500
3000-iPSC	2,665	91,340	13,026	107,031
600-iPSC-1*	44	30,849	108,889	139,782
600-iPSC-2	51	23,128	74,744	97,923
600-iPSC-3	58	29,752	97,900	127,710
600-iPSC-4 ^	93	113,200	189,220	302,513

Table 3.2. Reads surviving filtering by allele from all sequencing libraries. Short: < 7 CAGs. WT: wild type, a read with 7-29 CAGs. Expanded: a read with > 29 CAGs. CAGs counted by RD (restrictive profile). * Generated from the same 6 samples as 3000-iPSC. Libraries were sequenced on one SMRTcell 1M unless stated otherwise. ^ Sequenced on two SMRTcell. LBC: lymphoblastoid cells. PBMC: peripheral blood mononuclear cells. iPSC: induced pluripotent stem cells.

Table 3.3 shows a description of the samples sequenced in each library. Codes used are comprised of a patient or cell line identifier. Patient E16, for example appears twice in the list, once for the LBC sample coloured blue and once for the PBMC sample coloured red. 48 samples were sequenced in the first chip, followed by 47 in the next two chips, 3000-LBC-A and 3000-LBC-B, as 2 samples failed to generate sufficient DNA during library preparation. 8 samples failed to generate sufficient DNA in the preparation of library 3000-LBC-PBMC-iPSC. Higher read depth was required for subsequent libraries, hence fewer samples were run.

	Library name							
Sample number	**3000-LBC-PBMC**	**3000-LBC-A**	**3000-LBC-B**	**3000-LBC-PBMC-iPSC**	**3000-iPSC**	**600-iPSC-1**	**600-iPSC-2**	**600-iPSC-3**
1	E16	E01	E291	E13	11B11-P36	11B11-P36	L81	5F-P4-r1
2	E61	E02	E293	E15	5F-P6	5F-P6	L81	5F-P4-r2
3	E70	E03	E298	E24	5F-P33	5F-P33	E11	5F-P20-r1
4	E77	E04	L23	E28	109NI-P31	109NI-P31	E11	5F-P20-r2
5	E84	E05	L33	E29	109NI-P46	109NI-P46	E79	5F-P36-r1
6	E85	E07	L48	E40	N15-P36	N15-P36	E79	5F-P36-r2
7	E103	E08	L59	E51			5F-P20 (c20)	11B11-P4-r1
8	E104	E11	L63	E118			11B11-P20 (c20)	11B11-P4-r2
9	E116	E22	L75	E128			5F-P20 (c24)	11B11-P20-r1
10	E119	E33	L76	L52			11B11-P20 (c24)	11B11-P20-r1
11	E126	E35	L93	L77			5F-P20 (c28)	11B11-P36-r1
12	L54	E37	L106	L118			11B11-P20 (c28)	11B11-P36-r2
13	L56	E57	L115	E6				
14	L65	E75	L123	E13				
15	L96	E79	L124	E16				
16	L103	E86	L132	E24				
17	L116	E87	L133	E51		Cell type:		
18	L125	E91	L135	E77		LBCs		
19	E12	E106	L136	E126		PBMCs		
20	E14	E107	L140	E128		iPSCs		
21	E15	E115	L202	E136				
22	E28	E139	L205	L22				
23	E29	E201	L206	L25				
24	E40	E205	L209	L56				
25	E61	E206	L210	L103				
26	E70	E209	L217	E223				
27	E71	E213	L219	E263				
28	E84	E214	L221	E277				
29	E85	E217	L222	11N11-P36				
30	E103	E219	L223	11B11-P36				
31	E104	E220	L225	5F-P6				
32	E116	E223	L226	5F-P33				
33	E118	E225	L227	109NI-P31				
34	E119	E226	L229	109NI-P46				
35	E144	E236	L232	3H2-P8				
36	L2	E244	L233	9E-P5				
37	L15	E245	L235	9E-P29				
38	L21	E247	L239	N15-P4				
39	L31	E251	L243	N15-P36				
40	L52	E263	L250	N15-P40				
41	L54	E265	L251					
42	L65	E269	L252					
43	L77	E270	L264					
44	L96	E273	L265					
45	L116	E282	L267					
46	L118	E288	L270					
47	L125	E289	L271					
48	E37							

Table 3.3. List of samples sequenced in the first 8 PacBio sequencing libraries. E/L number: patient code. All iPCSs are either derived from the 109NI line or are the 109NI line. Family tree shown in Figure 3.13. The first part of iPSC name refers to the cell line. 109NI: parent line with an expanded *HTT* repeat of 109 CAGs and $HTT^{+/+}$. 11B11/11N11: isogenic subclone of 109NI. 5F: $HTT^{-/-}$ knockout of a 109NI subclone. 3H2: double wild type *HTT* repeat control. 9E: double wild type *HTT* repeat, $FAN1^{-/-}$ knockout. NI5: 109NI sublclone 5. P## refers to the number of passages cells were cultured for. (c##) numbers refer to the number of first-round PCR cycles. r# is the PCR replicate number.

3.3.1.3. Counting CAG tract length in the 3 kbp human *HTT* sequencing data

3.3.1.3.1. ScaleHD

To assess the accuracy of PacBio sequencing of *HTT* CAG repeats, I compared PacBio data from the 48-sample library ('3000-LBC-PBMC') with *HTT* CAG repeat sequencing data generated on the same samples using MiSeq ultra-high-depth short-read sequencing, a well validated method for determining *HTT* CAG repeat sequences and their lengths (Ciosi et al. 2019; Ciosi et al. 2021; McAllister et al. 2022). To generate the MiSeq data, Branduff McAllister employed targeted *HTT* CAG repeat sequencing of 500 patients at the extremes of residual age at motor onset, i.e., the age at motor onset predicted by uninterrupted CAG repeat length alone minus the actual age at motor onset. Early being defined as having a negative residual age at motor onset, and late defined as having a positive residual age at motor onset. In the study, the top 4% at each extreme were selected for whole exome sequencing and targeted *HTT* CAG repeat sequencing using MiSeq. Scale-HD (https://scalehd.readthedocs.io/en/latest/index.html) was used to call repeat sequences and lengths in the MiSeq sequencing data (Mcallister 2019), hence I used it to perform a baseline comparison of the PacBio and MiSeq data.

Briefly, ScaleHD works by aligning reads to a list of 4000 'canonical' *HTT* reference sequences with varying CAG and CCG repeat structures using the BWA-MEM alignment algorithm (Mcallister 2019). The reference sequence with the highest alignment score is given as the repeat structure for each read before reads are grouped into alleles and allele-specific attributes are calculated, such as modal CAG and Somatic Expansion.

Figure 3.6 shows an example of the CAG length frequency distribution of PacBio and MiSeq data of one HD patient sample counted by ScaleHD. Two peaks are visible in the data comprising the WT and expanded alleles respectively.

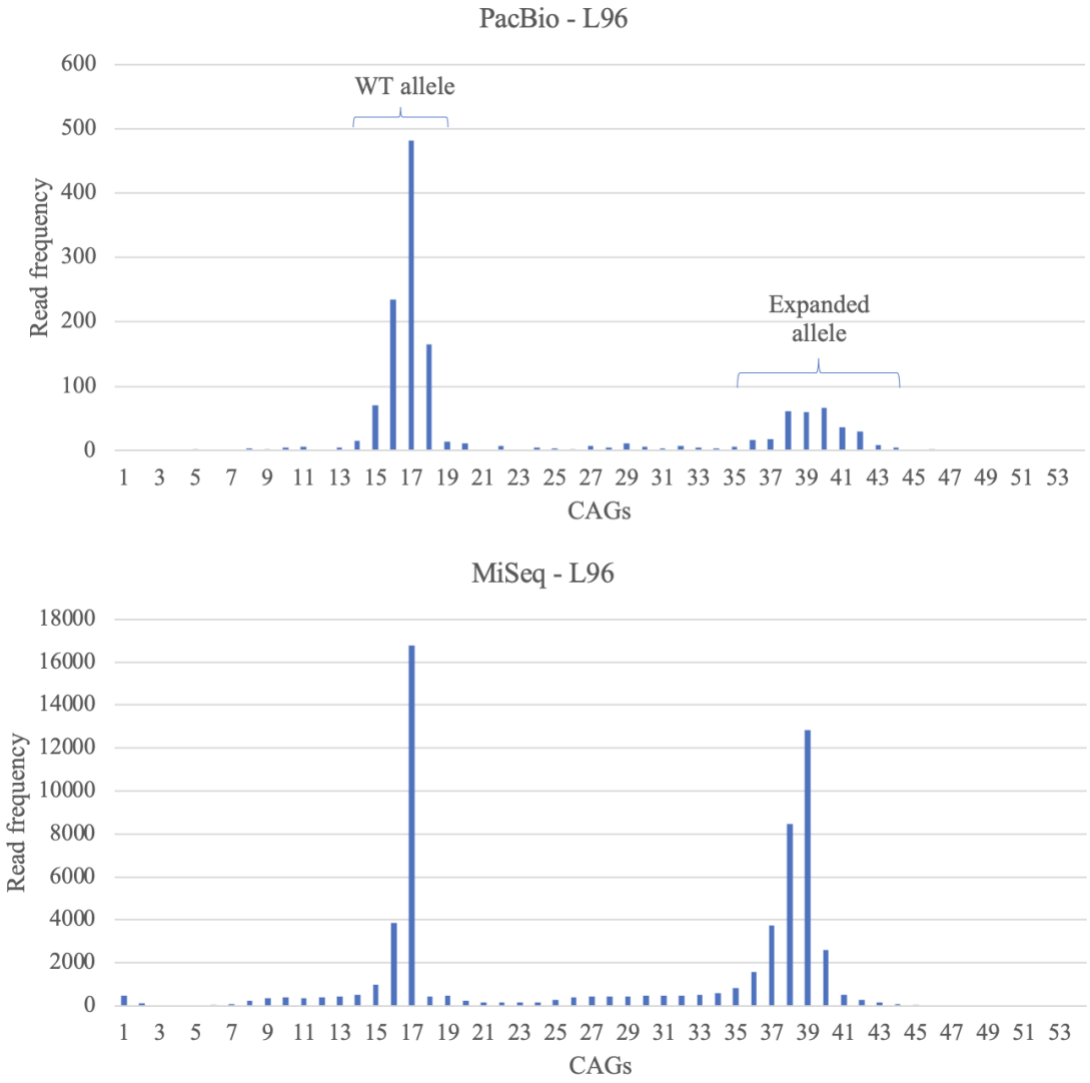


Figure 3.6. Example of CAG length frequency distributions of PacBio and MiSeq sequencing data of one HD patient PBMC sample. CAGs counted by ScaleHD. Sample L96 from library 3000-LBC-PBMC. WT: wild type. PBMC: peripheral blood mononuclear cell.

While the WT allele of the PacBio data shown in Figure 3.6 displays more variability than the equivalent MiSeq data, the modal WT peak at 17 CAGs is the same, which gives me confidence in this being 'correct'. There are other peaks around the WT peak, and they vary. Some of this could be due to mosaicism, however it is likely that most are PCR (or other) artefacts. It is not possible to distinguish the two in the current data, but we know from prior work (Murray et al. 1993; Veitch et al. 2007; Daunay et al. 2019) that peaks to the left of the modal peak are likely to be PCR stutter. Peaks to the right are not observed in MiSeq data, suggesting either the extra rounds of PCR in the PacBio method compared to MiSeq, or something inherent in PacBio's technology, result in additional artefacts. Examination of PacBio FASTQ files shows that 119 reads have an exact sequence match for the 18-uninterrupted-CAG sequence "TTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG

CAGCAGCAGCAACAGCCG", and manual inspection of 10 of these reads showed they were all normal in terms of sequence, length, and quality. 165 reads were assigned 18 CAGs by ScaleHD, showing that most are 'correct'. Despite these unexplained artefacts, the modal WT peak is clearly 17 CAGs and therefore consistent with MiSeq.

While the PacBio modal expanded peak in the example in Figure 3.6 is more than MiSeq – 40 compared to 39 – there is less than 100 reads in the modal peak itself in the PacBio data, and the peaks are broadly centred over the same number of CAGs, suggesting higher read depth may bring them into agreement. Examination of PacBio FASTQ files shows that 62 reads have an exact sequence match for the 40-uninterrupted-CAG sequence "CTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAACAGCCG", and manual inspection of 10 of these reads showed they were all normal in terms of sequence, length, and quality. 65 reads were assigned 40 CAGs by ScaleHD, showing that most are 'correct'.

There is much more variability around the modal peak at the pathogenic repeat lengths. In measuring the peaks around the WT and expanded alleles peaks I decided to measure expansions, as this does not involve peaks to the left of the modal peak, which are likely to be mostly stutter artefacts. To limit the amount of noise in this measurement I applied a 'modal CAG +10' threshold or a 10% frequency threshold depending on the samples involved, as discussed below.

To assess the degree of somatic expansion in HD patient alleles, ScaleHD calculates Somatic Expansion using a modified version of the method for calculating instability and expansion indices devised by Lee et al. (Lee et al. 2010). The method is illustrated in Figure 3.7. I decided to use this method rather than the original method as it is the default output from ScaleHD and because this method produced indices that were influenced far less by neighbouring alleles than the original method, which uses a frequency threshold cut-off. This is not a problem in 109NI iPSC alleles, as there is a much larger CAG size difference between them. For that data I used expansion and instability indices as it is a more robust measure of somatic expansion (example calculation shown in Figure 3.16).

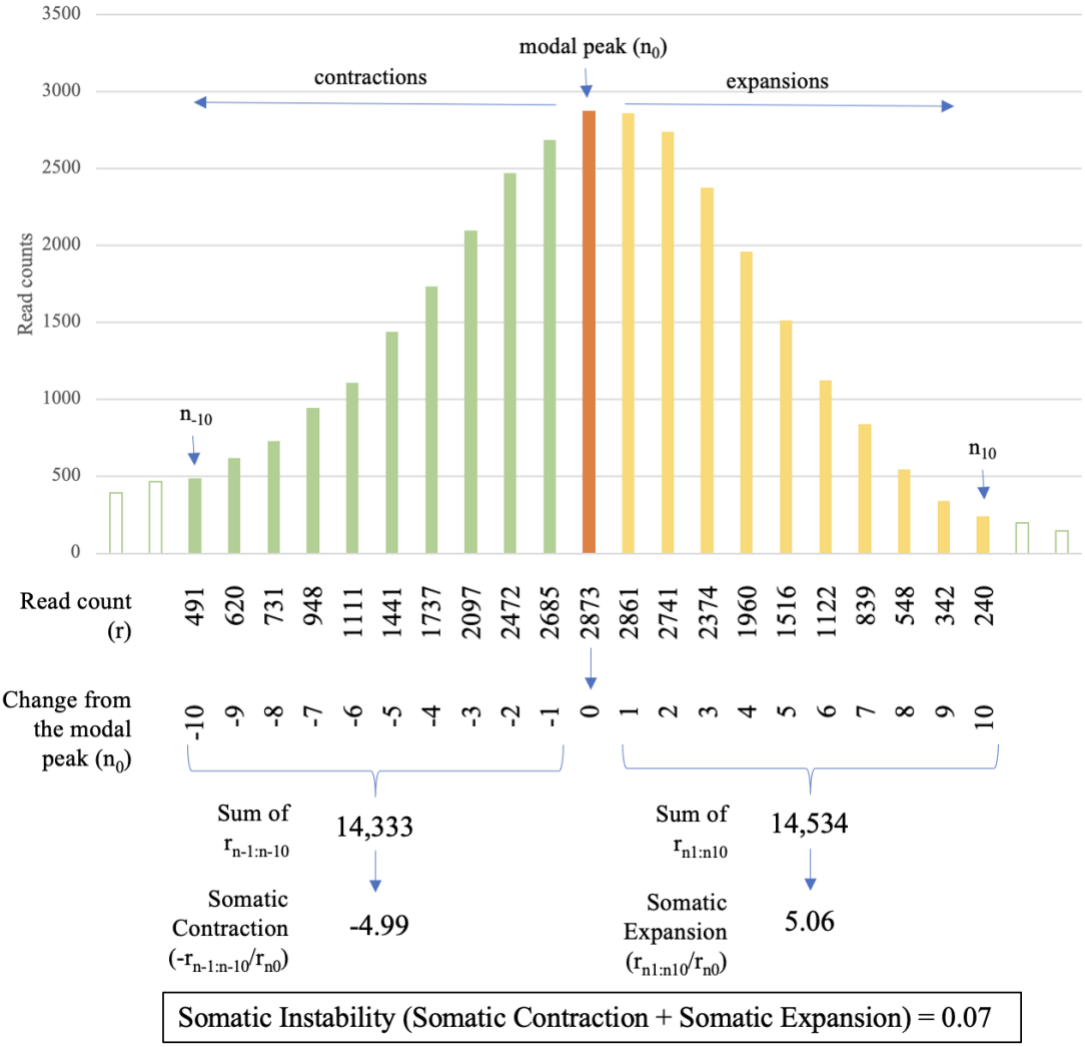


Figure 3.7. Example calculation of Somatic Expansion, Somatic Contraction, and Somatic Instability. Somatic Contraction is calculated by dividing the sum of the read frequencies in the range n-1 to n-10 by the read frequency of the modal peak (n0). Somatic Expansion is calculated by dividing the sum of the read frequencies in the range n1 to n10 by the read frequency of the modal peak (n0). Method and figure modified from Lee et al. 2010.

Because ScaleHD was designed to analyse Illumina short-read data, reads from the 3 kbp amplicon experiment had to be trimmed and filtered to make them compatible with ScaleHD analysis. This was done with Cutadapt, Seqkit and GNU Parallel (Martin 2011; Shen et al. 2016; Tange 2018) using a script written by Branduff McAllister. MiSeq and PacBio reads were then run on ScaleHD using the same parameters. Prior to alignment, ScaleHD performs a round of sequence trimming, which acts as another layer of quality control as reads are rejected if a 10 bp sequence flanking the repeat is not found. See 2.7.2 for the full parameters and scripts used with ScaleHD. Analysis failed due to low read count for a single allele in 27 out of 48 samples in the PacBio data – all had fewer than 200 reads in the modal CAG – and failed of both alleles in 5 samples. For single allele failures, modal CAG and Somatic

Expansion were extracted from the ScaleHD analysis output. All samples with atypical flanking sequences identified from the MiSeq analysis were checked for correct pure CAG length using the sequence viewer Tablet (Milne et al. 2013) and corrected manually where necessary. Alleles with fewer than 30 reads in the modal repeat, of which there were 3 (all expanded), were removed. After removing failures and low repeat counts there were 43 pairs of WT alleles and 40 pairs of expanded alleles to compare.

In comparing PacBio to MiSeq data, I wanted to test the prediction that PacBio sequencing will give the same repeat length as MiSeq using ScaleHD. If so, the mean modal CAG repeat sizes from each sequencing method will differ by no more than 1 CAG and paired values of modal CAG will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that PacBio sequencing will give comparable measures of somatic expansion as MiSeq using ScaleHD. If so, mean Somatic Expansion from each sequencing method will differ by no more than 1 and paired values of Somatic Expansion will be significantly positively correlated ($\alpha = 0.05$).

Table 3.4 summarises the comparison made between ScaleHD CAG count calls on MiSeq and PacBio data. Modal CAG lengths called for the WT allele on the PacBio data agreed with the MiSeq calls in all but 1 of the 43 available samples. In the sample that did not agree, the PacBio CAG length was 25 and the MiSeq CAG length 26. Modal CAG lengths called for the expanded allele on the PacBio data agreed with the MiSeq calls in 23 of the 40 available samples. Of the samples that showed a difference, the largest difference between them was 1 CAG. In all but one of the 17 alleles with a difference in modal CAG, the PacBio call was shorter by 1. Despite the relatively low read depth per sample for the PacBio expanded alleles (typically a few hundred reads with the modal CAG), they all fell within +/- 1 CAG of the MiSeq calls. This is within the error for diagnostic methods of quantifying CAG length (Losekoot et al. 2013).

Data from Lee et al. show that somatic expansion is clearly visible at an instability index of 5.8 (with few contractions) (Lee et al. 2010). This would translate to an expansion index higher than 5.8 as contractions are subtracted to calculate instability index (see Figure 3.7). Values of somatic expansion are comparably low in both MiSeq and PacBio data at 0.305 and 1.35 respectively, suggesting little expansion has occurred in these samples.

Comparison	Modal CAG		Somatic Expansion	
Allele	WT	Expanded	WT	Expanded
N	43	40	39†	39††
Mean MiSeq	18.9	42.5	0.0162	0.305
Mean PacBio	18.9	42.1	0.317	1.35
SD MiSeq	4.22	2.33	0.014	0.178
SD PacBio	4.18	2.11	0.276	0.413
Normal	No	Yes	No	Yes
R^2 or r_s	1.00	0.975	0.640	0.114
p-value	$7.85x10^{-86}$	$1.76x10^{-26}$	$1.13x10^{-5}$	0.490

Table 3.4. Comparison of ScaleHD calls of MiSeq and PacBio data of the *HTT* locus of library 3000-LBC-PBMC. N represents the number of samples analysed. † 4 outliers removed; †† 1 outlier removed. Outlier: any value more than 2.5 SDs from the mean. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. rs is the Spearman's rank correlation coefficient. R^2 or rs: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. WT: wild type.

Modal CAG and Somatic Expansion (see Figure 3.7 for calculation of Somatic Expansion and Contraction) correlations were performed between ScaleHD counts of MiSeq and PacBio data for WT and expanded alleles (Table 3.4). After testing each set of Modal CAG values and Somatic Expansion for normality (Shapiro-Wilk, $P \leq 0.05$) (see 2.8.3) and removing any outliers (more than 2.5 standard deviations from the mean), Pearson correlation coefficients (R) were calculated if normally distributed and Spearman's rank correlation coefficients (r_s) if not (see 2.8.3).

As shown in Table 3.4., Modal CAG counts were strongly significantly positively correlated in both the WT and expanded allele, and the difference between mean MiSeq and PacBio counts was less than 1 (0.02 and 0.38 for the WT and expanded alleles respectively), suggesting the two methods give the same repeat length and adding weight to the use of PacBio data in CAG repeat analyses.

Differences between the mean Somatic Expansion were less than one for the WT allele and slightly more than 1 for the expanded allele. PacBio was higher than MiSeq by 0.301 for the WT allele and 1.05 for the expanded allele. The larger than predicted difference in the expanded allele is likely to be driven in part by the difference in read depth between the two sequencing methods – and therefore signal-to-noise ratio – but may also be driven by the number of PCR cycles in the library preparation. MiSeq library preparation involved 28 PCR cycles, while PacBio involved 50. In STRs, more

PCR cycles is associated with more PCR stutter. PCR stutter typically results in a leftward skew in repeat length distributions. This issue is discussed further in the discussion of this chapter.

Somatic Expansion was significantly positively correlated for the WT allele but not correlated for the expanded allele. This is likely to be an effect of the difference in signal-to-noise ratio between the two alleles compared to their MiSeq equivalents. Somatic Expansion standard deviations are greater in the PacBio in both alleles, suggesting there is more noise in this data, but particularly in the expanded allele, and this is likely to be driven by the differences in read depth. The number of reads equal to the WT modal CAG is on average 13,344 for MiSeq, 30-fold higher than PacBio at 440; the number of reads equal to the expanded modal CAG is on average 9,713, 51-fold higher than PacBio at 189.

ScaleHD has some limitations when applied to PacBio long-read sequencing data of *HTT* CAG repeats. Because it is designed to work with short or medium read lengths (< 400 bp), long PacBio reads must be trimmed and formatted prior to analysis, and even then, calling often fails when there are < 200 reads in the modal peak. In addition, because ScaleHD analysis aligns reads to a finite number of reference sequences, it is not capable of detecting novel interruption structures, the presence of which must be ascertained by checking the data manually. Furthermore, the reference library upon which ScaleHD relies must be adapted for very long reads (e.g. from the Q109 cells) and this is a very laborious and inefficient way of calling repeat lengths.

3.3.1.3.2. Python-based CAG counting methods

To investigate whether alternative repeat calling methods not dependent on alignment to reference sequences could be applied to PacBio data, I tested several other algorithms. Firstly, I extracted the CAG tracts from each read by deploying a string searching pattern coded in Python that locates 12 bp sequences that flank the CAG-CCG repeat (see 2.8.1 and Figure 2.1). Trimming removes the flanks, leaving the CAG-CCG repeat only. Then the first instance of "CAGCAG" from 3' to 5' within the extracted CAG-CCG is marked as the 3' end of the pure CAG tract, which is then extracted.

First, I counted the occurrences of "CAG" in each pure CAG tract using a string searching function (Figure 3.8A). Modal CAG lengths for the WT and expanded allele

were consistently 2 or 3 repeats shorter than the equivalent MiSeq-ScaleHD counts, respectively. This difference could be due to the use of fewer PCR cycles in the MiSeq data but could result from the fact that this method only counts exact "CAG" matches. Deletion, insertion, and substitution errors within individual CAG tracts result in reduced CAG lengths. The presence of 1 or more of these errors per read was common, likely leading to some count deflation of modal CAG counts.

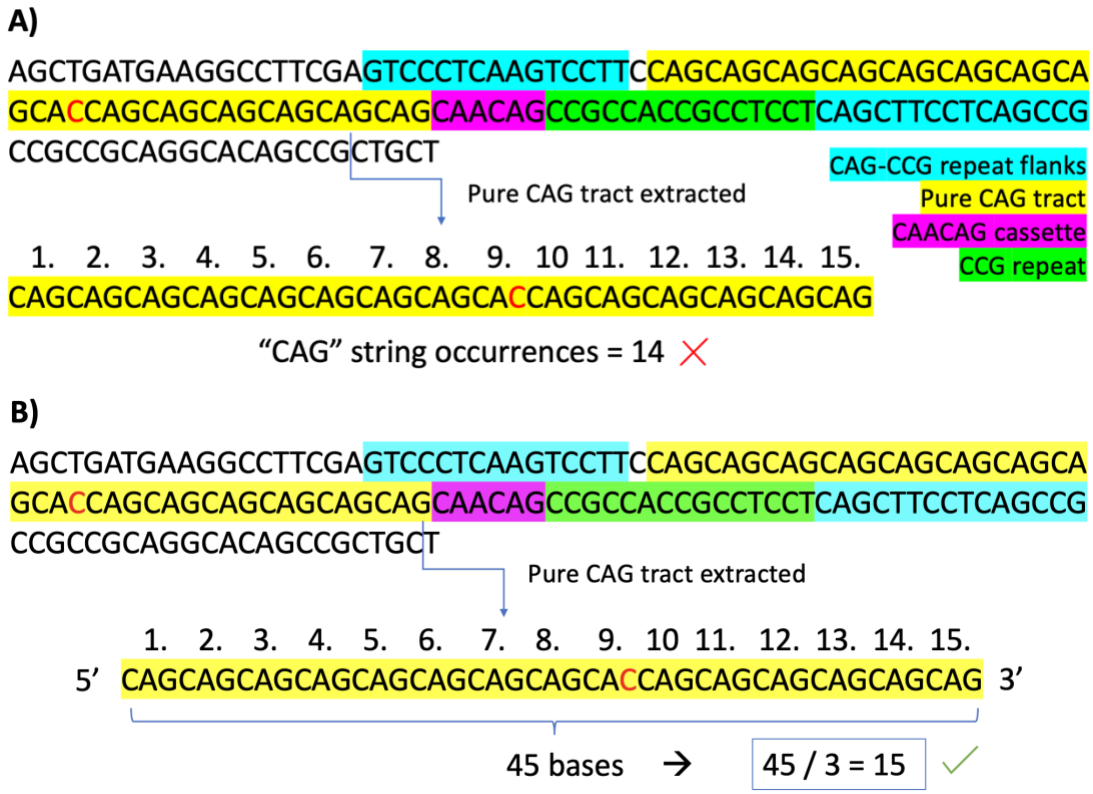


Figure 3.8. Pure CAG length determination methods tested on PacBio sequencing data. (A) "CAG" string occurrences. 12 bp flanking sequences are used to locate the CAG-CCG repeat. The pure CAG tract is extracted and the number "CAG" string matches in the extract is given as the CAG length. (B) Pure CAG tract length divided by 3. As in (A) but the length of the extract in base pairs divided by 3 is given as the CAG length. Substitutions (red text) results in 'count deflation' in (A) but not in (B).

In the second method tested, I divided the length of the pure CAG tract strings by 3 to give an approximate CAG count (Figure 3.8B). While the modal counts for the WT allele were consistently the same as MiSeq and the counts of the expanded allele were broadly centred around the MiSeq modal CAG count, the expanded allele's modal count itself was often 1 or 2 more or less than MiSeq modal CAG count. As this method relies on the base pair length of the repeat, any insertions or deletion errors will inflate or deflate the overall CAG count and will do so by a third of a CAG per error, leading to many instances of non-integer CAG counts. This sensitivity to errors may explain why many of the modal CAGs did not match up to MiSeq-ScaleHD data.

ScaleHD is better able to handle insertions and deletions by assigning reads CAG lengths based on alignment score.

A weakness of both methods is that the rely on accurate extraction of the pure CAG tract, which in turn relies on error-free flanking sequences and an error-free "CAGCAG" at the 3' end of the pure CAG tract.

3.3.1.3.3. RepeatDecoder

Next, I tested a novel repeat counting algorithm developed by Thierry Schuepback and Vincent Dion at the University of Lausanne called RepeatDecoder (RD) (see 2.7.1). RD works by aligning short tandem repeats (in the example in Figure 3.9 the repeating unit is "CAG") of increasing length to a target read, giving each a score based on how closely the sequence matches the read's pure CAG tract. The repeat length with the highest score is given as the CAG length. Scoring profiles can be modified to change RD's counting behaviour. In the example in Figure 3.9B, the permissive profile penalises the "CAA" less harshly than the restrictive profile so that the CAACAG cassette is included, changing the top scoring repeat length from 14 to 16. Profiles with different behaviours can be combined to accurately determine the location and sequence of repeat flanking sequences (Figure 3.9C).

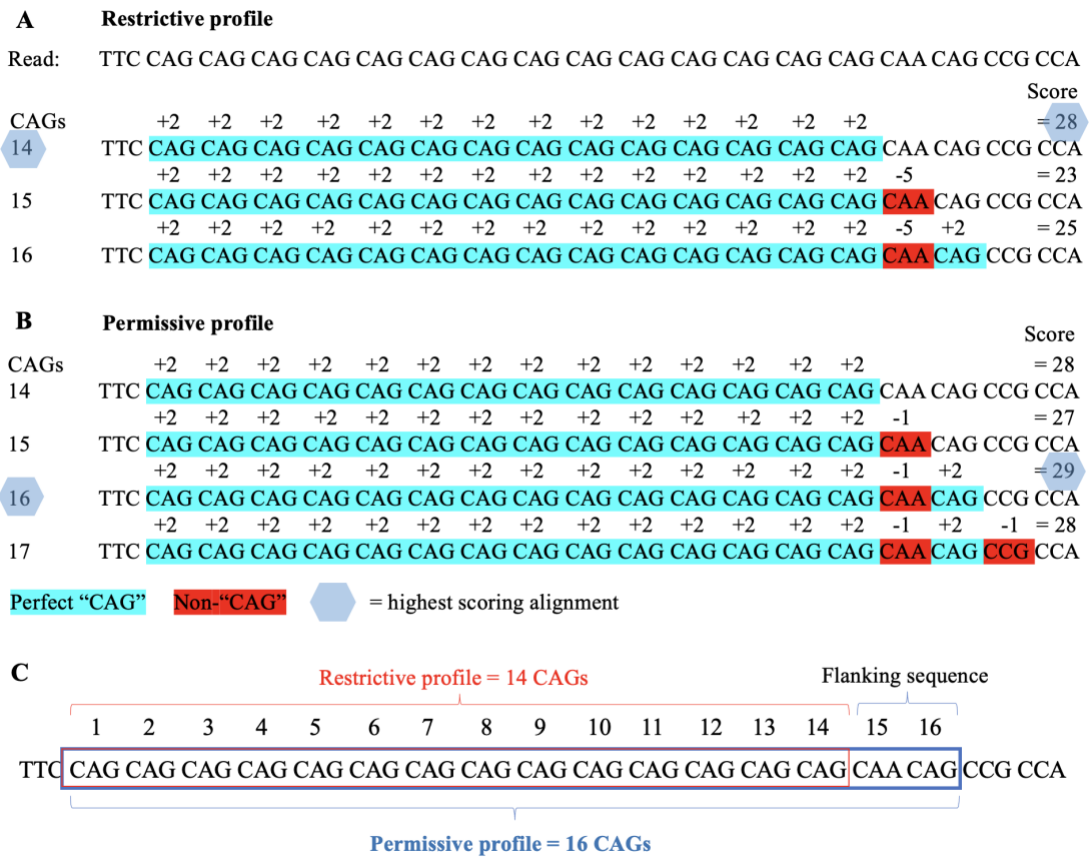


Figure 3.9. Illustration of how the RD counting method works. (A) The restrictive profile measures the uninterrupted polyCAG length. The CAG length with the highest alignment score is given as the CAG length. (B) The permissive profile measures the polyQ length. (C) The sequence between the 3' ends of the restrictive and permissive profiles is the 'flanking sequence'.

While the restrictive profile is sensitive to substitutions and indels in the CAG repeat, particularly at the 5' and 3' ends, the penalty scores have been set to tolerate non-CAGs within the repeat tract. In addition, prior extraction of the pure CAG tract is not needed, meaning reads with substitutions or indels in the CAG-CCG flanks are assigned CAG counts.

To validate RD, and to further validate PacBio sequencing, data from the PacBio library 3000-LBC-PBMC and matching MiSeq samples (the same data used in the ScaleHD comparison in Table 3.4) was analysed in a pipeline utilising RD and compared. Restrictive profile counts were selected for this purpose as this counts the pure CAG tract length, or polyCAG length. It has been shown that this is a better predictor of HD symptom onset age than polyglutamine length (Lee et al. 2019). RD data was then imported into my Python analysis pipeline and paired with FASTQ data before applying the filtering steps detailed in Figure 2.1.

To call the modal CAGs of both alleles from RD CAG length distribution data, I split the data into two groups. Reads with 35 CAGs were categorised as WT and 36 or more categorised as expanded, as the longest WT allele CAG length was 34 and the shortest expanded allele 38. The Read with the highest frequency in each group were called the WT allele and expanded allele modal CAGs respectively.

Figure 3.10 shows a strong agreement between RD's CAG counts of PacBio and MiSeq data, despite the relatively low number of PacBio reads. The position and shape of peaks are well matched for both the WT and the expanded alleles. MiSeq modal peak read frequency percentages are consistently higher, presumably due to better signal to noise ratio afforded by the higher read depth and accuracy. MiSeq read depths are typically 10 to 20-fold higher for the WT allele and 15 to 30-fold higher for the expanded allele, when compared with PacBio.

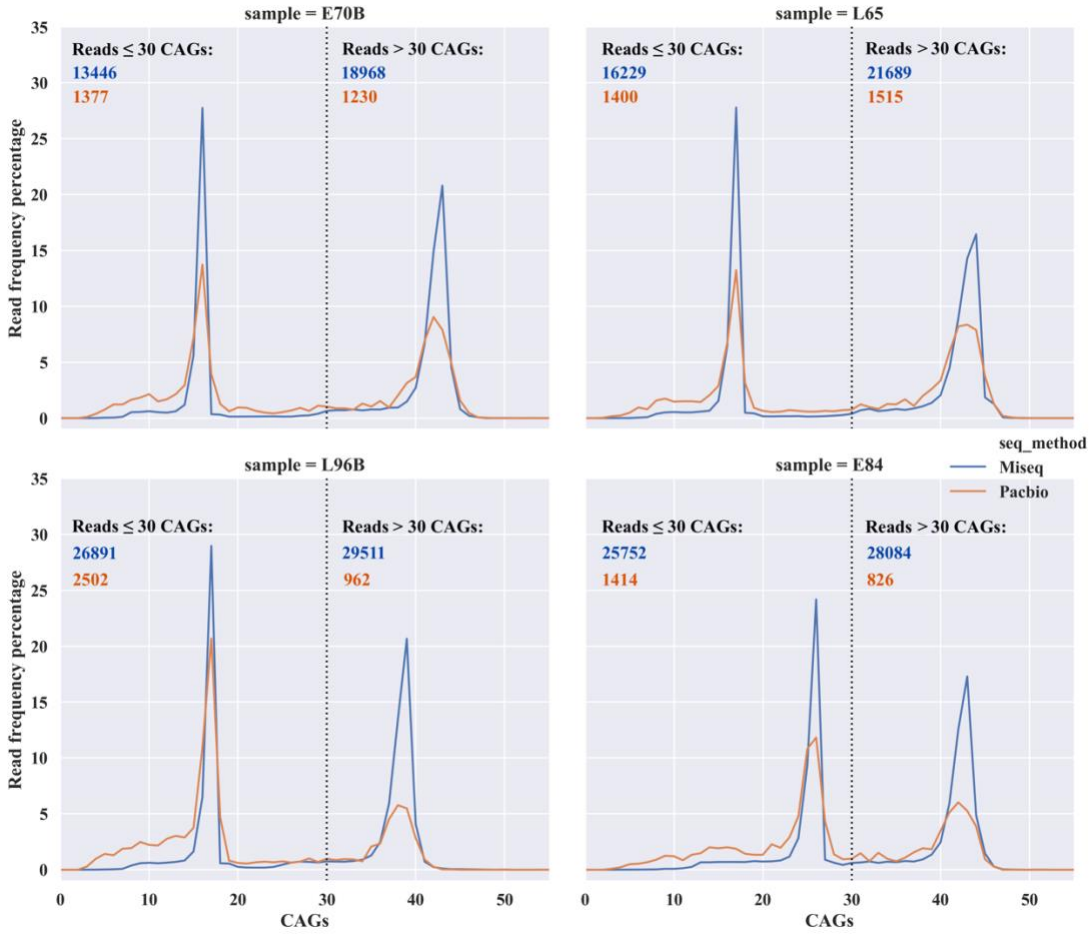


Figure 3.10. Comparison of RD restrictive CAG counts of MiSeq and PacBio data of the *HTT* locus for four representative samples. RD CAG count frequencies were calculated for all 48 patient DNA samples sequenced by two sequencing methods. Read frequency percentage represents the frequency of reads at a given CAG length as a percentage of all reads in that sample (for that sequencing method). The raw frequencies of reads greater than 30 and less than or equal to 30 are given for each sample for each sequencing method. MiSeq data is represented by blue lines and numbers, PacBio data by orange lines and numbers. Four representative samples were chosen for the figure.

In comparing PacBio to MiSeq data, I wanted to test the prediction that PacBio sequencing will give the same repeat length as MiSeq when using RD to count CAGs as it did using ScaleHD. If so, the mean modal CAG repeat sizes from each sequencing method will differ by no more than 1 CAG and paired values of modal CAG will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that PacBio sequencing will give comparable measures of somatic expansion as MiSeq using RD. If so, mean Somatic Expansion from each sequencing method will differ by no more than 1 and paired values of Somatic Expansion will be significantly positively correlated ($\alpha = 0.05$).

RD called the modal CAG length of the WT allele consistently between PacBio and MiSeq in 48 out of 48 samples. The expanded alleles of 2 PacBio samples had modal

peaks with fewer than 30 reads, these alleles were removed from the analysis. Of the remaining 46 pairs of expanded alleles, the modal CAG lengths were identical in 16 (34.8%). In the other 30 pairs (65.2%), all PacBio calls were one CAG shorter than their MiSeq equivalent. Even though there are a higher proportion of expanded allele modal CAG agreements between PacBio and MiSeq with ScaleHD than with RD (57.5% vs 34.8%), the mean modal CAG difference is 0.65 for RD – compared to 0.40 for ScaleHD.

To further assess the consistency between the two sequencing technologies and the reliability of RD, I ran RD on the same data used to produce the correlations in Table 3.4. My custom analysis pipeline identified the WT and expanded allele modal peaks and calculated Somatic Expansion for each. Normality tests were used to decide the correlation method, as before, and outliers, defined as a value +/- 2.5 standard deviations from the mean, were removed.

Comparison	**Modal CAG**		**Somatic Expansion**	
Allele	WT	Expanded	WT	Expanded
N	48	46	43†	45††
Mean MiSeq	18.9	42.5	0.149	0.487
Mean PacBio	18.9	41.9	0.751	1.43
SD MiSeq	4.21	2.12	0.0795	0.192
SD PacBio	4.21	2.27	0.352	0.389
Normal	No	No	No	No
r_s	1	0.973	0.502	0.0403
p-value	0	9.53×10^{-30}	6.02×10^{-4}	0.793

Table 3.5. Comparison of RD calls of MiSeq and PacBio data of the *HTT* locus. N represents the number of samples analysed. † 5 outliers removed; †† 1 outlier removed. Outlier: any value more than 2.5 SDs from the mean. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. rs is the Spearman's rank correlation coefficient. p-values derived from a 2-tailed t-test of the correlation coefficient. WT: wild type.

As shown in Table 3.5, correlations for modal CAG were again strong, positive, and significant using RD and the difference between means is less than one in both alleles (0.01 for WT, 0.65 for expanded allele). These results are in line with the prediction that PacBio and MiSeq will give the same repeat length using RD. It is worth noting that the mean PacBio modal CAG count is lower in all 4 direct comparisons compared with the CAG count given by MiSeq.

Somatic Expansion was within 1 between sequencing methods for both alleles. WT Somatic Expansion was significantly positively correlated between the sequencing technologies but expanded allele Somatic Expansion was not. So, the prediction that these sequencing methods will give the same somatic expansion using RD is valid for the WT allele but not for the expanded allele. RD is the same as ScaleHD in this respect. The number of samples in the PacBio expanded allele data with fewer than 200 reads in the modal CAG was 25, which, while fewer than the ScaleHD output, is still high compared to the WT allele with 4.

3.3.1.4. Direct comparison of CAG counting methods: RD vs ScaleHD

3.3.1.4.1. MiSeq data

I then compared different CAG counting methods. To compare CAG counting of RD restrictive profile against ScaleHD directly, first I ran them on identical, validated input data, namely that of the MiSeq (same 48 samples used in Table 3.5). Normality tests were used to decide the correlation method, as before, and outliers (outside +/- 2.5 SDs from the mean) were removed. Table 3.6 summarises the results.

In comparing RD to ScaleHD repeat length measures, I wanted to test the prediction that RD will give the same repeat length as ScaleHD on MiSeq data. If so, the mean modal CAG repeat sizes from each counting method will differ by no more than 1 CAG and paired values of modal CAG will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that RD will give comparable measures of somatic expansion to ScaleHD on MiSeq data. If so, mean Somatic Expansion from each counting method will differ by no more than 1 and paired values of Somatic Expansion will be significantly positively correlated ($\alpha = 0.05$).

Comparison	Modal CAG		Somatic Expansion	
Allele	WT	Expanded	WT	Expanded
N	48	47	46†	48
Mean SHD	18.9	42.5	0.0204	0.355
Mean RD	18.9	42.5	0.146	0.518
SD SHD	4.21	2.08	0.0173	0.287
SD RD	4.21	2.08	0.0772	0.269
Normal	No	No	No	Yes
R^2 or r_s	1	1	0.636	0.960
p-value	0	0	2.05×10^{-6}	5.76×10^{-27}

Table 3.6. Comparison of ScaleHD and RD calls of MiSeq data of the *HTT* locus. N represents the number of samples analysed. † 2 outliers removed. Outlier: any value more than 2.5 SDs from the mean. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. rs is the Spearman's rank correlation coefficient. R^2 or r_s: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. SHD: ScaleHD, RD: RepeatDecoder, WT: wild-type.

Modal CAG counts were identical between counting methods for all samples in both the WT and expanded alleles (ScaleHD counts corrected for the polyCAG length), meaning mean modal CAGs were within 1 and modal CAGs were significantly positively correlated. This is in line with the prediction that the two counting methods will give the same repeat length on MiSeq data.

Somatic Expansion was within 1 and significantly positively correlated between counting methods for both the WT and expanded allele. This is in line with the prediction that RD will give comparable measures of somatic expansion to ScaleHD on MiSeq data. While Somatic Expansion is greater in RD in both alleles, the difference in the WT allele is more pronounced, with RD being 7-fold higher. The standard deviation is 4.5-fold higher. This difference is likely to be driven by the different sensitivity to substitutions/indels in the two counting methods. While RD can tolerate some mismatches, penalties applied to CAGs near the start/ends of CAG repeats are too high to be overcome by the positive scores of CAGs between the mismatch and the nearest end, resulting in truncated CAG counts. Using a reference sequence alignment, ScaleHD doesn't have this issue.

At a relatively low rate, as in the MiSeq data, mismatches don't affect RD's modal CAG counts but do affect individual read counts. A small proportion of expanded allele reads are truncated by RD, skewing the distribution towards the WT allele,

which is in turn driving up Somatic Expansion values for the WT allele. The difference in variation likely results from the variation in distances between the alleles between samples and therefore the number of reads falling inside the modal CAG +10 calculation window.

3.3.1.4.2. PacBio data

To further assess the reliability of the CAG counting methods I compared the output of ScaleHD and RD on the same PacBio dataset (same 48 samples used in all correlations above). Normality tests were used to decide the correlation method, as before, and outliers (values outside +/- 2.5 SDs from the mean) were removed. Table 3.7 summarises the results.

In comparing RD to ScaleHD repeat length measures, I wanted to test the prediction that RD will give the same repeat length as ScaleHD on PacBio data, as it does on MiSeq data. If so, the mean modal CAG repeat sizes from each counting method will differ by no more than 1 CAG and paired values of modal CAG will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that RD will give comparable measures of somatic expansion to ScaleHD on PacBio data. If so, mean Somatic Expansion from each counting method will differ by no more than 1 and paired values of Somatic Expansion will be significantly positively correlated ($\alpha = 0.05$).

Comparison	**Modal CAG**		**Somatic Expansion**	
Allele	WT	Expanded	WT	Expanded
N	43	40	38†	38††
Mean SHD	18.9	42.1	0.317	1.35
Mean RD	18.9	41.9	0.654	1.40
SD SHD	4.18	2.33	0.0845	0.402
SD RD	4.22	2.29	0.250	0.368
Normal	No	No	No	No
r_s	1.00	0.958	0.221	0.292
p-value	7.85×10^{-86}	3.16×10^{-22}	0.181	0.0754

Table 3.7. Comparison of ScaleHD and RD calls of PacBio data of the *HTT* locus. N represents the number of samples analysed. † 5 outliers removed; †† 2 outliers removed. Outlier: any value more than 2.5 SDs from the mean. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. r_s is the Spearman's rank correlation coefficient. p-values derived from a 2-tailed t-test of the correlation coefficient. SHD: ScaleHD, RD: RepeatDecoder, WT: wild type.

The difference between RD and ScaleHD mean modal CAG counts is 0 for the WT allele and 0.22 for the expanded allele, i.e., both are within 1. Modal CAG standard deviations are also very similar. Modal CAG counts were significantly positively correlated between the two counting methods for both the WT and expanded alleles. These results are in line with the prediction that RD will give the same repeat length as ScaleHD on PacBio data.

Somatic Expansion was within 1 not significantly positively correlated between counting methods for both the WT and expanded allele. This is not consistent with the prediction that RD will give comparable measures of somatic expansion to ScaleHD on MiSeq data. As with MiSeq data, RD has higher values of Somatic Expansion, particularly for the WT allele, although the ratio of difference between counting methods is far lower. Somatic Expansion standard deviations are similar in the expanded allele but approximately 3-fold higher for the WT allele.

I decided to use RD for *HTT* sequence analysis in all subsequent analyses for four main reasons. Firstly, it performs comparably with Scale HD in terms of CAG length determination. Secondly, it can be used to detect novel flanking sequence structures, unlike ScaleHD. Thirdly, because it is immediately compatible with long reads. And fourthly, because it can tolerate substitutions and indels in and around CAG repeat without giving non-integer CAG counts.

3.3.1.5. Comparison of CAG counts of lymphoblastoid cells and peripheral blood mononuclear cells

Once PacBio-RD analysis had been shown to produce broadly accurate repeat counts on HD patient *HTT* repeats, I used these methods to compare repeats from different cell types, namely PBMCs and LBCs derived from the same individual. PBMCs, i.e. cells from blood, are routinely sampled in clinic and show small amounts of repeat instability (Ciosi et al. 2019). LBCs are preferrable for analysis as they are more easily stored and used and are renewable, enabling a potentially limitless supply of DNA. What remains unclear is whether they show similar modal CAGs and expansion to PBMCs, and therefore, whether they can be used to follow somatic instability in patients, so I compared the two in individuals where I had sequencing data from both samples.

MiSeq *HTT* sequence data of 41 patients with both PBMC and LBC samples was used to conduct a comparison of ultra-high depth sequencing of the two DNA sources (Table 3.8) before a similar comparison of 30 patient samples was conducted using PacBio-RD data as only 30 pairs of patient DNA were sequenced on PacBio.

In comparing LBC to PBMC repeat length measures, I wanted to test the prediction that LBCs will give the same repeat length as PBMCs using MiSeq-ScaleHD counts. If so, the mean modal CAG repeat sizes from each cell type will differ by no more than 1 CAG and paired values of modal CAG will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that LBCs will give comparable measures of somatic expansion to PBMCs on using MiSeq-ScaleHD counts. If so, mean Somatic Expansion from each counting method will differ by no more than 1 and paired values of Somatic Expansion will be significantly positively correlated ($\alpha = 0.05$).

Comparison	**Modal CAG**		**Somatic Expansion**	
Allele	WT	Expanded	WT	Expanded
N	39	39	39	39
Mean LBC	18.4	42.0	0.0187	0.341
Mean PBMCs	18.4	42.4	0.0211	0.359
SD LBC	3.77	2.14	0.0121	0.354
SD PBMCs	3.81	2.15	0.0194	0.162
Normal	No	No	No	No
r_s	1.00	0.987	0.844	0.273
p-value	0	9.11×10^{-31}	1.50×10^{-11}	0.0931

Table 3.8. Comparison of MiSeq-ScaleHD calls of the *HTT* repeat locus in LBC and PBMC samples. N represents the number of samples analysed. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. rs is the Spearman's rank correlation coefficient. p-values derived from a 2-tailed t-test of the correlation coefficient. LBC: lymphoblastoid cell, PBMC: peripheral blood mononuclear cell, WT: wild type.

41 patients had data from both LBC and PBMC samples. Two patients had WT and expanded allele modal CAGs within 11 CAGs of each other. A small difference between alleles skews Somatic Expansion due to overlapping distributions so these samples were removed from the analysis. Somatic Expansion was calculated as per Figure 3.7. None of the data were normal for both LBCs and PBMCs.

The difference in mean modal CAG between cell types was less than 1 for both WT and expanded alleles. Paired CAG counts were significantly positively correlated in both the WT and expanded alleles. This is in line with the prediction that LBCs will give the same repeat length as PBMCs using MiSeq-ScaleHD counts.

The difference in Somatic Expansion was less than 1 for both WT and expanded alleles. Paired values of Somatic Expansion were correlated in the WT allele but not the expanded allele, so the prediction that LBCs will give comparable measures of somatic expansion to PBMCs using MiSeq-ScaleHD data holds for the WT allele but not the expanded allele.

Very little expansion is observed in either allele in either line, meaning any correlation between samples is essentially capturing the correlation in noise. Scatter plots from the data are shown in Figure 3.11.

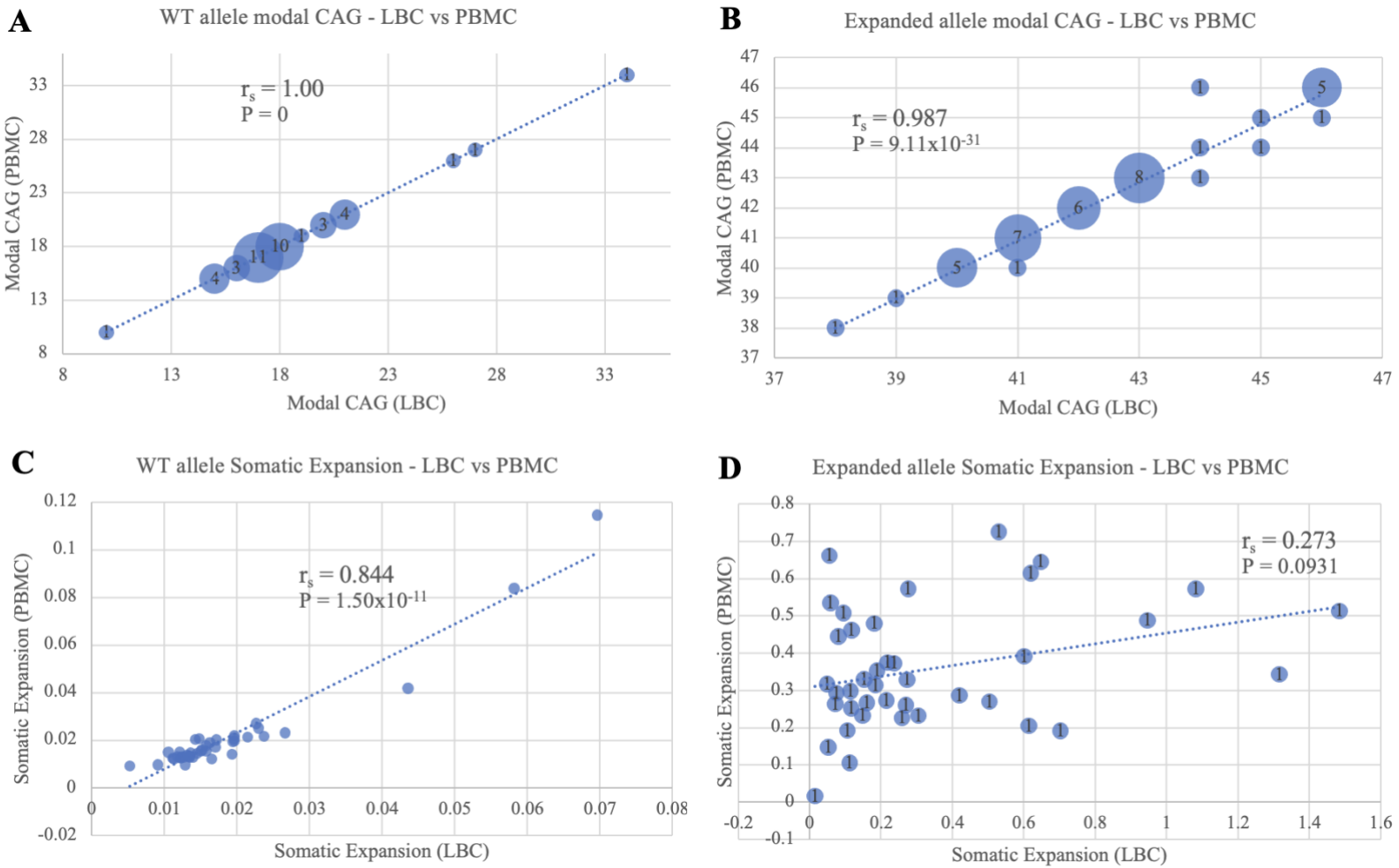


Figure 3.11. Comparison of MiSeq-ScaleHD calls of the *HTT* repeat locus in LBC and PBMC samples. Bubble area and number reflects the number of pairs of data points represented. All points in (C) represent a single LBC-PBMC pair. LBC: lymphoblastoid cell, PBMC: peripheral blood mononuclear cell, rs: Spearman's rank correlation coefficient. P: p-value. Dashed line: line of best fit.

There were 30 patients with both LBC and PBMC samples in our PacBio *HTT* sequencing data. Polished CCS reads were counted using RD and run through our standard analysis pipeline for quality control. The CAG with the highest frequency in each allele is taken to be the modal CAG. Somatic Expansion were calculated as per Figure 3.7.

In comparing LBC to PBMC repeat length measures, I wanted to test the prediction that LBCs will give the same repeat length as PBMCs using PacBio-RD counts, as they do using MiSeq-ScaleHD counts. If so, the mean modal CAG repeat sizes from each cell type will differ by no more than 1 CAG and paired values of modal CAG will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that LBCs will give comparable measures of somatic expansion to PBMCs on using MiSeq-ScaleHD counts. If so, mean Somatic Expansion from each counting method will differ by no more than 1 and paired values of Somatic Expansion will be significantly positively correlated ($\alpha = 0.05$).

Comparison	**Modal CAG**		**Somatic Expansion**	
Allele	WT	Expanded	WT	Expanded
N	28	28	27†	28
Mean LBC	18.1	42.0	0.0927	0.385
Mean PBMC	18.1	42.5	0.0946	0.393
SD LBC	2.93	2.20	0.0319	0.153
SD PBMC	2.92	2.70	0.0336	0.235
Normal	No	No	No	No
r_s	0.677	0.809	0.0598	-0.107
p-value	7.63×10^{-5}	1.91×10^{-7}	0.767	0.587

Table 3.9. Comparison of PacBio-RD calls of the *HTT* repeat locus in LBC and PBMC samples. N represents the number of samples analysed. † 1 outlier removed. Outlier: any value more than 2.5 SDs from the mean. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. rs is the Spearman's rank correlation coefficient. p-values derived from a 2-tailed t-test of the correlation coefficient. LBC: lymphoblastoid cell, PBMC: peripheral blood mononuclear cell, WT: wild type.

2 patients had WT and expanded alleles within 11 CAGs of each other and were removed from the analysis due to overlapping CAG distributions. One of the remaining 28 patients had an outlying value for WT Somatic Expansion (> mean +/- 2.5 SD) and was removed from the analysis. None of data summarised in Table 3.9 were normally distributed for LBC or PBMC.

The difference in mean modal CAG between cell types was less than 1 for both WT and expanded alleles. Paired CAG counts were significantly positively correlated in both the WT and expanded alleles. This is in line with the prediction that LBCs will give the same repeat length as PBMCs using PacBio-RD counts.

The difference in Somatic Expansion was less than 1 for both WT and expanded alleles. Paired values of Somatic Expansion were correlated in neither the WT allele nor the expanded allele, so the prediction that LBCs will give comparable measures of somatic expansion to PBMCs using PacBio-RD data holds for neither the WT allele nor the expanded allele. Scatter plots of these data are shown in Figure 3.12.

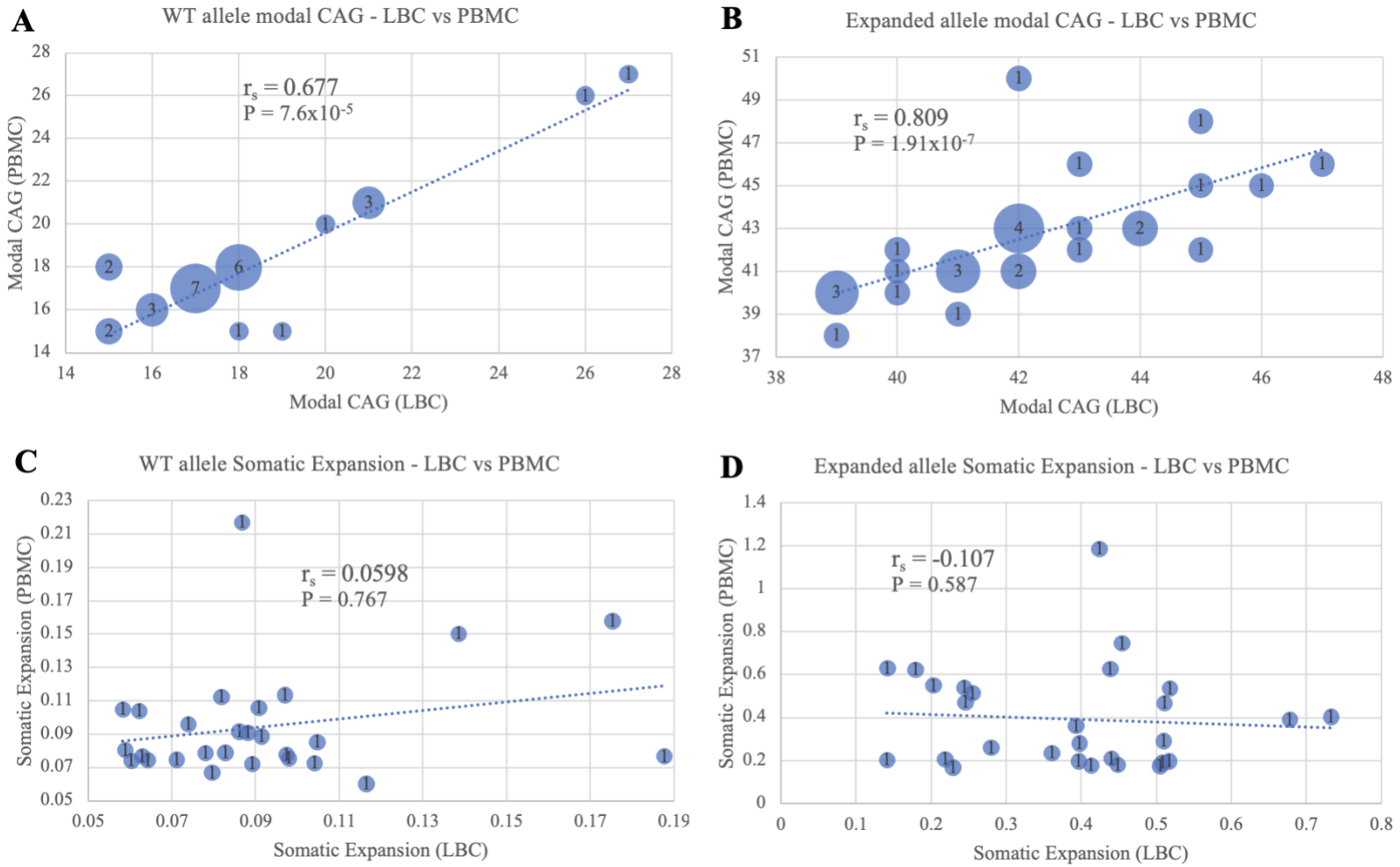


Figure 3.12. Comparison of PacBio-RD calls of the *HTT* repeat locus in LBC and PBMC samples. Bubble area and number reflects the number of pairs of data points represented. rs: Spearman's rank correlation coefficient. P: p-value. LBC: lymphoblastoid cell, PBMC: peripheral blood mononuclear cell. Dashed line: line of best fit.

3.3.2. Sequencing Cell Models with 130 CAG repeats and Increasing Depth

To investigate whether the methods developed above could be used to examine longer repeats, such as those found in cell and animal models of HD, I sequenced the *HTT* CAG repeat of induced pluripotent stem cells (iPSCs) derived from a HD patient originally with 109 CAGs (The HD iPSC Consortium et al. 2012). These lines show somatic expansion over time in culture and in our laboratory now have around 130 CAGs.

iPSCs are a type of stem cell which have been reprogrammed from an adult somatic cell for pluripotency, i.e., it can be differentiated into many different cell types. Like embryonic stem cells, iPSCs, are renewable and can be differentiated into complex, disease relevant cell types such as MSNs. Since the method for their generation was first described in 2006 (Takahashi and Yamanaka 2006), iPSCs, enabled by gene editing techniques such a CRISPR, have been used extensively to model human neurological diseases (Rowe and Daley 2019). In 2012, the HD iPSC consortium generated 14 iPSC lines from HD patients and controls, including the 109 CAG line (The HD iPSC Consortium et al. 2012).

To examine whether *FAN1* has an influence on somatic expansion in this cell model, Jasmine Donaldson generated isogenic *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ lines from the 109N1 parent line using CRISPR-Cas9 gene editing (Donaldson 2019; McAllister et al. 2022) (see 2.1). A family tree of the different lines available is shown in Figure 3.13.

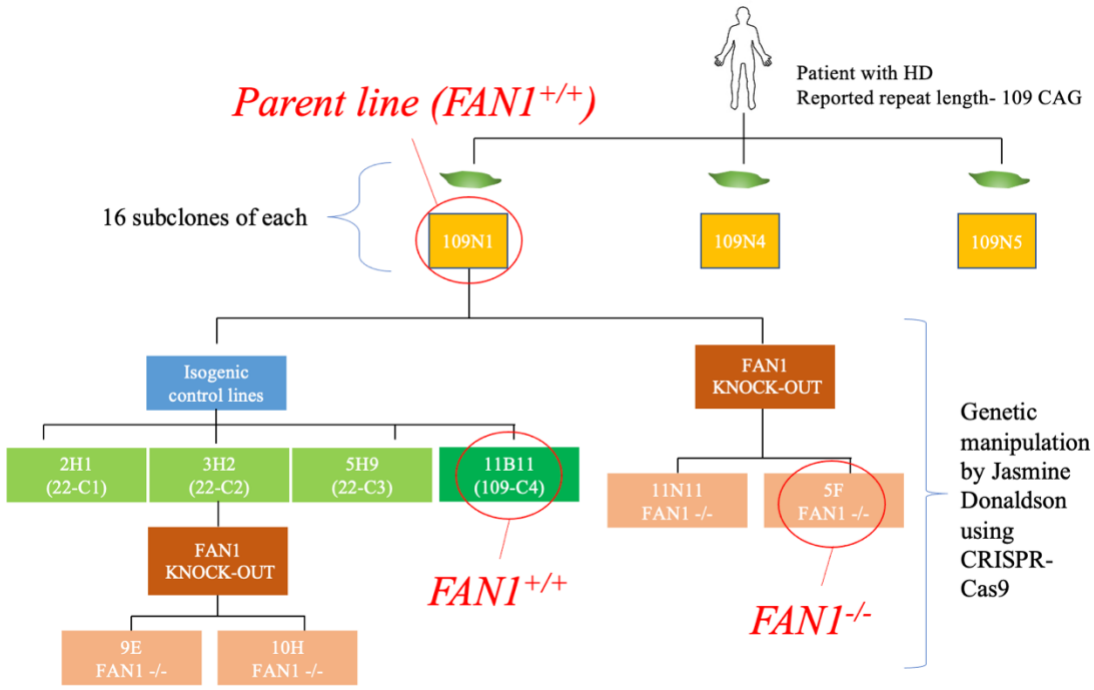


Figure 3.13 iPSC family tree, including generation of isogenic *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ 109NI lines. Figure adapted with permission from Donaldson 2019.

3.3.2.1. Culture of cell lines

The iPSCs were cultured in Essential 8 Flex medium and cells were passaged every 3-4 days, at a confluency of ~ 70%. A sample was taken at each passage and genomic DNA was extracted as described in methods section (2.3.1)

3.3.2.2. Library preparation

To establish firstly whether PacBio could read through the entire 130 CAG repeat of Jasmine's iPSC models, and secondly whether PacBio could accurately call the repeat lengths, I chose a range of 109NI-derived samples with different repeat lengths, as determined by fragment analysis. Thirdly, I wanted to see if PacBio data was sensitive enough to detect rare large repeat expansions, so I limited the number of samples in this experiment to 6 to obtain a high read depth.

Library 3000-iPSC was comprised entirely of iPSC samples, from either the parent line, 109N1, or lines derived from 109N1. Sample details are shown in Table 3.10. The family tree in Figure 3.13 shows the relationship of the different iPSC lines used. 109NI-5F is a homozygous *FAN1* knockout line of a subclone of 109N1. 109NI-11B11 is an isogenic subclone of 109NI which is homozygous wild type for *FAN1*. 109NI-SC5 is a different subclone of 109NI.

Sample number	Cell line	P	*FAN1 genotype*	FA modal CAG	ng of DNA in PacBio library
1	109NI-11B11	36	$^{+/+}$	133	772
2	109NI-5F	6	$^{-/-}$	126	1144
3	109NI-5F	32	$^{-/-}$	129	1030
4	109N1	31	$^{+/+}$	117	916
5	109N1	45	$^{+/+}$	128	445
6	109N1-SC5	36	$^{+/+}$	127	400

Table 3.10. PacBio SMRTbell library 3000-iPSC sample details. P: passage number. FA: fragment analysis. ng: nanograms. SC5: subclone 5. See Figure 3.13 for cell line family tree.

3000-iPSC library preparation was conducted using the same protocol used to generate the 3000-LBC-PBMC library (see 2.5.1). The library was run on a capillary electrophoresis chip after SMRTbell adapter ligation to inspect the integrity of the library (Figure 3.14). Amplicons of approximately 3.0 kbp and 3.3 kbp were expected (after allowing for PCR barcoding and adapter ligation) from the WT and expanded alleles respectively (Table 2.4). The major peak at 2,611 bp is within the +/- 15% error for the expected WT allele size for this technology (Agilent Technologies' website: https://www.agilent.com/cs/library/usermanuals/public/quick-guide-dnf-464-large-fragment-50kb-kit-SD-AT000127.pdf ; accessed 23 Dec 2021). The peak at 4,909 is more likely to contain artefacts of library preparation (PCR chimera or ligation-induced dimers) than the expanded allele as it is approximately double the size of the main peak. By contrast, the expanded allele products should be roughly 330 bp larger than the WT allele, meaning it is likely to be contained within the shoulder of the major peak at 2,611 bp.

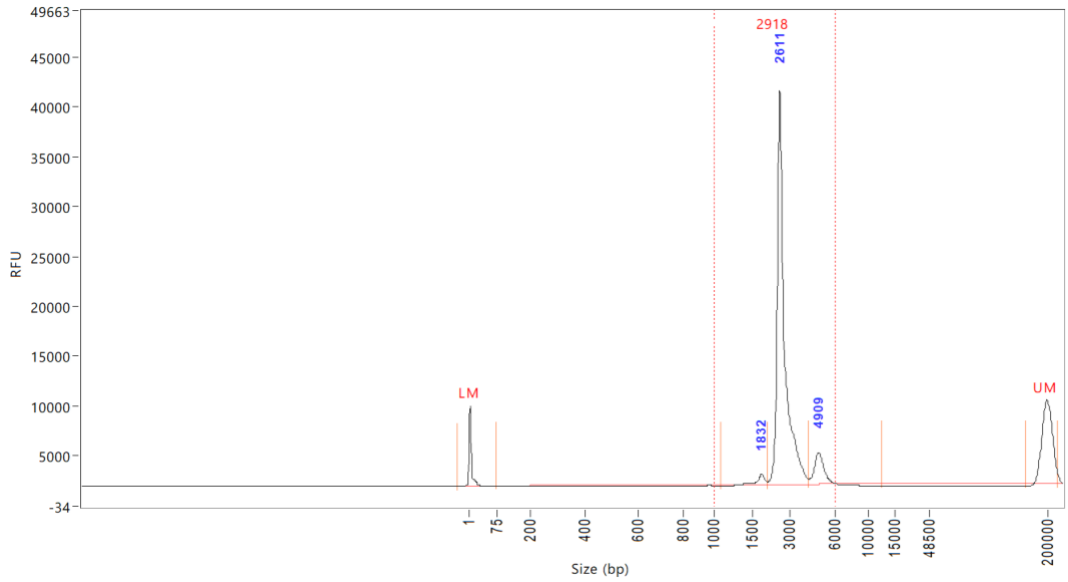


Figure 3.14 – Capillary electrophoresis trace of SMRTbell library 3000-iPSC. RFU: relative fluorescence units, LM: lower marker, UM: upper marker, bp: base pairs. Blue numbers correspond to the size of the peak in bp. Red number (top) is the mean fragment size of the material within the red dashed lines.

3.3.2.3. Sequencing and data quality control

Like 3000-LBC-PBMC, the 3000-iPSC library was sequenced on one chip and run with identical loading parameters. The loading was near optimal with 66% of ZMWs occupied by a single template-polymerase complex, and the run produced sequencing metrics comparable to 3000-LBC-PBMC. 182,789 CCS reads were generated, slightly more than the 166,701 CCS reads generated in the 3000-LBC-PBMC library. Median read quality was Q26, which is the same as 3000-LBC-PBMC (see section 1.6.2 for an explanation of Q scores). 94% of read qualities are between Q10 and Q40 in 3000-iPSC, 95% for 3000-LBC-PBMC (Figures 3.5A and 3.15A). Most reads are in the peak at 3 kbp, with a small shoulder to the right and a peak at 4 kbp, broadly consistent with the capillary electrophoresis trace in Figure 3.14 and library 3000-LBC-PBMC (Figure 3.5B).

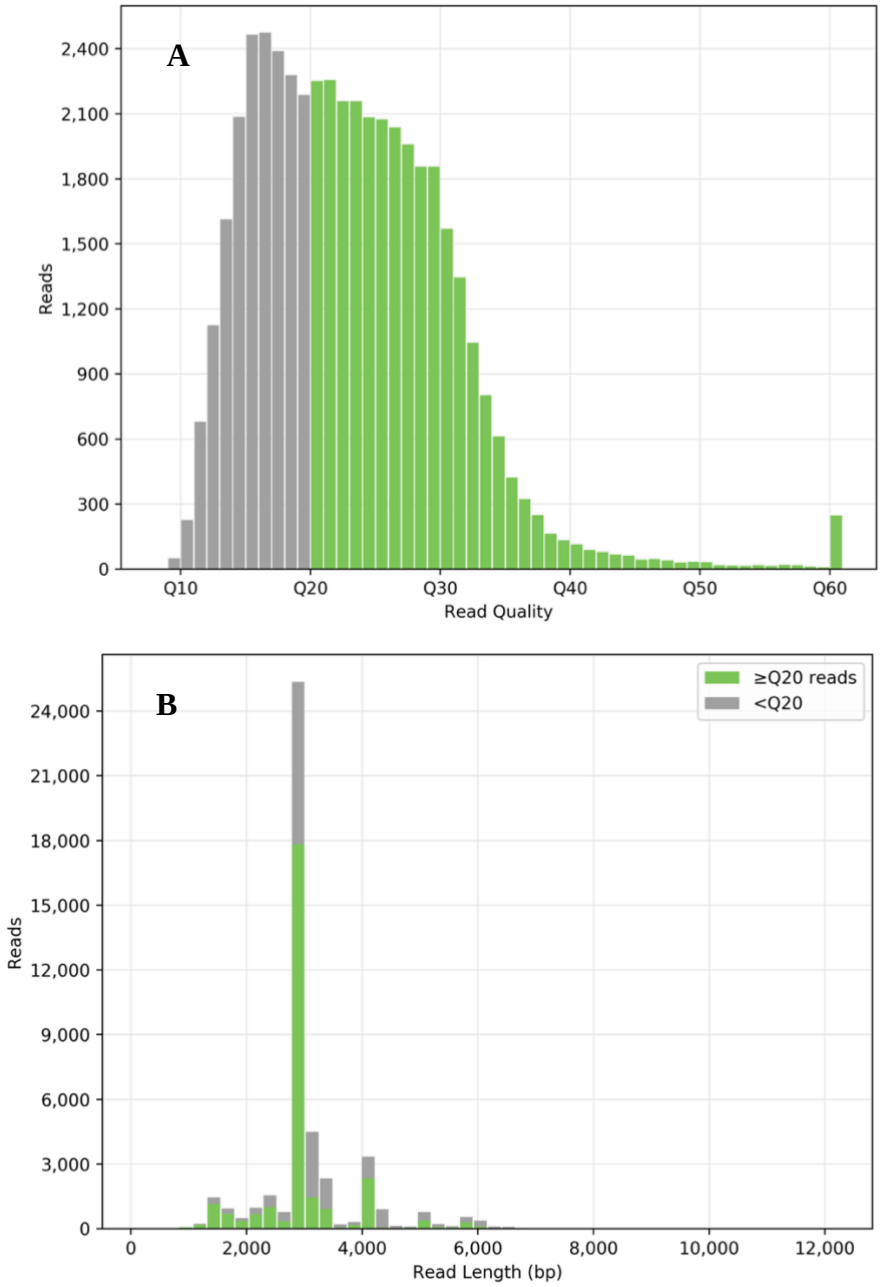


Figure 3.15. Read quality and length distributions for all circular consensus reads in library 3000-iPSC. Q: Phred quality score. HiFi reads (Q20 or higher) shown in green, non-HiFi reads in grey. (A) Read quality distribution. (B) Read length distribution. bp: base pairs.

After filtering, the number of WT and expanded allele reads left was 104,366, slightly more than the 93,069 left for the 3000-LBC-PBMC library after identical filtering. The proportion of WT reads is far higher in the iPSC library at 83.0% of all reads, compared to 48.0% in 3000-LBC-PBMC. The proportion of WT alleles in the iPSC library before *any* filtering is 81.2%, so whatever the mechanism of WT allele enrichment, it must occur before the analysis pipeline. Given the inherent equimolar ratio of WT and expanded alleles in native genomic DNA, the observed bias is likely

due to the preferential amplification of the shorter WT repeat tracts. This issue is addressed in section 3.3.3. with the physical enrichment of expanded alleles during the library preparation.

3.3.2.4. Comparison of 3 kbp iPSC sequencing CAG counts to fragment analysis

As the alleles in the 109NI iPSC samples had alleles with distributions that did not overlap, being centred on approximately 20 and 130 CAGs, I decided to use a different, more robust way of measuring somatic expansion, contraction, and skew. Figure 3.16, adapted from Lee et al. 2010, shows how expansion, contraction and instability indices are calculated. The critical difference between this set of measures and Somatic Expansion, Somatic Contraction and Somatic Instability (calculation shown in Figure 3.7) is the use of a 10% frequency threshold as opposed to a CAG change threshold. This means that any expansion or contraction beyond +/- 10 CAGs will be captured if it is more frequent than 10% of the modal CAG frequency. This both helps to reduce the amount of noise contributing to the measures and enables the capture of bulk expansion/contraction.

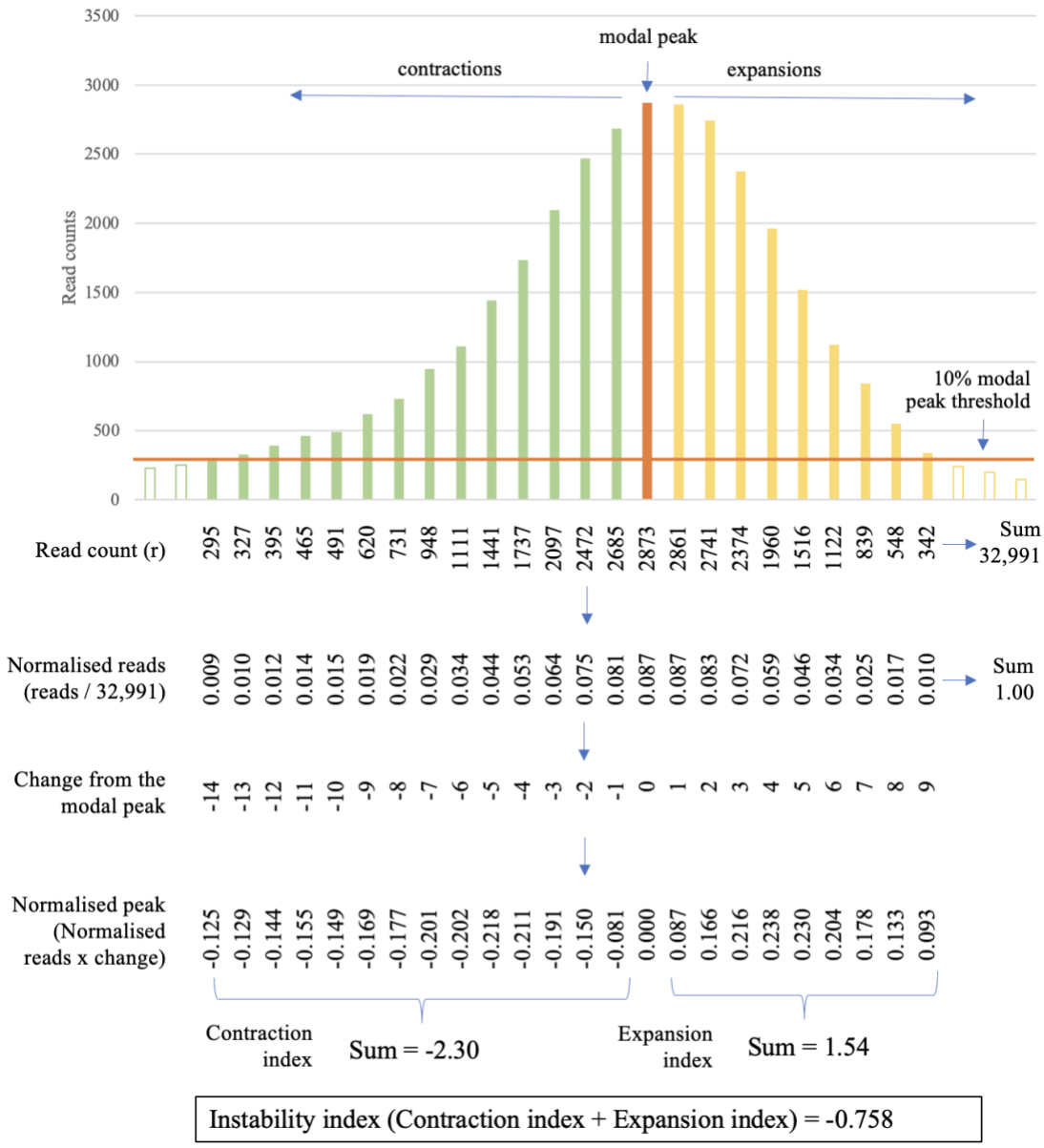


Figure 3.16. Example calculation of expansion index, contraction index and instability index. Contraction index is the sum of the negative products of Normalised reads and Change from the modal peak for those reads with a frequency greater than 10% of the modal peak frequency. Expansion index is the equivalent of the contraction index for the positive products. Instability index is the sum of the contraction and expansion indices. Figure adapted from Lee et al. 2010.

While there are no MiSeq sequencing data to validate the CAG counts of the 3 kbp iPSC libraries against, as MiSeq reads are not long enough to span 130 CAG repeats, fragment analysis provides a clinically validated *HTT* repeat sizing method for comparison. In addition, some samples have been sequenced multiple times, which allows us to assess the variability of iPSC expanded allele counts across technical replicates and sequencing runs. Nearly all the iPSC samples sequenced on PacBio have been sized by fragment analysis. PacBio reads were CAG-counted by RD restrictive profile. QC and further analysis was conducted as described in section

3.3.1.2. Table 3.11 shows a comparison between fragment analysis and PacBio counts of 109NI samples from library 3000-iPSC.

In comparing PacBio-RD calls to fragment analysis data, I wanted to test the prediction that PacBio-RD will give the same repeat length as fragment analysis on 130 CAG cell model repeats. If so, the mean modal CAG repeat sizes from each method will differ by no more than 1 for the WT allele, no more than 3 for the expanded allele (error limits set by the European Molecular Genetics Quality Network (EMQN) for fragment analysis CAG repeat sizing are +/- 1 for alleles < 40 repeats and +/- 3 repeats for alleles > 39 repeats (Losekoot et al. 2013)) and paired modal CAG values will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that PacBio-RD will give comparable measures of somatic expansion as fragment analysis. If so, mean expansion indices from each method will differ by no more than 1 and paired expansion indices will be significantly positively correlated ($\alpha = 0.05$).

Comparison	**Modal CAG**		**Expansion Index**	
Allele	WT	Expanded	WT	Expanded
N	6	6	6	6
Mean FA	18.8	125	0	1.65
Mean PB	19.2	134	0.121	1.11
SD FA	0.839	5.56	0	0.773
SD PB	0.408	6.22	$3.83x10^{-3}$	0.512
Normal	No	Yes	No	Yes
R^2 or r_s	-0.655	0.969	N/A	0.324
p-value	0.158	$1.40x10^{-3}$	N/A	0.508

Table 3.11. Comparison of fragment analysis and PacBio-RD calls from library 3000-iPSC of the *HTT* repeat locus in 109NI iPSC samples. N represents the number of samples analysed. FA: fragment analysis, PB: PacBio-RD. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. r_s is the Spearman's rank correlation coefficient. R^2 or r_s: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. The mean number PacBio reads with the modal CAG was 5555 for the WT allele, and 175 for the expanded allele. WT: wild type.

The difference in mean modal CAG between PacBio and fragment analysis was less than 1 for the WT allele, however, paired modal CAG counts were not significantly positively correlated, which runs counter to the prediction that the two methods will give the same repeat lengths. This may be due to the relatively low spread in the WT

alleles: fragment analysis counts had a range of 2, while PacBio counts a range of 1. The mean difference between counting methods was 0.64 CAGs.

Also counter to the prediction that PacBio and fragment analysis will give the same repeat length, the difference in expanded allele mean modal CAG was more than 3 – the mean modal CAG was 8.8 higher for PacBio than fragment analysis, however, paired modal CAG counts were significantly positively correlated. While EMQN error limits may be appropriate for typical expanded allele repeat lengths, repeats longer than 100 CAGs are not routinely observed in clinic and measures are likely to vary more than typical expanded alleles. Despite this, PacBio modal CAG measures were consistently between 7-12 CAGs higher than their fragment analysis counterparts suggesting a systematic difference is present between the two repeat counting methods. This is discussed further at the end of this chapter.

The difference in mean expansion indices between PacBio and fragment analysis was less than 1 for both the WT and expanded alleles. Fragment analysis WT expansion index was 0 for all 6 samples, hence no correlation test was conducted. PacBio and fragment analysis expanded allele expansion indices were not significantly positively correlated. This runs against the prediction that the two methods will give comparable measures of somatic expansion. Because these are cross-sectional samples, expansion indices have a low spread (SDs are less than 1 in both methods), which will contribute to the lack of correlation. This subject is discussed further at the end of this chapter.

Further comparisons were made between PacBio sequencing data of 109NI iPSCs and equivalent fragment analysis traces. As shown in Table 3.3, Library 3000-LBC-PBMC-iPSC contained all samples present in 3000-iPSC, with the addition of 6 others derived from 109N1, 3 of which were from lines in which the expanded allele had been edited to a WT allele length of 22 CAGs (Donaldson 2019). 3000-LBC-PBMC-iPSC also contained 28 samples from human PBMC/LBCs, which were analysed for the LBC-PBMC comparison in section 3.3.1.5. Table 3.12 summarises the comparison of the iPSC data in 3000-LBC-PBMC-iPSC to equivalent fragment analysis data. 3 of the samples are from lines which effectively have two WT alleles. These were excluded from expanded allele calculations.

Comparison	**Modal CAG**		**Expansion Index**	
Allele	WT	Expanded	WT	Expanded
N	11†	8††	11†	8††
Mean FA	18.8	126	0	1.68
Mean PB	19.3	133	0.17	1.32
SD FA	0.852	5.24	0	0.914
SD PB	0.452	5.96	0.105	0.694
Normal	No	Yes	No	Yes
R^2 or r_s	-0.358	0.917	N/A	0.313
p-value	0.280	$1.35x10^{-3}$	N/A	0.450

Table 3.12. Comparison of fragment analysis and PacBio-RD calls from library 3000-LBC-PBMC-iPSC of the *HTT* repeat locus in 109NI iPSC samples. N represents the number of samples analysed**.** FA: fragment analysis, PB: PacBio-RD, WT: wild type. † The FA trace of sample N15-P4 had no clear signal. †† 3 further samples were from lines with two WT alleles, these were excluded from expanded allele calculations. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. r_s is the Spearman's rank correlation coefficient. R^2 or r_s: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. The mean number PacBio reads with the modal CAG was 978 for the WT allele, and 38.5 for the expanded allele.

The difference in mean modal CAG between PacBio and fragment analysis was less than 1 for the WT allele, however, paired modal CAG counts were not significantly positively correlated, which runs counter to the prediction that the two methods will give the same repeat lengths. This may be due to the relatively low spread in the WT alleles: fragment analysis counts had a range of 2, while PacBio counts a range of 1. The mean difference between counting methods was 0.75 CAGs.

Also counter to the prediction that PacBio and fragment analysis will give the same repeat length, the difference in expanded allele mean modal CAG was more than 3 – the mean modal CAG was 8.0 higher for PacBio than fragment analysis. See error limits discussion below Table 3.11. Despite this, paired modal CAG counts were significantly positively correlated and PacBio measures were consistently between 6-12 CAGs higher than their fragment analysis counterparts, further suggesting a systematic difference is present between the two repeat counting methods. This is discussed further at the end of this chapter.

The difference in mean expansion indices between PacBio and fragment analysis was less than 1 for both the WT and expanded alleles. Fragment analysis WT expansion index was 0 for all 6 samples, hence no correlation test was conducted. PacBio and fragment analysis expanded allele expansion indices were not significantly positively

correlated. This runs against the prediction that the two methods will give comparable measures of somatic expansion. Because these are cross-sectional samples, expansion indices have a low spread (SDs are less than 1 in both methods), which will contribute to the lack of correlation. This subject is discussed further at the end of this chapter.

3.3.2.5. Comparison of different read depth sequencing

A comparison of PacBio data at different read depths was conducted. Library 3000-iPSC is comprised of 6 samples and a mean number of 5555 modal CAG reads for the WT allele and 175 reads for the expanded allele and is considered as the high-depth data in this comparison. Library 3000-LBC-PBMC-iPSC shares the same 6 samples as 3000-iPSC but 34 additional samples and is thus of lower depth. Of the 6 shared samples, the mean number of 3000-LBC-PBMC-iPSC reads with the modal CAG is 969 for the WT allele and 30.5 for the expanded allele. A summary of the comparison is shown in Table 3.13.

In comparing high and low depth PacBio-RD calls, I wanted to test the prediction that PacBio-RD will give the same repeat length at different read depths on 130 CAG cell model repeats. If so, the mean modal CAG repeat sizes from each dataset will differ by no more than 1 for the WT allele, no more than 3 for the expanded allele (see guidelines on error limits above Table 3.11) and paired modal CAG values will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that PacBio-RD will give comparable measures of somatic expansion at different read depths. If so, mean expansion indices will differ by no more than 1 and paired expansion indices will be significantly positively correlated ($\alpha = 0.05$).

Comparison	Modal CAG		Expansion Index	
Allele	WT	Expanded	WT	Expanded
N	6	6	6	6
Mean 3000-iPSC	19.2	134	0.121	1.11
Mean 3000-LBC-PBMC-iPSC	19.2	134	0.130	1.15
SD 3000-iPSC	0.408	6.22	$3.84x10^{-3}$	0.512
SD 3000-LBC-PBMC-iPSC	0.408	5.54	$2.73x10^{-3}$	0.539
Normal	No	Yes	Yes	Yes
R^2 or r_s	1.00	0.941	0.331	-0.220
p-value	0	$5.21x10^{-3}$	0.521	0.675

Table 3.13. Comparison of PacBio-RD calls of the *HTT* repeat locus in 109NI iPSC samples from libraries 3000-iPSC and 3000-LBC-PBMC-iPSC. N represents the number of samples analysed. FA: fragment analysis, PB: PacBio-RD, WT: wild type. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. r_s is the Spearman's rank correlation coefficient. R^2 or r_s: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. The mean number 3000-iPSC reads with the modal CAG was 5555 for the WT allele, and 175 for the expanded allele. The mean number 3000-LBC-PBMC-iPSC reads with the modal CAG was 969 for the WT allele, and 30.5 for the expanded allele.

The difference in mean modal CAG between datasets was less than 1 for the WT allele and paired modal CAG counts were significantly positively correlated, in line with the prediction that different read depths will give the same repeat lengths. The difference in expanded allele mean modal CAG was less than 3 and paired modal CAG counts were significantly positively correlated, which is also in line with the prediction that different read depths will give the same repeat length.

The difference in mean expansion indices between datasets was less than 1 for both the WT and expanded alleles. Despite this, expansion indices were not significantly positively correlated in either the WT or expanded alleles. This goes against the prediction that different read depths will give comparable measures of somatic expansion. Because these are cross-sectional samples, expansion indices have a low spread (SDs are less than 1 in both datasets), which will contribute to the lack of correlation. This subject is discussed further at the end of this chapter.

3.3.3. Sequencing Shorter Amplicons to Generate Increased Expanded Allele Read Depth

One of the aims of this project was to establish whether PacBio sequencing data could be used to reliably quantify the expanded *HTT* CAG repeats in 109NI iPSC models. Considering the relatively low read depth in the foregoing experiments using a 3000 bp amplicon, I decided to investigate whether a shorter, 600 bp amplicon, combined with an enrichment step for the expanded allele would improve read depth of the expanded allele.

3.3.3.1. Library Preparation

4 sets of primers were designed to generate 109NI expanded allele amplicons approximately 600 bp in length (Table 2.4). This length is easier to amplify as it is shorter, produces CCS reads with higher quality (more passes) and therefore more HiFi reads per sample. Also, the difference between a 250 bp WT allele and a 600 bp expanded allele enables efficient size selection using paramagnetic beads, unlike 3 kbp and a 3.3 kbp amplicons (https://core-genomics.blogspot.com/2012/04/how-do-spri-beads-work.html). This property of paramagnetic beads enables physical enrichment of the expanded allele which results in a higher proportion of expanded allele reads and thus greater expanded allele read depth.

Figure 3.17 shows PCR products of primers ANT1/2, ANT3/4, ANT5/6 and ANT7/8 used with 109NI iPSC gDNA as a template, specifically 11N11, a *FAN1* KO line with WT and expanded *HTT* alleles containing a pure repeat of 20 and 140 CAGs respectively, as measured by fragment analysis. Expected product sizes are shown in Table 3.14.

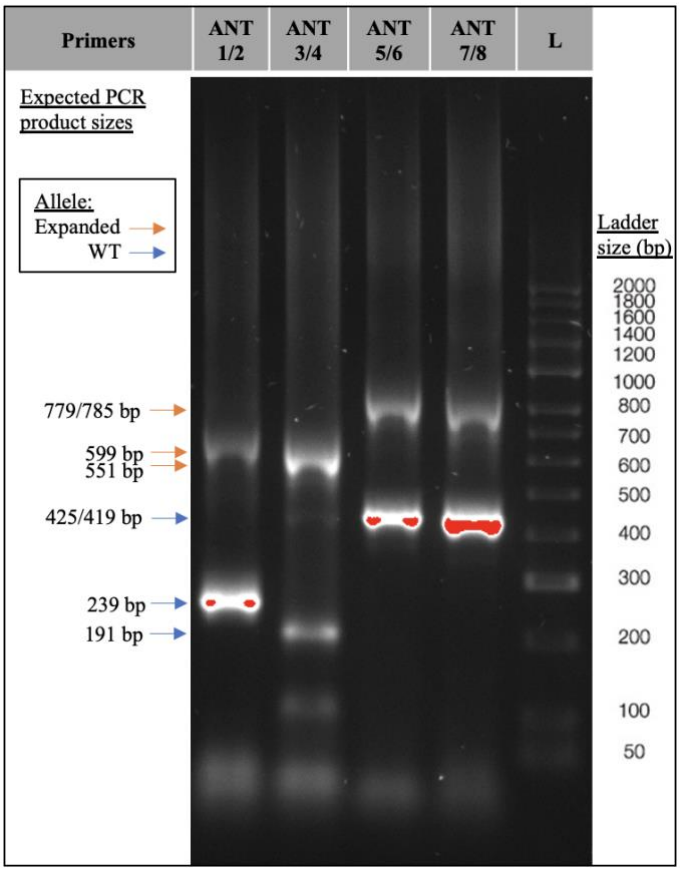


Figure 3.17. Gel electropherogram showing 600 bp PCR amplicons of the *HTT* locus. Four primer pairs were tested on 109NI iPSC (11N11) gDNA. L: Hyperladder II. bp: base pairs. WT: wild type. PCR run for 30 cycles, annealing at 61.7 °C, 25 ng of template DNA per 10 ul reaction. 1% Agarose-TBE gel run at 100V for 60 minutes.

Amplicon size (bp)	**Allele**	
Primers	**WT**	**Expanded**
ANT1/2	239	599
ANT3/4	191	551
ANT5/6	425	785
ANT7/8	419	779

Table 3.14. Expected PCR product sizes for ANT primers used on 109NI iPSC (11N11) DNA. bp: base pairs. WT: wild type.

See methods section 2.5.1.3 for full PCR protocol details. All PCR products were of the expected size (Figure 3.17). Primer pairs ANT1/2 and ANT3/4 were taken forward for PCR optimisation due to the relatively short WT allele compared to ANT5/6 and ANT7/8, which gave a greater difference in size between alleles for bead-based size selection.

Further PCR trials were conducted to determine which primers to use, what bead concentration to use for enriching the expanded allele, and to determine the optimal annealing temperature and the minimum number of PCR cycles required in the first

round of PCR. Reducing the number of PCR cycles should minimise the amount of PCR bias in the resulting data.

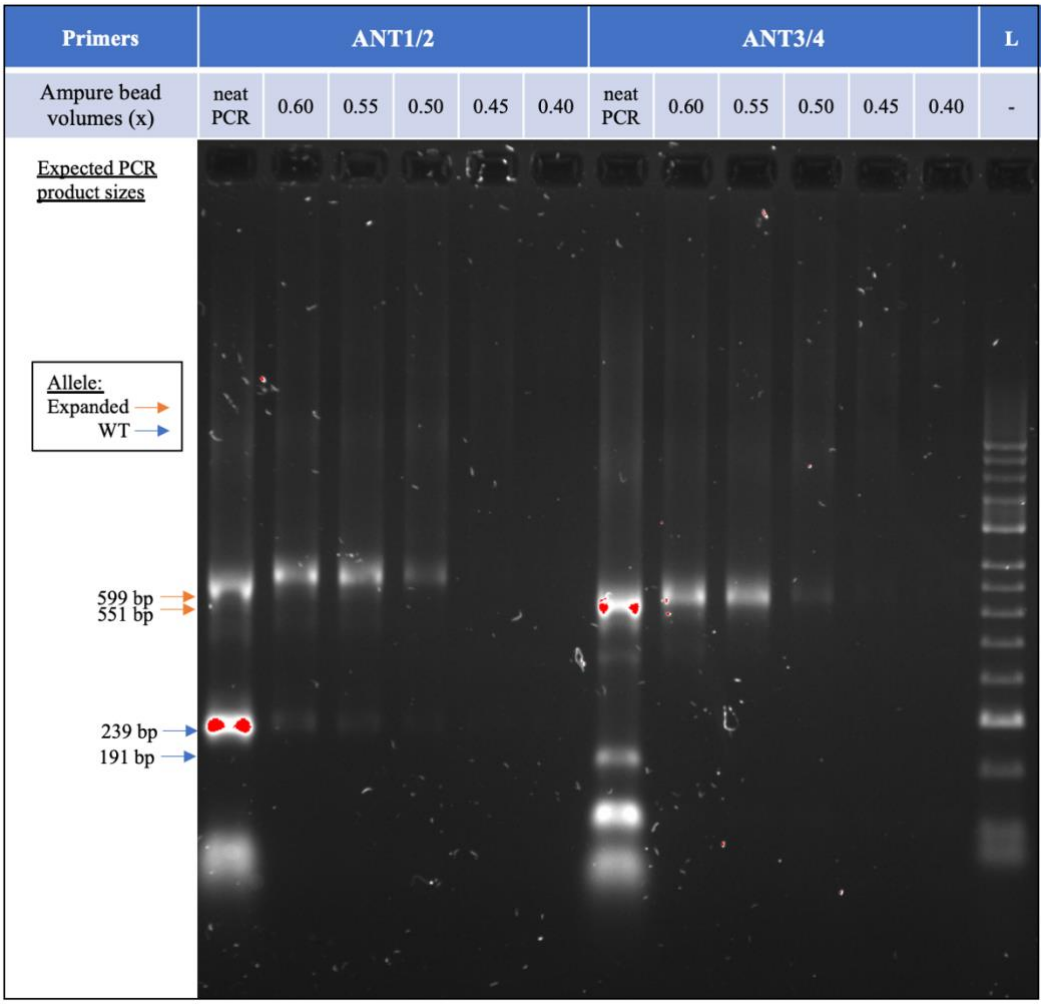


Figure 3.18. Gel electropherogram of Ampure PB bead purification of 600 bp amplicons of the *HTT* locus at a range of bead concentrations. Two primer pairs were tested on 109NI iPSC (11N11) gDNA. L: Hyperladder II. bp: base pairs. WT: wild type. PCR run for 30 cycles, annealing at 61.7 °C, 25 ng of template DNA per 10 ul reaction. 1% Agarose-TBE gel run at 100V for 60 minutes.

As show in Figure 3.18, purification with Ampure PB beads (method section 2.5.1.5) results in the removal of most of the WT allele band. The beads are coated with carboxyl molecules that reversibly bind DNA in the presence of a "crowding agent", polyethylene glycol (PEG), and salt. The concentration of PEG determines the size of DNA that can bind to the beads meaning the exact ratio of DNA to beads used is critical. A ratio of 0.60 volumes of beads to PCR products was the most effective concentration tested in terms of retaining the expanded allele, while removing the WT allele (tested: 0.60, 0.55, 0.50, 0.45, 0.40).

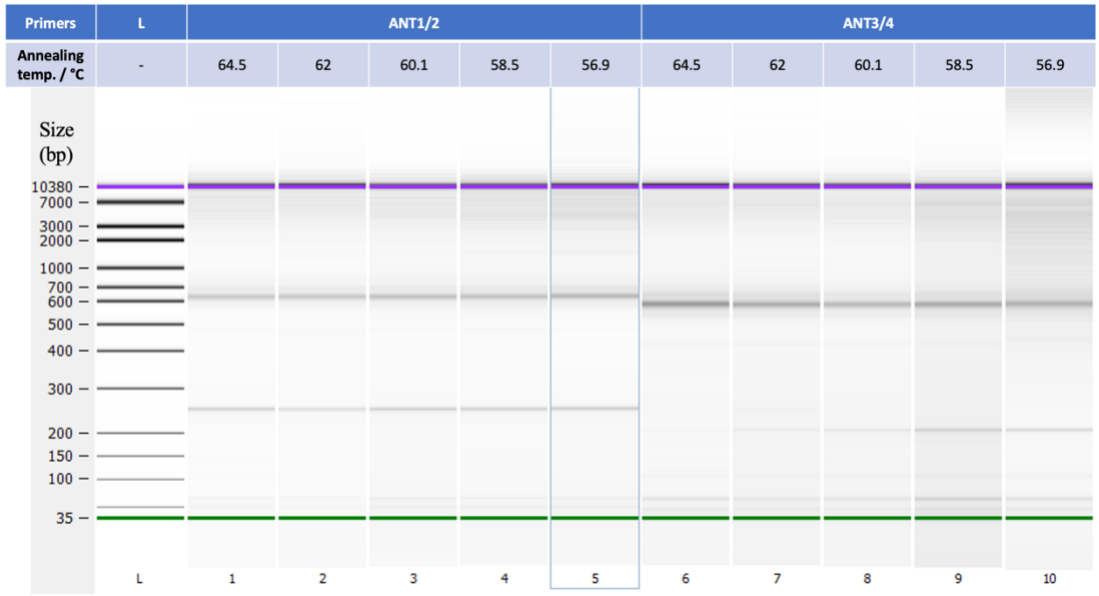


Figure 3.19. Capillary electropherogram of amplification of the *HTT* locus at a range of annealing temperatures. Bioanalyzer High Sensitivity kit.

As shown in Figure 3.19, amplification of the *HTT* locus is effective at a range of annealing temperatures. 62 °C was chosen as it gave clean expanded allele products in both ANT1/2 and ANT3/4 primers (tested: 64.5, 62, 60.1, 58.5 and 56.9 °C). Primers ANT1/2 were chosen as they had slightly cleaner expanded allele bands than ANT3/4.

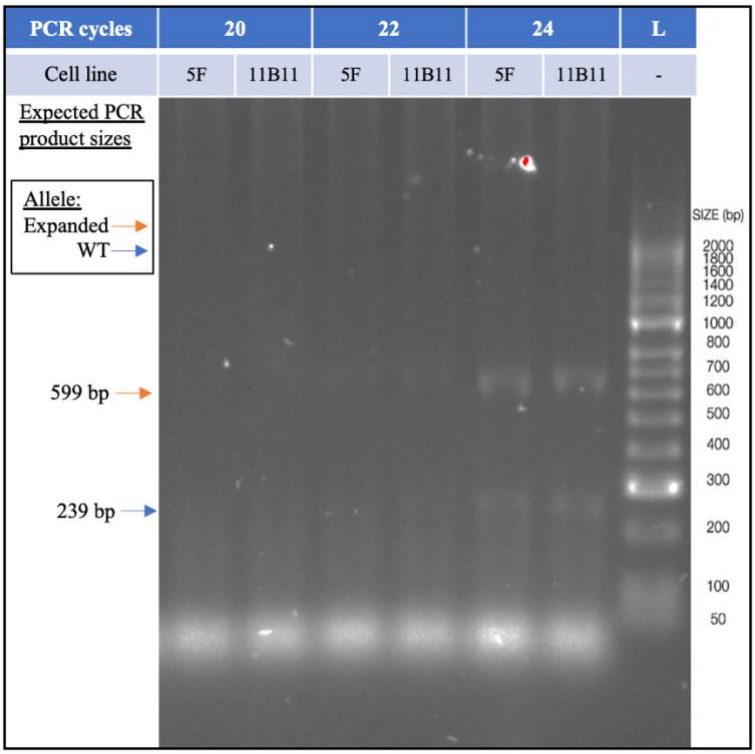


Figure 3.20. Gel electropherogram of 600 bp PCR amplicons of the *HTT* locus across a range of PCR cycle numbers. ANT1/2 primers amplifying 109NI-5F and 109-11B11 iPSC gDNA. L: Hyperladder II. bp: base pairs. WT: wild type. Annealing at 62°C, 25 ng of template DNA per 10 ul reaction. 1% Agarose-TBE gel run at 100V for 60 minutes.

Figure 3.20 shows the amplification of 109NI-5F and 109NI-11B11 DNA at 3 different PCR cycle numbers. The minimum number of first round PCR cycles deemed viable was 24, as PCR products were too dilute to visualise at 20 and 22 cycles.

Figure 3.21 shows the degree of enrichment of the expanded allele when purifying PCR products with 0.6x volumes of Ampure PB beads. Before purification the molarity of the WT and expanded alleles was 305 and 31.7 pmol/L respectively, giving a ratio of 9.61 WT alleles to every expanded allele. After purification the molarity of the WT and expanded alleles was 141 to 246 pmol/L, giving a ratio of 0.573 WT alleles to every expanded allele. This represents an expanded allele enrichment factor of 16.8.

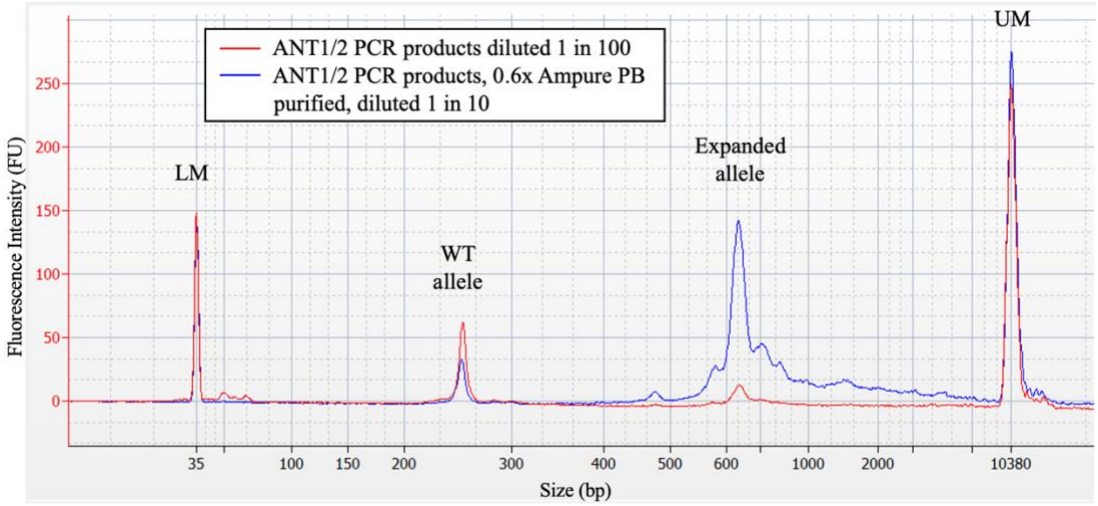


Figure 3.21. Capillary electrophoresis trace showing 109NI-5F *HTT* expanded allele enrichment using Ampure PB beads. WT allele at 254 bp. Expanded allele at 634 bp. LM: lower marker, UM: upper marker, WT: wild type, bp: base pairs.

Subsequent library preparation was conducted in the same way as for the 3 kbp libraries, except for DNA purification after the second round of PCR, for which 0.6x volumes of Ampure PB beads were used as opposed to the 1.0x volumes previously, introducing a second size selection step into the library preparation. Figure 3.22 shows capillary electrophoresis traces of the pooled 600-iPSC-1, 600-iPSC-2 and 600-iPSC-3 libraries.

Sample details of all three 600-iPSC libraries are shown in Table 3.18. 600-iPSC-1 is a 600 bp equivalent to 3000-iPSC, which allowed a direct comparison between the two amplicon lengths (comparison results in section 3.3.3.2) and comparison to fragment analysis data (section 3.3.3.3). Further analysis appears in section 3.3.3.4.1. 600-iPSC-2 is comprised of 12 samples, split into two experiments of 6 samples each. The results of the first experiment, looking at the effect of PCR duplicates on CAG repeats are summarised in section 3.3.3.4.2. The results of the second experiment, looking at the effect of PCR cycle number on CAG repeats are summarised in section 3.3.3.4.3. 600-iPSC-3 consisted of 12 samples that comprise one experiment designed to determine the effect of iPSC maturity and *FAN1 genotype* on modal CAG and repeat expansion. The results of this experiment are summarised in section 4.3.1.1. As the samples in 600-iPSC-3 have previously been analysed by fragment analysis an additional validatory comparison was possible (see section 3.3.3.3).

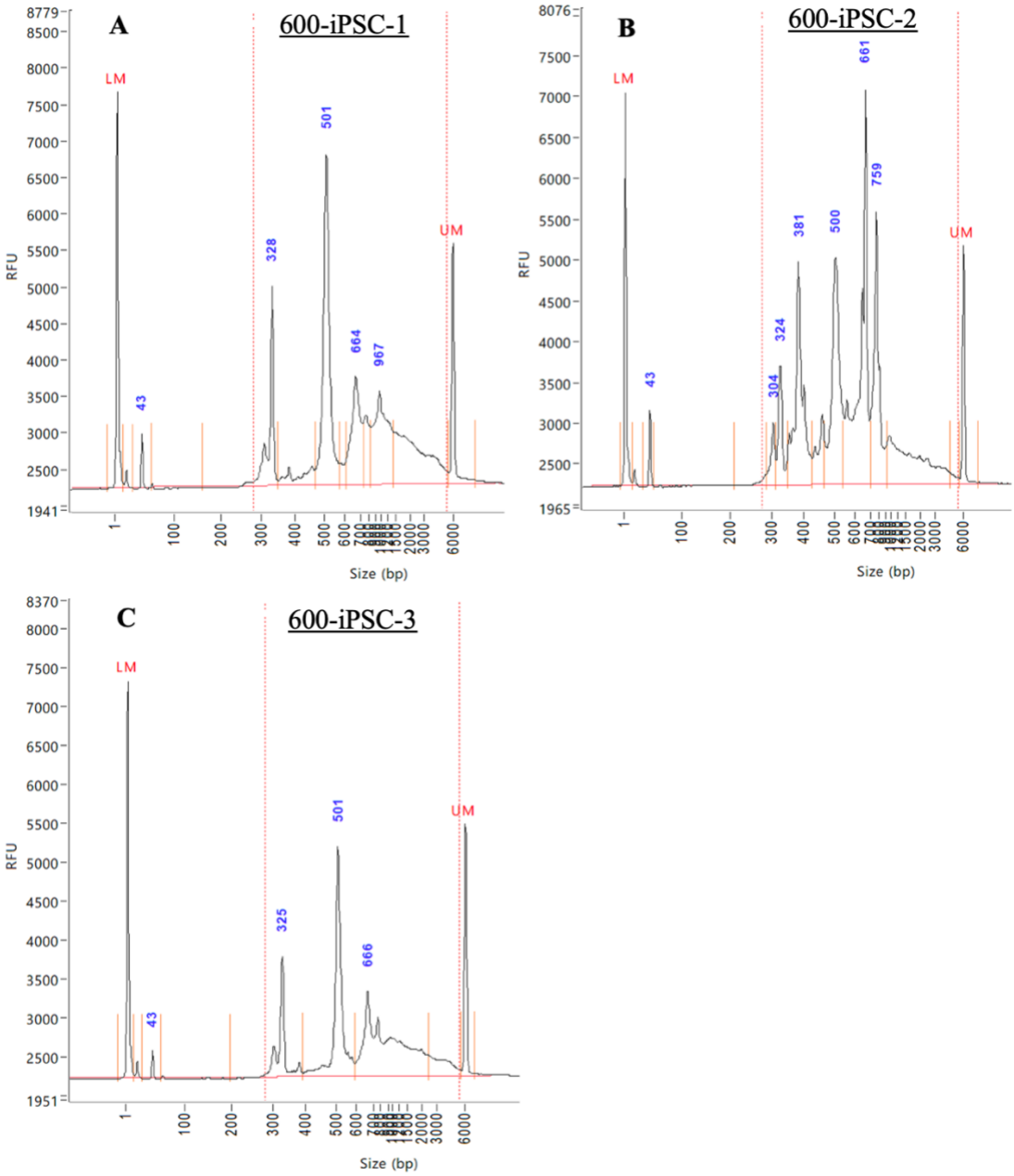


Figure 3.22. Capillary electrophoresis traces of pooled 600 bp *HTT* SMRTbell libraries, 600-iPSC-1, 600-iPSC-2 and 600-iPSC-3. RFU: relative fluorescence intensity, bp: base pairs, LM: lower marker, UM: upper marker. Blue numbers indicate the size of peaks in bp.

As shown in Figure 3.22, library 600-iPSC-1 contains peaks at 328 and 501 bp, which correspond to the WT and expanded alleles of 109NI samples respectively. Additional peaks at 664 and 967 likely represent PCR chimera in the sample. Library 600-iPSC-2 contains WT peaks at 304, 324. The peak at 381 is approximately 60 bases longer than the main WT peak and is therefore likely to be expanded alleles from the six HD patient PBMC samples present in the library. The peak at 500 corresponds to expanded alleles from the six 109NI samples, while peaks at 661 and 759 are probably PCR chimera. Library 600-iPSC-3 contains peaks at 325 and 501 bp, which correspond to

the WT and expanded alleles of 109NI samples respectively. Additional peaks at 666 and above likely represent PCR chimera in the sample.

3.3.3.2. Comparing 3 kbp and 600 bp Sequence Data

Once I had optimised the 600 bp amplicon methods, I generated data for some of the samples previously analysed using the 3000 bp amplicon method to assess the characteristics of the different methods. 3000-iPSC and 600-iPSC-1 were generated from the same six 109N1 iPSC samples, however, 600-iPSC-1 is a 600 bp library which underwent two 0.6x Ampure purifications, whereas 3000-iPSC is a 3 kbp library which underwent two 1.0x Ampure purifications. 0.6x Ampure purification has been shown to physically enrich for the expanded allele (Figure 3.21), thus I would expect the proportion of PacBio reads expanded alleles to be higher in 600-iPSC-1 than in 3000-iPSC. As shown in Table 3.2, 91,340 WT and 13,026 expanded allele 3000-iPSC reads survived filtering, giving a ratio of 7.01 WT:expanded. 30,849 WT and 108,889 expanded allele 600-iPSC-1 reads survived filtering, giving a ratio of 0.28 WT: expanded. This represents an expanded allele enrichment factor of 24.8. A large majority of reads, 77.9%, are now expanded alleles. A similar proportion is seen in the other library comprising only 109N1 iPSC samples, 600-iPSC-3, where 97,900 reads out of the 127,652 (76.7%) are classified as expanded.

Another consequence of the shorter amplicon length is that the median read quality has improved from Q26 (3000-iPSC), or approximately 25 errors per 1000 bases, to Q40 (600-iPSC-1), approximately 1 error per 1000 bases. Figures 3.15A and 3.23A show the distribution of quality scores in libraries 3000-iPSC and 600-iPSC-1 respectively. Q50 is the maximum average Phred quality score assigned in Figure 3.23A (1 error per 10,000 bases), hence the pileup (Phred quality scores explained in section 1.6.2.). 89% of reads in 600-iPSC-1 are ≥ Q20 (HiFi reads), where just over half were HiFi reads in 3000-iPSC.

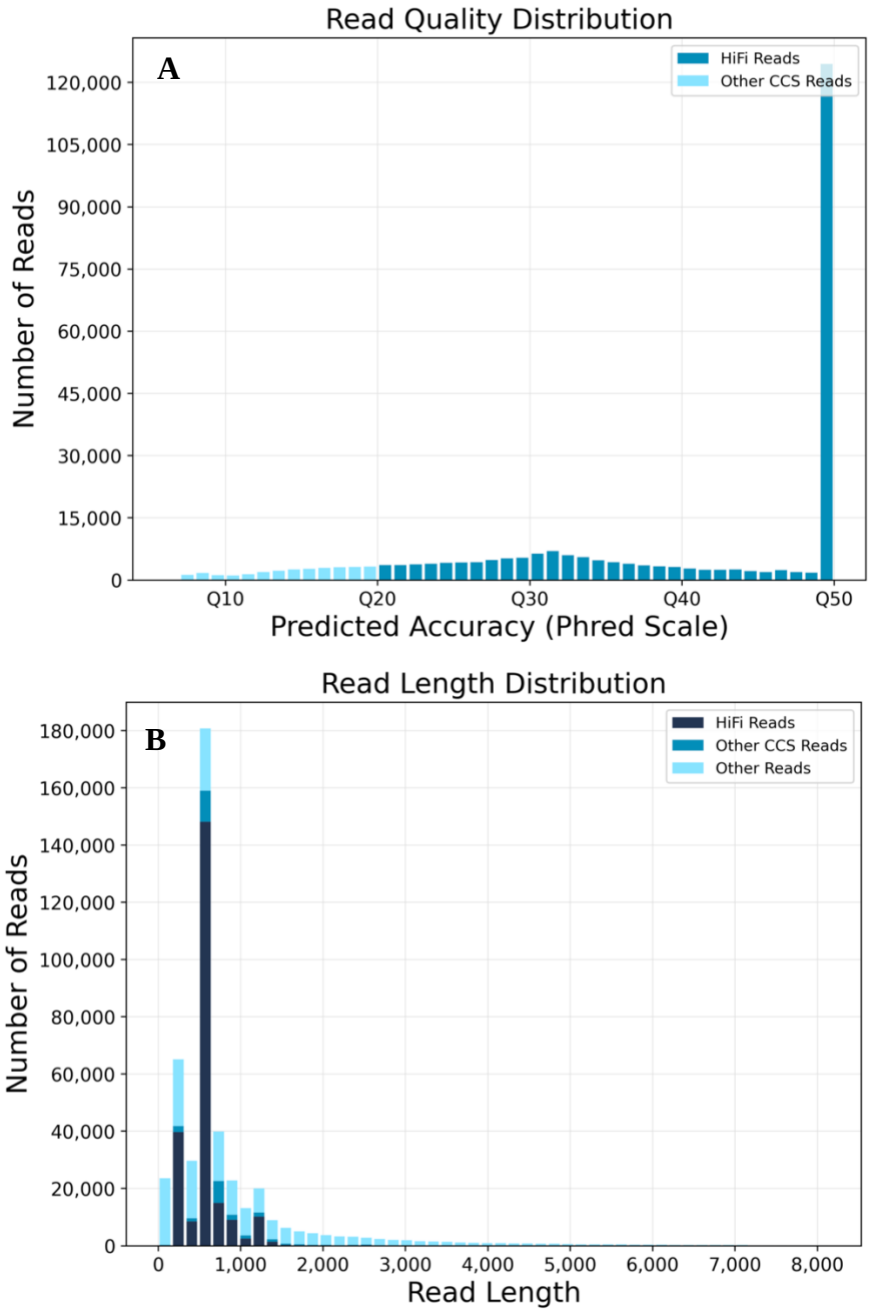


Figure 3.23. Read quality and length distributions for all reads in 600 bp iPSC library 600-iPSC-1. Q: Phred quality score. HiFi reads (Q20 or higher) shown in green, non-HiFi reads in grey. (A) Read quality distribution. (B) Read length distribution. bp: base pairs.

The distribution of read lengths of 600-iPSC-1, shown in Figure 3.23B is broadly comparable to the capillary electrophoresis trace shown in Figure 3.22A.

3000-iPSC and 600-iPSC-1 consisted of the same samples and were analysed in the same way, meaning a further comparison of summary metrics derived from the sequencing data could be made. The six samples of 600-iPSC-1 had a mean number of reads with the modal CAG of 2,289 for the WT allele and 1,292 for the expanded allele. The six equivalent samples of 3000-iPSC had a mean number of reads with the

modal CAG of 5,555 for the WT allele and 175 for the expanded allele. A comparison of modal CAG and expansion indices is shown in Table 3.15.

In comparing PacBio-RD calls of 3000 bp and 600 bp data, I wanted to test the prediction that the two amplicon lengths will give the same repeat lengths. If so, the mean modal CAG repeat sizes from each dataset will differ by no more than 1 for the WT allele, no more than 3 for the expanded allele (see guidelines on error limits above Table 3.11) and paired modal CAG values will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that PacBio-RD will give comparable measures of somatic expansion at different amplicon lengths. If so, mean expansion indices will differ by no more than 1 and paired expansion indices will be significantly positively correlated ($\alpha = 0.05$).

Comparison	**Modal CAG**		**Expansion Index**	
Allele	WT	Expanded	WT	Expanded
N	6	6	6	6
Mean 3000-iPSC	19.2	134	0.121	1.11
Mean 600-iPSC-1	19.2	130	0.058	1.33
SD 3000-iPSC	0.408	6.22	3.84×10^{-3}	0.512
SD 600-iPSC-1	0.408	4.88	3.29×10^{-3}	0.558
Normal	No	Yes	Yes	Yes
R^2 or r_s	1.00	0.960	-0.359	-0.048
p-value	0	2.42×10^{-3}	0.485	0.928

Table 3.15. Comparison of PacBio-RD calls of the *HTT* repeat locus in 109NI iPSC samples from libraries 3000-iPSC and 600-iPSC-1. N represents the number of samples analysed. FA: fragment analysis, PB: PacBio-RD, WT: wild type. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. r_s is the Spearman's rank correlation coefficient. R^2 or r_s: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. The mean number 3000-iPSC reads with the modal CAG was 5555 for the WT allele, and 175 for the expanded allele. The mean number 600-iPSC-1 reads with the modal CAG was 2,289 for the WT allele, and 1,292 for the expanded allele.

The difference in mean modal CAG between datasets was less than 1 for the WT allele and paired modal CAG counts were significantly positively correlated, in line with the prediction that different amplicon lengths will give the same repeat lengths. At 3.3, the difference in expanded allele mean modal CAG was more than 3, and thus runs counter to the prediction that different amplicon lengths will give the same repeat lengths. While a difference of 3 is a conservative limit, this is unexpected given that

PacBio-RD counts on data of differing read depths produce the same repeat lengths (see Table 3.13) and may be related to the introduction of size selection steps for the 600 bp library. Despite this, paired modal CAG counts were significantly positively correlated, which is in line with the prediction that different amplicon lengths will give the same repeat length.

The difference in mean expansion indices between datasets was less than 1 for both the WT and expanded alleles. Despite this, expansion indices were not significantly positively correlated in either the WT or expanded alleles. This goes against the prediction that different amplicon lengths will give comparable measures of somatic expansion. Because these are cross-sectional samples, expansion indices have a low spread (SDs are less than 1 in both datasets), which will contribute to the lack of correlation. This subject is discussed further at the end of this chapter.

3.3.3.3. Comparison of 600 bp iPSC sequencing CAG counts to fragment analysis

The same 6 samples sequenced in library 3000-iPSC comprise library 600-iPSC-1, however the amplicon length is shorter at 600 bp compared to 3 kbp. 600-iPSC-3 is a 600 bp amplicon library of 12 samples including two isogenic 109NI lines (*FAN1*$^{+/+}$ and *FAN1*$^{+/+}$) harvested at 3 different passages amplified in duplicate.

Table 3.16 shows a comparison of 600-iPSC-1 to fragment analysis data of the same samples. In comparing PacBio-RD calls to fragment analysis data, I wanted to further test the prediction that PacBio-RD will give the same repeat length as fragment analysis on 130 CAG cell model repeats. If so, the mean modal CAG repeat sizes from each method will differ by no more than 1 for the WT allele, no more than 3 for the expanded allele (see error limits discussion above Table 3.11) and paired modal CAG values will be significantly positively correlated ($\alpha = 0.05$). I also wanted to test the prediction that PacBio-RD will give comparable measures of somatic expansion as fragment analysis. If so, mean expansion indices from each method will differ by no more than 1 and paired expansion indices will be significantly positively correlated ($\alpha = 0.05$).

Comparison	Modal CAG		Expansion Index	
Allele	WT	Expanded	WT	Expanded
N	6	6	6	6
Mean FA	18.8	125	0	1.64
Mean PB	19.2	130	0.06	1.33
SD FA	0.839	5.56	0	0.773
SD PB	0.408	4.88	$3.29x10^{-3}$	0.558
Normal	No	Yes	No	Yes
R^2 or r_s	-0.655	0.957	N/A	0.624
p-value	0.158	$2.79x10^{-3}$	N/A	0.185

Table 3.16. Comparison of fragment analysis and PacBio-RD calls from library 600-iPSC-1 of the *HTT* repeat locus in 109NI iPSC samples. N represents the number of samples analysed. FA: fragment analysis, PB: PacBio-RD, WT: wild type. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. rs is the Spearman's rank correlation coefficient. R^2 or r_s: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. The mean number PacBio reads with the modal CAG was 2,289 for the WT allele, and 1,292 for the expanded allele.

The difference in mean modal CAG between PacBio and fragment analysis was less than 1 for the WT allele, however, paired modal CAG counts were not significantly positively correlated, which runs counter to the prediction that the two methods will give the same repeat lengths. This may be due to the relatively low spread in the WT alleles: fragment analysis counts had a range of 2, while PacBio counts a range of 1. The mean difference between counting methods was 0.64 CAGs.

Also counter to the prediction that PacBio and fragment analysis will give the same repeat length, the difference in expanded allele mean modal CAG was more than 3 – the mean modal CAG was 5.5 CAGs higher for PacBio than fragment analysis, however, paired modal CAG counts were significantly positively correlated. While EMQN error limits may be appropriate for typical expanded allele repeat lengths, repeats longer than 100 CAGs are not routinely observed in clinic and measures are likely to vary more than typical expanded alleles. Despite this, PacBio modal CAG measures were consistently between 3-8 CAGs higher than their fragment analysis counterparts, further suggesting a systematic difference is present between the two repeat counting methods. This is discussed further at the end of this chapter.

The difference in mean expansion indices between PacBio and fragment analysis was less than 1 for both the WT and expanded alleles. Fragment analysis WT expansion index was 0 for all 6 samples, hence no correlation test was conducted. PacBio and

fragment analysis expanded allele expansion indices were not significantly positively correlated. This runs against the prediction that the two methods will give comparable measures of somatic expansion. Because these are cross-sectional samples, expansion indices have a low spread (SDs are less than 1 in both methods), which will contribute to the lack of correlation. This subject is discussed further at the end of this chapter.

Comparison	**Modal CAG**		**Expansion Index^**	
Allele	WT	Expanded	WT	Expanded
N	12	12	12	12
Mean FA	18.4	126	0	3.56
Mean PB	19.5	132	0.0627	3.20
SD FA	0.943	2.97	0	2.79
SD PB	0.522	3.15	5.08×10^{-3}	2.32
Normal	No	No	No	No
R^2 or r_s	-0.655	0.957	N/A	0.937
p-value	0.202	1.14×10^{-4}	N/A	7.22×10^{-6}

Table 3.17. Comparison of fragment analysis and PacBio-RD calls from library 600-iPSC-3 of the *HTT* repeat locus in 109NI iPSC samples. N represents the number of samples analysed. FA: fragment analysis, PB: PacBio-RD, WT: wild type. ^ passage 4-anchored expansion indices. SD: standard deviation. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. R^2 is the Pearson correlation coefficient squared. rs is the Spearman's rank correlation coefficient. R^2 or r_s: where data is Normal, Pearson coefficient is used, otherwise Spearman coefficient is used. p-values derived from a 2-tailed t-test of the correlation coefficient. The mean number PacBio reads with the modal CAG was 978 for the WT allele, and 38.5 for the expanded allele.

Table 3.17 shows a comparison of 600-iPSC-3 to fragment analysis data of the same samples. The difference in mean modal CAG between PacBio and fragment analysis was more than 1 for the WT allele and paired modal CAG counts were not significantly positively correlated, which runs counter to the prediction that the two methods will give the same repeat lengths. The difference in modal CAG is unexpected given previous comparisons have been within 1 and is driven by unusually short measures in four fragment analysis samples. The lack of correlation may be due to this and to the relatively low spread in the WT alleles: fragment analysis counts had a range of 2, while PacBio counts a range of 1. The mean difference between counting methods was 0.64 CAGs.

Also counter to the prediction that PacBio and fragment analysis will give the same repeat length, the difference in expanded allele mean modal CAG was more than 3 – the mean modal CAG was 5.6 CAGs higher for PacBio than fragment analysis. Despite this, paired modal CAG counts were significantly positively correlated and

were consistently between 4-8 CAGs higher than their fragment analysis counterparts, further suggesting a systematic difference is present between the two repeat counting methods. This is discussed further at the end of this chapter.

The difference in mean expansion indices between PacBio and fragment analysis was less than 1 for both the WT and expanded alleles. Fragment analysis WT expansion index was 0 for all 12 samples, hence no correlation test was conducted. PacBio and fragment analysis expanded allele expansion indices were significantly positively correlated, which is in line with the prediction that the two methods will give comparable measures of somatic expansion.

In contrast to the cross-sectional iPSC samples compared previously (see Tables 3.11, 3.12, 3.13, 3.15 and 3.16), this dataset is comprised of longitudinal samples which meant that longitudinal expansion indices could be calculated. These have a higher spread (SDs are more than 2 in both methods) than previous cross-sectional measures, which may explain why this expansion index correlation is significant where previous ones were not. Longitudinal measures of somatic expansion are calculated using modal CAG of the first time point in the series so if the first time point is passage 4, expansion indices will be referred to as a 'passage 4-anchored expansion indices' to highlight this fact.

3.3.3.4. Analysis of 600 bp amplicon data CAG repeats

Once I established that the 600 bp amplicon method gave an increased read depth of the expanded allele, CAG repeat data from the expanded alleles (>29 CAGs) of libraries 600-iPSC-1, 600-iPSC-2 and 600-iPSC-3 were analysed further. A summary of this analysis is shown in Table 3.18.

3.3.3.4.1. 600-iPSC-1

Starting with 600-iPSC-1, it was notable that one of the samples, 11B11-P36, had 1,364 expanded allele reads, more than 10 times fewer than any other. Prior to filtering, the number of expanded allele reads in this sample was 5144, over 4 times fewer than the mean of the other 5 samples (23,785), despite having the highest proportion of expanded reads (87.8%). The first filter applied (read possesses 12-bp flanks), removed very few expanded allele reads (30), however the second filter (read is non-chimeric) (Figure 2.1), removed 3,663 of the remaining 5114 (71.6%), leaving 1451, 87 more than the final count. Inspection of the sequence files confirmed that

most of the reads in this sample had a chimeric structure, although the reason is unclear as library preparation was conducted in the same way as the other samples, which had a range of 0.4 to 16.8% of reads removed by the chimeric read filter.

Library	Sample number	Cell line	Passage	Replicate	PCR cycles	Modal CAG	Expansion index	Contraction index	Expanded allele reads	reads < modal CAG	Modal peak reads	reads > modal CAG	reads > modal CAG +30	% reads < modal CAG	% modal peak reads	% reads > modal CAG	% reads > modal CAG +30
600-iPSC-1	1	11B11	36*	N/A	24	137	2.36	-2.52	1364	670	93	601	24	49.1	6.82	44.1	1.76
	2	5F	6	N/A	24	128	1.41	-1.56	31986	15125	2457	14404	627	47.3	7.68	45.0	1.96
	3	5F	33	N/A	24	133	1.03	-2.06	16962	9786	1131	6045	392	57.7	6.67	35.6	2.31
	4	109NI	31	N/A	24	123	0.75	-2.51	18094	11520	1260	5314	352	63.7	6.96	29.4	1.95
	5	109NI	46	N/A	24	132	1.09	-2.05	23389	12496	1694	9199	558	53.4	7.24	39.3	2.39
	6	SC5	36	N/A	24	128	1.35	-3.22	17094	9087	1114	6893	286	53.2	6.52	40.3	1.67
Library	Sample number	Cell line	Passage	Replicate	PCR cycles	Modal CAG	Expansion index	Contraction index	Expanded allele reads	reads < modal CAG	Modal peak reads	reads > modal CAG	reads > modal CAG +30	% reads < modal CAG	% modal peak reads	% reads > modal CAG	% reads > modal CAG +30
600-iPSC-2	1	L81 - LBC	early†	1	24	39	0.28	-1.09	9457	4780	2226	2451	0	50.5	23.5	25.9	0.00
	2	L81 - LBC	early†	2	24	39	0.26	-1.33	9566	5067	2117	2382	0	53.0	22.1	24.9	0.00
	3	E11 - LBC	early†	1	24	42	0.25	-0.99	6954	3704	1639	1611	1	53.3	23.6	23.2	0.01
	4	E11 - LBC	early†	2	24	42	0.22	-2.60	14243	8536	2873	2834	1	59.9	20.2	19.9	0.01
	5	E79 - LBC	early†	1	24	48	0.37	-1.11	10283	5669	1869	2745	8	55.1	18.2	26.7	0.08
	6	E79 - LBC	early†	2	24	48	0.32	-2.63	4845	2919	751	1175	1	60.2	15.5	24.3	0.02
	7	5F	20	N/A	20	131	0.53	-2.64	2129	1303	193	633	39	61.2	9.07	29.7	1.83
	8	11B11	20	N/A	20	134	1.63	-1.64	3612	1670	301	1641	68	46.2	8.33	45.4	1.88
	9	5F	20	N/A	22	131	0.70	-2.35	2105	1274	155	676	39	60.5	7.36	32.1	1.85
	10	11B11	20	N/A	22	134	1.85	-1.47	5527	2532	417	2578	118	45.8	7.54	46.6	2.13
	11	5F	20	N/A	24	131	0.69	-2.34	4734	2983	362	1389	98	63.0	7.65	29.3	2.07
	12	11B11	20	N/A	24	135	0.99	-1.73	1289	755	94	440	66	58.6	7.29	34.1	5.12
Library	Sample number	Cell line	Passage	Replicate	PCR cycles	Modal CAG	Expansion index^	Contraction index^	Expanded allele reads	reads < modal CAG	Modal peak reads	reads > modal CAG	reads > modal CAG +30	% reads < modal CAG	% modal peak reads	% reads > modal CAG	% reads > modal CAG +30
600-iPSC-3	1	5F	4	1	24	129	1.03	-2.61	8511	4906	654	2951	213	57.6	7.68	34.7	2.50
	2	5F	4	2	24	128	1.42	-1.89	10512	5202	859	4451	219	49.5	8.17	42.3	2.08
	3	5F	20	1	24	130	1.89	-1.81	8901	4731	640	3530	273	53.2	7.19	39.7	3.07
	4	5F	20	2	24	130	2.43	-1.42	4765	2614	330	1821	157	54.9	6.93	38.2	3.29
	5	5F	36*	1	24	131	2.52	-1.30	12232	6536	890	4806	245	53.4	7.28	39.3	2.00
	6	5F	36*	2	24	131	2.94	-1.10	9301	5006	754	3541	111	53.8	8.11	38.1	1.19
	7	11B11	4	1	24	130	0.96	-2.23	8105	4596	696	2813	195	56.7	8.59	34.7	2.41
	8	11B11	4	2	24	130	0.94	-2.77	9653	5689	787	3177	268	58.9	8.15	32.9	2.78
	9	11B11	20	1	24	136	5.01	-0.40	9876	5776	725	3375	253	58.5	7.34	34.2	2.56
	10	11B11	20	2	24	135	4.89	-0.60	6445	3499	454	2492	180	54.3	7.04	38.7	2.79
	11	11B11	36*	1	24	136	7.13	-2.63	3937	1977	186	1774	199	50.2	4.72	45.1	5.05
	12	11B11	36*	2	24	137	7.26	-0.96	5662	3052	296	2314	218	53.9	5.23	40.9	3.85

Table 3.18. CAG repeat summary statistics from expanded alleles (>29 CAGs) of PacBio libraries 600-iPSC-1, 600-iPSC-2 and 600-iPSC-3. 11B11: 109NI-11B11. 5F: 109NI-5F. SC5: 109NI-SC5. RD restrictive profile CAG counts. *Non-proliferating cells. † Exact passage number unknown. ^ passage 4-anchored. LBC: lymphoblastoid cell line. PCR cycles: First round PCR cycles shown only, an additional 20 cycles was performed on all samples in the second round of PCR.

Contraction indices are greater in magnitude than expansion indices in all 600-iPSC-1 samples, however most of these will be PCR artefacts. The percentages of reads with fewer CAGs than the modal CAG was higher than the percentages of reads with more CAGs than the modal CAG in all 600-iPSC-1 samples. Despite this there are 2,239 'greater than modal CAG plus 30' reads in 600-iPSC-1, with at least 1.67% in all samples. 10 'greater than modal CAG plus 30' reads were chosen at random for a visual sequence inspection. All reads had a typical FLANK-REPEAT-FLANK structure. 594 of the 108,889 expanded allele reads in 600-iPSC-1 had over 200 CAGs (0.55%), and 18 had over 300 CAGs (0.017%).

3.3.3.4.2. The effect of duplicate PCRs on long read sequencing CAG repeats

The first 6 samples of 600-iPSC-2 are 3 patient LBC samples amplified in duplicate. The purpose of this experiment was twofold – to determine the effect of duplicate PCRs on CAG repeats (i.e., how much variation exists between technical replicates of patient samples?) and to see if rare large repeat expansions exist in patient alleles at high depth.

The 6 LBC samples consisted of 3 patient lines. These lines were initially chosen based on their inclusion in the HD exome sequencing project (McAllister et al. 2022). For library 600-iPSC-2, I prepared 1 extreme late onset line, L81, and 2 extreme early onset lines, E11 and E79, in duplicate. See section 3.3.1.3.1 for definitions of late and early onset.

Modal CAG was identical between duplicates. Expansion index is also well matched between duplicates with the biggest difference 0.05 (E79). Contraction index is more variable between duplicates in all lines, with the biggest difference 1.59 (E11), and the smallest difference 0.22 (L81).

More than 50% of the reads in all samples had CAG counts below the modal CAG. Only the two early onset lines had any 'greater than modal CAG plus 30' reads, although 3 of the 4 samples had only 1 read in this category. E79 replicate 1 had 8 reads greater than the modal CAG +30, accounting for 73% of the reads in this category.

Library 600-iPSC-3 also contained PCR duplicates of multiple samples (6). Modal CAG was identical in 3 of these and in the remainder the difference between duplicates

was 1. Anchored expansion index also appears to be well correlated between duplicates, especially in the 11B11 samples (biggest difference 0.13).

Overall, the effect of PCR duplicates on modal CAG seems to be more pronounced with longer repeats but within +/- 1 CAG up to 130 CAGs. PCR appears to have a larger effect on contraction index than on expansion index. The effect of PCR duplicates on expansion index also seems to be more pronounced with longer repeats but within +/- 0.5 on samples up to 130 CAGs.

3.3.3.4.3. The effect of changing the PCR cycle number on long read sequencing

The second 6 samples in library 600-iPSC-2 consisted of an experiment to determine the effect of PCR cycle number on modal CAG, repeat expansion and contraction of 109NI expanded alleles. DNA from two cell lines (109NI-5F and 109NI-11B11) underwent first round amplification with 20, 22 and 24 cycles. All samples underwent 20 cycles in the second round of PCR, which translates to 40, 42 and 44 cycles in total. The PCR was conducted on passage 20 samples from two cell lines, the *FAN1*$^{+/+}$ line 11B11 and the *FAN1*$^{-/-}$ line 5F. There were no changes in modal CAG with increasing PCR cycles, except for the 11B11-24 cycle sample, which had an increase of 1. Figure 3.24 shows the percentage of reads across cell lines and PCR cycle numbers.

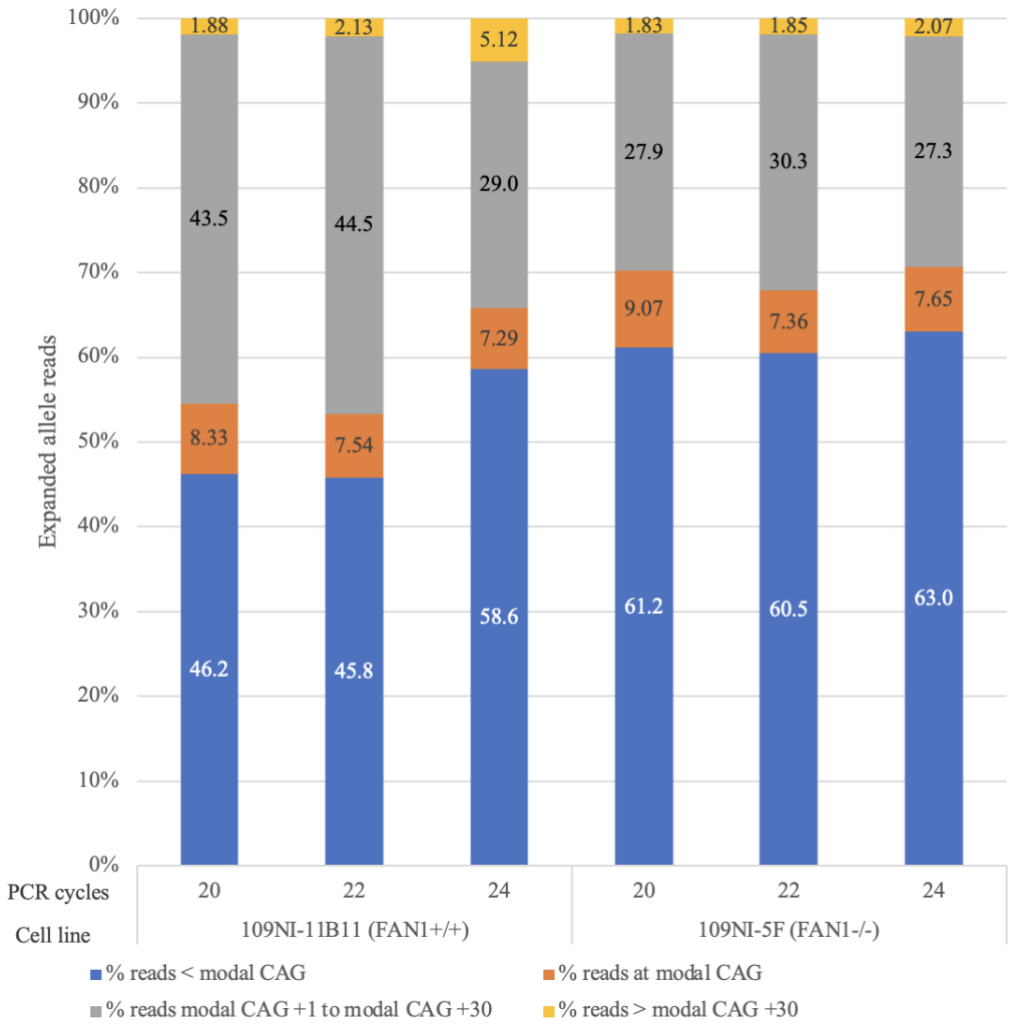


Figure 3.24. Percentage of expanded allele reads by *HTT* CAG repeat length category across cell line and number of first round PCR cycles. Library 600-iPSC-2 samples 7-12. PacBio-RD restrictive profile counts of filtered expanded allele (>29 CAGs) reads.

Few trends emerged in relation to expansion and contraction indices, however the percentage of reads with fewer CAGs than the modal CAG was markedly increased in the 11B11-24 cycle sample at 58.6%, compared with 46.2 and 45.8% in the 11B11-20 and 22 cycle samples respectively. A smaller increase is seen in the 5F-24 cycle sample at 63.0%, compared to 61.2 and 60.5% in the 5F-20 and 22 cycle samples respectively. The 11B11-24 cycle sample also showed a large increase in the percentage of 'greater than modal CAG +30' reads at 5.12%, compared to 1.88 and 2.13% for the 11B11-20 and 22 cycle samples respectively but a large decrease in percentage of reads greater than the modal CAG overall at 29.0%, compared to 43.5 and 44.5% for the 11B11-20 and 22 cycle samples respectively. A much smaller increase in the percentage of reads 'greater than modal +30' was seen in the 5F-24 cycle sample, at 2.07%, compared to 1.83 and 1.85% for the 20 and 22 cycle samples respectively.

3.4. Discussion

Short tandem repeats over 200 base pairs in total length, including the expanded *HTT* CAG repeat found in cell models of Huntington's disease, cannot at present be sequenced with standard sequencing-by-synthesis technologies. Even those adapted for longer read lengths like Illumina's MiSeq generate reads do not span the entire repeat tract of the longest repeats in cell models and therefore cannot be uniquely mapped to a reference. While it is possible to reliably quantify the length of short tandem repeats above 200 base pairs using capillary electrophoresis methods like fragment analysis, they provide no information on the sequence of the repeat, including the 3' repeat flanking sequence, which has been associated with altered age at onset of HD after accounting for CAG tract length (Lee et al. 2019; Wright et al. 2019; McAllister et al. 2022).

Long-read sequencing technologies, on the other hand, routinely generate reads many kilobases in length. However, their uptake has historically been hindered by lower throughputs, accuracies and higher cost compared to their short-read counterparts. Recent advances in PacBio's SMRT sequencing technology, among others, have demonstrated that long-read sequencing is now highly accurate, and that high depth sequencing can be obtained, though the technologies are still expensive compared with shorter read sequencing (Pollard et al. 2018). It is likely that these costs will reduce as the technology matures.

In the work presented in this chapter I have shown that PacBio long-read sequencing can be used to count the CAG repeats in *HTT* and that the modal counts are well matched to validated repeat quantification methods. Read depths were more than an order of magnitude lower in the PacBio data compared with the MiSeq data. This meant that measures of somatic expansion were generally far higher because of the reduced signal-to-noise ratios of the resulting peaks. To increase the read depth of the expanded alleles for the iPSC samples, I reduced the number of samples per run from 48 to 6. Whilst the depth per sample increased, the 2.5x gain in read depth was not proportional to the 8x reduction in the number of samples. The larger difference of the expanded repeat of the 109NI line may result in greater enrichment of the WT allele during or prior to the sequencing itself.

Read depth of 130-CAG cell model expanded alleles with this technology can be increased by approximately 7-fold at no extra cost by using shorter amplicons and reducing the ratio of beads used during PCR clean-up to physically enrich for the expanded allele. A by-product of the shorter amplicon method is – for the same sequencing run time – increased median read quality: Q26 with no allele enrichment of 3000 bp amplicons to Q40 with two allele enrichments of 600 bp amplicons. While some genomic context is lost when sequencing a shorter amplicon, allelic phasing would require amplicons much larger, even, than 3000 bp and the disease-relevant variation at this locus is located at the repeat itself. For example, Svrzikapa et al. sequence a 10 kb cDNA encompassing 10 exonic SNPs in a 160 kb region around *HTT* to generate a haplotype for phasing the expanded repeats of individuals with HD (Svrzikapa et al. 2020).

RD produced comparable CAG repeat calls to ScaleHD, which has in turn been shown to produce comparable repeat calls – on MiSeq sequencing data (Ciosi et al. 2021)– to PCR fragment analysis, the method used to count *HTT* CAG repeats in predictive clinical testing for HD (Losekoot et al. 2013). Unlike ScaleHD, RD is compatible with the long reads produced by SMRT sequencing without trimming – including those from cell models where repeat lengths are much longer than in most adult-onset HD samples – and can detect novel repeat flanking structures. By analysing reads with multiple scoring profiles, it can detect the location of repeat flanking sequences within reads and, in doing so, their sequence.

PacBio modal CAG repeat length calls of patient expanded alleles were consistently shorter than their MiSeq equivalents using both ScaleHD and RD, with a difference of 0.38 and 0.65 CAGs respectively. This may be a consequence of the higher number of PCR cycles used in the library preparation with 28 used for MiSeq and 50 used for PacBio – PCR is known to generate stutter artefacts when amplifying repetitive DNA, which results in frameshift products that are generally shorter than the template strand (Murray et al. 1993; Daunay et al. 2019). This effect scales with the number of PCR cycles and may explain the difference in the modal peaks between sequencing methods. The difference between CAG counting methods is likely to be because ScaleHD is more tolerant of imperfect CAGs than RD.

The difference between PacBio-RD calls and fragment analysis calls was greater and more variable than the difference between PacBio and MiSeq calls and appears to

scale with repeat length. PacBio-RD mean modal CAG calls were higher than their fragment analysis equivalents by 0.35-1.07 for the WT allele and 5.5-8.8 for ~130 CAG expanded allele of 109NI iPSCs. A possible explanation for this difference may be that because CAG repeats form secondary structures which migrate faster in electrophoresis than heterogeneous sequence, CAG sizing by fragment analysis may produce fragment sizes that are systematically shorter than the true length. This would explain the observation of the margin of difference scaling with the repeat size, however, is yet to be established experimentally.

			p-value			
	Comparison		Modal CAG		Somatic Expansion / Expansion index	
Table	Dataset 1	Dataset 2	WT	Expanded	WT	Expanded
3.4	MiSeq-SHD	PacBio-SHD	****	****	****	ns
3.5	MiSeq-RD	PacBio-RD	****	****	***	ns
3.6	MiSeq-SHD	MiSeq-RD	****	****	****	****
3.7	PacBio-SHD	PacBio-RD	****	****	ns	ns
3.8	MiSeq-SHD-LBC	MiSeq-SHD-PBMC	****	****	****	ns
3.9	PacBio-SHD-LBC	PacBio-SHD-PBMC	****	****	ns	ns
3.11	FA	3000-iPSC	ns	**	N/A	ns
3.12	FA	3000-LBC-PBMC-iPSC	ns	**	N/A	ns
3.13	3000-iPSC	3000-LBC-PBMC-iPSC	****	**	ns	ns
3.15	3000-iPSC	600-iPSC-1	****	**	ns	ns
3.16	FA	600-iPSC-1	ns	**	N/A	ns
3.17	FA	600-iPSC-3	ns	***	N/A	****

Table 3.19. Summary of correlations made of *HTT* CAG count data. p-values: ns: non-significant. *: ≤ 0.05. ** < 0.01. *** < 0.001. **** < 0.0001. WT: wild type allele. Expanded: expanded allele. SHD: ScaleHD. RD: RepeatDecoder. FA: fragment analysis. LBC: Lymphoblastoid cells. PBMC: peripheral blood mononuclear cells. iPSC: induced pluripotent stem cells.

Modal CAG correlations made between datasets are summarised in table 3.19. Paired PacBio and MiSeq WT allele modal CAG counts (Tables 3.4 and 3.5) were significantly positively correlated, while paired PacBio and fragment analysis WT allele modal CAG counts (Tables 3.11, 3.12, 3.16 and 3.17) were not significantly positively correlated. This is likely due to the spread of the modal CAG counts in the respective datasets. All PacBio-MiSeq correlations were conducted on patient samples

which had a wide range of WT allele counts (15-34), while all PacBio-fragment analysis correlations were performed on 109NI cell lines which had a very narrow range of WT allele counts (19-21). Routine error of +/- 1 CAG in both PacBio and fragment analysis accounts for a lack of correlation in data with a range of 2 CAGs.

Where valid comparisons could be made, measures of WT allele somatic expansion were correlated in those comparisons which did not involve 2 PacBio datasets (Tables 3.4, 3.5, 3.6 and 3.8). There is relatively limited instability at these repeat lengths (Ciosi et al. 2019), hence it is only with very high depth sequencing (MiSeq) that the level of signal:noise is sufficient to give reliable measures of somatic expansion.

Expanded allele modal CAG counts were significantly positively correlated in all comparisons, however somatic expansion/expansion index was only correlated in two comparisons, that of MiSeq-SHD vs MiSeq-RD (Table 3.6) and FA vs 600-iPSC-3 (Table 3.17). Again, due to the limited instability of typical HD patient length repeats (36-55), correlated measures of somatic expansion were only observed with counts from identical samples sequenced at ultra-high depth. Even at 109NI cell line repeat lengths, PacBio's read depth and sensitivity was only sufficient to reliably detect the large changes in expansion index seen in the longitudinal data of 600-iPSC-3 (Table 3.17). Expansion over time is expected in these lines and expansion indices are therefore more spread than with previous data which are cross-sectional. Indeed, the standard deviations reflect the difference in the spread of expansion indices in this data with 2.79 and 2.32 for the fragment analysis and PacBio data respectively in the 600-iPSC-3 comparison, compared to SDs in the other 3 PacBio-fragment analysis comparisons which are all less than 1 (Tables 3.11, 3.12 and 3.16).

Other factors which may be contributing to the lack of correlation seen in the other PacBio-fragment analysis expansion index comparisons include the low sensitivity of fragment analysis and the different levels of PCR stutter resulting from library preparations with different numbers of PCR cycles (fragment analysis 30, PacBio 44-50). Also, all PacBio-fragment analysis correlation tests were performed on a small number of samples (either 6 or 12), which limits their power to detect a relationship.

Read depth of the WT allele was typically between 2 to 10 times higher than that of the expanded allele in non-size selected libraries and probably explains why

equivalent WT allele comparisons of Somatic Expansion (Tables 3.4-3.9) are more commonly correlated than expanded allele comparisons.

Of the LBC-PBMC comparisons, while modal CAG counts are well correlated on counts from both sequencing technologies, only the WT allele of MiSeq data had correlated values for Somatic Expansion (Table 3.8). Expanded allele Somatic Expansion was approaching significance in the MiSeq data but the scatter plot shows a weak association (Figure 3.11). The data makes clear that LBCs and PBMCs display relatively low levels of instability with similar means. Expansion index means are 0.34 and 0.36 for LBCs and PBMCs respectively for MiSeq data and 0.39 and 0.39 for LBCs and PBMCs respectively for the PacBio data. The low level of instability observed in these cell types may explain to some extent why expanded allele expansion indices were not correlated in data from either sequencing method.

Overall, these comparisons showed that PacBio data will give the same *HTT* CAG repeat length as existing and established methods of measuring repeat loci on typical patient repeat lengths. While measures of cell model-length repeats are systematically longer than one existing method, this is likely down to capillary electrophoresis-specific biases that can be calibrated for (see section 5.1 for further discussion). Importantly, PacBio measures of 130-CAG repeats are significantly correlated to fragment analysis measures and are reproducible (Table 3.18). Furthermore, given sufficient read depth and spread in the data, PacBio gives the same expansion indices as fragment analysis on cell model-length CAG repeats (Table 3.17).

The effect of PCR cycles was examined in 6 samples from 600-iPSC-2, however the data were largely inconclusive, with no clear trends emerging. Based on previous literature (Murray et al. 1993; Daunay et al. 2019), I expected the proportion of reads shorter than the modal CAG to increase with increasing PCR cycles, and while this was observed in the 11B11 line sample, the proportion only increased for the 24 cycle sample (by 12.4% from 20 cycles), not the 22 cycle sample. There was a smaller change in the 5F line from 20 cycles to 24 cycles (increase of 1.8%). Since the PCR was conducted on purified DNA, and the starting PCR template was identical across samples, I would not expect to see a difference between cell lines. If I had more time, I would have liked to repeat this experiment with a wider range of PCR cycle number and included multiple technical replicates to improve reliability.

As touched on several times in this discussion, one of the limitations of this work is the reliance on PCR amplification to generate libraries. PCR introduces errors (Fungtammasan et al. 2015), some of which are likely to have an effect on CAG repeat calling. Indeed, polymerases are known to amplify trinucleotide repeats and high GC content DNA poorly (Mamedov et al. 2008; Hommelsheim et al. 2015), and the CAG repeat is both of these. To mitigate this, as few amplification cycles as possible should be used in future library preparations. In addition, it may be possible to perform CAG count error correction or forgo the use of PCR altogether: PacBio and Oxford Nanopore have commercially available PCR-free sequencing kits; however, the read depth is currently in the order of several hundred per sample, too few to accurately assess somatic expansion in highly mosaic repeats, and library preparation requires input DNA in the order of several micrograms (Höijer et al. 2018; Giesselmann et al. 2019; Wieben et al. 2019).

Despite the limitations of PCR sequencing-based methods for *HTT* CAG quantification, the work here demonstrates that PacBio sequencing can be used to quantify the 130 CAG repeats of 109NI iPSC lines. Furthermore, longitudinal samples, those from experiments conducted over multiple time points, can generate meaningful measures of somatic expansion that are correlated to fragment analysis, a clinically accepted repeat sizing method.

Having optimised amplification, sequencing, and analysis of the *HTT* CAG repeat in a variety of samples, I am now going to apply this knowledge to experiments which will investigate the relationship between repeat sequence, *FAN1 genotype* and cell maturation in cell models of HD. In chapter 4 I study iPSCs which have been differentiated to medium spiny neurons, which are implicated in HD pathology.

Chapter 4 : CAG repeat dynamics in iPSC models of HD

4.1 Introduction

In the previous chapter, long read sequence data was used to measure *HTT* CAG repeat length and repeat length variation in cell models of HD with an uninterrupted CAG tract of approximately 130 in the *HTT* gene. Here I use this validated technique to investigate the levels of variation in individual alleles. I examine the effect of knocking out *FAN1*, a genetic modifier of the onset of HD (GeM-HD Consortium 2019), which has been shown to protect against somatic expansion in HD mice and neuronal iPSCs (Loupe et al. 2020; McAllister et al. 2022), on *HTT* CAG repeat length and sequence in post-mitotic neurons sampled at 4 time points covering 55 days.

The cell model used in these experiments was derived from the fibroblasts of an individual with 109 uninterrupted *HTT* CAG repeats. A subclone of the parent cell line underwent genetic manipulation by Jasmine Donaldson using CRISPR-Cas9 to introduce a homozygous knockout mutation in *FAN1*. Lines with $FAN^{+/+}$ and $FAN1^{-/-}$ genotypes have been terminally differentiated to striatal spiny neurons, the cell type most vulnerable in HD (Donaldson 2019). Over time the *HTT* CAG repeat has expanded and now has approximately 130 uninterrupted CAGs.

Variation in the expanded CAG repeat and its surrounding sequence have been observed in HD patients (Ciosi et al. 2019; Lee et al. 2019; Wright et al. 2019; McAllister et al. 2022) and we were interested to see whether we detected altered sequence in and around the repeat in our model. In this chapter I explore the effect of *FAN1 genotype* and CAG expansion length on the rate of flanking sequence alterations in individual reads.

Lastly, I look at whether large repeat expansions observed in my long-read sequencing data can be detected using small pool PCR, a technique with high sensitivity for detecting rare expansions (Ciosi et al. 2021).

4.2 Chapter aims

In this chapter I aim to answer the following research questions:

1. Can 109NI iPSC model expanded repeats be sequenced at sufficient depth to perform novel experiments relating to repeat expansion, stability, and sequence variation?
2. What is the effect of *FAN1 genotype* and cell age in iPSC models of post-mitotic neurons on expanded *HTT* CAG repeat length, stability, and flanking sequence?
3. Do reads with altered flanking sequences have altered repeat lengths in cell models compared to reads with non-altered flanking sequences?
4. Are the large CAG repeat expansions observed in long-read PacBio sequencing of iPSC models also observed in small pool PCR?

4.3 Results

4.3.1 Sequencing 109NI iPSC expanded repeats at sufficient depth for novel experiments

4.3.1.1 The effect of passage number and *FAN1 genotype* on CAG repeats

The first aim in this chapter is to establish whether iPSC model expanded repeats can be sequenced at sufficient depth to perform novel experiments relating to repeat expansion, stability, and sequence. To answer this question, I further analysed data from the 12 109NI samples of library 600-iPSC-3, partially analysed in the previous chapter (section 3.3.3.4). Details of the samples are shown in Table 3.18. Library 600-iPSC-3 was comprised of DNA from cells passaged to 4, 20 and 36 times from the *FAN1*$^{-/-}$ line 109NI-5F and the *FAN1*$^{+/+}$ line 109NI-11B11 and was amplified in duplicate. Details of library preparation can be seen in section 3.3.3.1.

The mean number of expanded allele reads surviving filtering was 8,158 per sample (min. 3,937, max. 12,232). 3 of the 6 pairs of duplicates had identical modal CAG, with the remaining 3 having a difference of 1 CAG. Figure 4.1 shows the frequency distribution of reads by CAG length from replicate 1 of each of the 6 cell line-harvest time combinations.

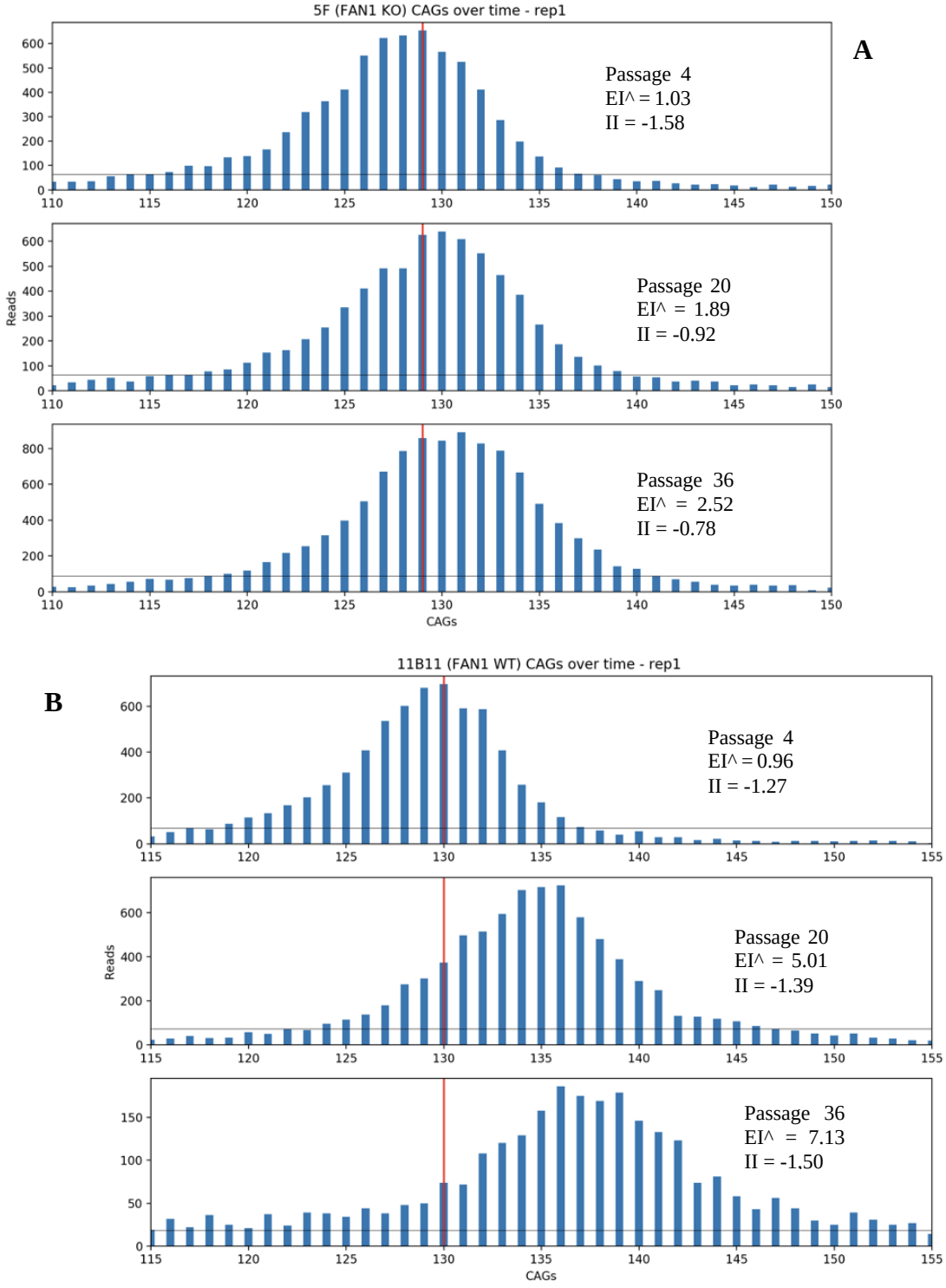


Figure 4.1. Expansion of the *HTT* CAG repeat over time in *FAN1*$^{-/-}$ and $^{+/+}$ cell lines using data from long read PacBio sequencing of library 600-iPSC-3. (A) *FAN1*$^{-/-}$ cells. (B) *FAN1*$^{+/+}$ cells. RepeatDecoder restrictive profile counts. Replicate 1 sample shown only. x-axes are aligned. Red lines indicate the modal CAG at passage 4. Horizontal black line represents the 10% threshold used to calculate EI^ and II. EI^: passage 4-anchored expansion index. II: instability index.

Figure 4.1 illustrates the difference in expansion over time in these cell lines. While there is a shift in the entire peak of about 1 CAG per time point in the knockout line, the shift in the WT line is much more pronounced, with around 5 CAGs from Passage

4 to Passage 20 and a further 2-3 CAGs to Passage 36. This is reflected in the passage 4-anchored expansion indices, plotted in Figure 4.2, and is likely to be due to the selection of a WT line with a particularly high expansion rate, as discussed at the end of this chapter and further in section 5.7. Also, repeats from WT line passage 36 appear to be more unstable, however it is uncertain whether this is a true reflection of the sample, noise in the data (the number of expanded allele reads from this sample was lower than the rest), or a combination of the two.

Figure 4.2 shows the change in modal CAG, passage 4-anchored expansion index and instability index over time in both cell lines.

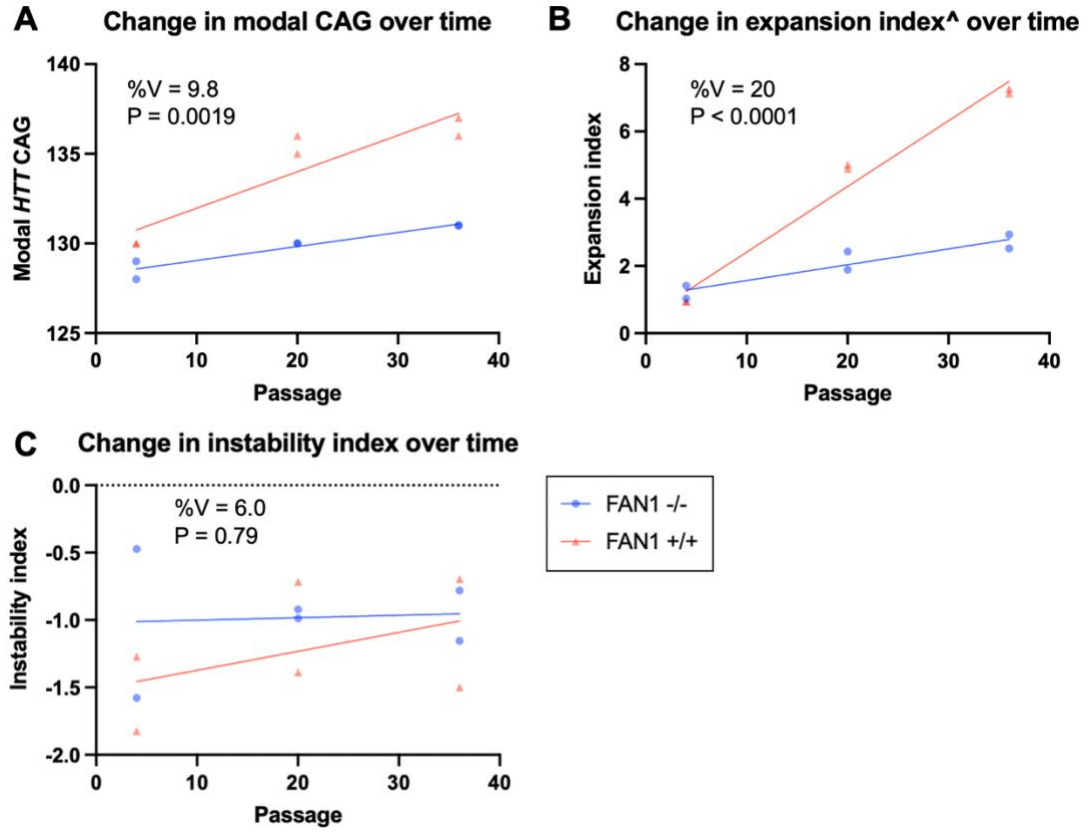


Figure 4.2. Change in modal CAG, passage 4-anchored expansion index and instability index over time in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines using data from long-read PacBio sequencing of library 600-iPSC-3. RepeatDecoder restrictive profile counts. Expanded alleles are those with restrictive profile counts of 30 or more. ^ passage 4-anchored. %V: percentage of total variation explained by the interaction between passage and *FAN1 genotype* in 2-way ANOVA. P: p-value of the interaction between passage and *FAN1 genotype* in 2-way ANOVA.

Modal CAG increases progressively over time in both cell lines (Figure 4.2A). The rate of increase in modal CAG over time was significantly higher in *FAN1*$^{+/+}$ cells compared with *FAN1*$^{-/-}$ cells (p = 0.0019; 2-way ANOVA). The *FAN1* genotype of the cells explained 9.8% of the variance in modal CAG observed. Again, this result is

likely due to the selection of a WT line with a particularly high expansion rate, as discussed at the end of this chapter.

Passage 4-anchored expansion index also increases progressively over time in both cell lines (Figure 4.2B). The rate of increase in passage 4-anchored expansion index over time was significantly higher in *FAN1*$^{+/+}$ cells compared with *FAN1*$^{-/-}$ cells ($p < 0.0001$; 2-way ANOVA). The *FAN1* genotype of the cells explained 20% of the variance in passage 4-anchored expansion index observed.

Instability index increases progressively in the *FAN1*$^{+/+}$ but not the FAN$^{-/-}$ line. Despite this, the rate of increase in expansion index over time was not significantly higher in *FAN1*$^{+/+}$ cells compared with *FAN1*$^{-/-}$ cells ($p = 0.79$; 2-way ANOVA). The *FAN1* genotype of the cells explained 6.0% of the variance in expansion index observed. All values of instability index were negative in both lines, meaning all distributions were negatively skewed in both lines.

Figure 4.3 shows that the proportion of expanded allele reads in each CAG length category changes little over time in the *FAN1*$^{-/-}$ line, with none of the categories deviating by more than 1% from the passage 4 proportions at passage 20 and passage 36. By contrast, the *FAN1*$^{+/+}$ line shows much larger changes, with the number of reads with a larger repeat than the modal CAG growing by 8.8% from passage 4 to passage 36. The proportion of reads ‘greater than the modal CAG plus 30’ almost doubles in that time (passage 4: 2.61%, passage 36: 4.35%). This is accompanied by equivalent reductions in the percentage of reads equal to and shorter than the modal CAG.

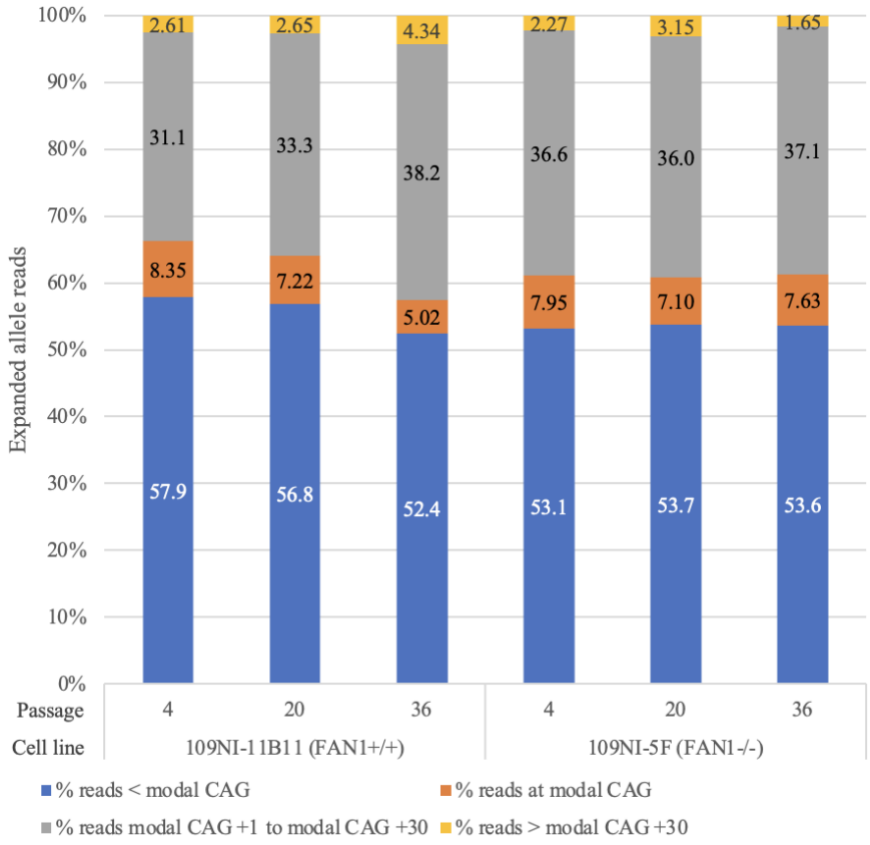


Figure 4.3. Percentage of 600-iPSC-3 reads by *HTT* CAG repeat length category across cell line and passage number. PacBio-RD restrictive profile counts of filtered expanded allele (>29 CAGs) reads. Read counts from replicates are combined to produce overall percentages.

4.3.1.2 *HTT* CAG repeat flanking sequence analysis

Part of the first aim for this chapter was to determine if PacBio can be used to perform novel experiments related to the sequence of the *HTT* CAG repeat. To this end I wanted to examine whether the flanking sequence at the 3' end of the repeat varied between reads of the same sample and, if so, how much variation existed between samples. Reads from 600-iPSC-3 were analysed with both the restrictive and permissive profiles in RD to generate coordinates for the 3' end of the polyCAG and polyglutamine repeats respectively. These co-ordinates were then used to extract the 'flanking sequence' string from each read.

Table 4.1 shows the read frequency of repeat flanking sequences in expanded alleles of all samples in library 600-iPSC-3. The most common flanking sequence in all 12 samples in 600-iPSC-3 library was "CAACAG", which is the typical *HTT* CAG repeat flanking sequence and the same flanking sequence as the parent line (as determined by Sanger sequencing). The frequency of this sequence in all reads in 600-iPSC-3 was 84.8%, with a range of 3.3% between samples (maximum 86.0%, minimum 82.7%).

15.2% of all reads were identified to have an altered flanking sequence. The top 5 ranking sequences account for 90.3% of all reads. 1,749 unique flanking sequences were identified in 97,900 reads, 1,615 of which (92.3%) were observed less than ten times.

The second and third top ranking sequences in all 600-iPCS-3 samples was either “C” or loss of the CAACAG. These sequences shared between 1.91 and 2.77% of the total reads in all samples. The fourth and fifth top ranking sequences in all 600-iPCS-3 samples was either “CAACAACAG” or “CAACAGCAACAG”. These sequences shared between 0.32 and 0.69% of the total reads in all samples. The origin of these variations is not known, however PCR-free approaches would allow you to eliminate DNA amplification as the source of this and the per-base accuracies provided by PacBio sequencing give likelihood that they arise during the sequencing itself. This subject is discussed further at the end of this chapter.

Sample number	Cell line	Passage	Rep	Flank 1	Flank 2	Flank 3	Flank 4	Flank 5	% Flank 1	% Flank 2	% Flank 3	% Flank 4	% Flank 5	% Flank 1-5
1	*FAN1-/-*	4	1	CAACAG	C		CAACAACAG	CAACAGCAACAG	84.3	2.28	2.13	0.529	0.505	89.7
2	*FAN1-/-*	4	2	CAACAG	C		CAACAGCAACAG	CAACAACAG	85.1	2.35	2.16	0.447	0.409	90.5
3	*FAN1-/-*	20	1	CAACAG	C		CAACAGCAACAG	CAACAACAG	84.5	2.36	2.21	0.483	0.483	90.0
4	*FAN1-/-*	20	2	CAACAG		C	CAACAGCAACAG	CCAGCAACAG	84.5	2.64	2.54	0.483	0.399	90.5
5	*FAN1-/-*	36	1	CAACAG		C	CAACAACAG	CAACAGCAACAG	86.0	2.14	2.13	0.466	0.433	91.1
6	*FAN1-/-*	36	2	CAACAG	C		CAACAGCAACAG	CAACAACAG	85.6	2.23	1.92	0.667	0.323	90.8
7	*FAN1+/+*	4	1	CAACAG		C	CAACAGCAACAG	CAACAACAG	85.2	2.39	1.99	0.629	0.568	90.7
8	*FAN1+/+*	4	2	CAACAG		C	CAACAGCAACAG	CAACAACAG	84.7	2.33	1.91	0.559	0.528	90.0
9	*FAN1+/+*	20	1	CAACAG		C	CAACAGCAACAG	CAACAACAG	85.0	2.42	2.32	0.618	0.415	90.8
10	*FAN1+/+*	20	2	CAACAG		C	CAACAGCAACAG	CAACAACAG	83.8	2.67	2.23	0.497	0.481	89.7
11	*FAN1+/+*	36	1	CAACAG		C	CAACAACAG	CAACAGCAACAG	84.3	2.46	2.46	0.559	0.432	90.2
12	*FAN1+/+*	36	2	CAACAG	C		CAACAACAG	CAACAGCAACAG	82.7	2.77	2.74	0.689	0.495	89.4
Mean									84.6	2.42	2.23	0.552	0.456	90.3

Table 4.1. Top 5 most frequent flanking sequences and their normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ neuronal cell lines by sample using data from long-read PacBio sequencing of library 600-iPSC-3. Flanking sequence: sequence between the 3' ends of RepeatDecoder restrictive and permissive profile repeat sequences. Expanded alleles are those with restrictive profile counts of 30 or more. Rep: PCR replicate. Flank 1 is most frequent flanking sequence, Flank 2 is the second most frequent flanking sequence and so on.

Table 4.2 shows the 15 most frequent flanking sequences in the two cell lines used in library 600-iPSC-3. Differences in the ranking of sequences exist between cell lines, however, differences in the percentage share of reads of each sequence is at most 0.7% between lines. Manual inspection of a random sample of 30 sequences associated with the flanking sequence "C" from library 600-iPSC-4 (see Appendix 1B) suggested they are primarily a mixture of insertion and deletion errors around the junction between the polyCAG repeat and the polyproline repeat with variable basecall quality scores.

A **109NI-11B11 (FAN1+/+)**

	Repeat flanking sequence	Reads	% of total
1	CAACAG	36874	84.42
2		1082	2.48
3	C	972	2.23
4	CAACAGCAACAG	243	0.56
5	CAACAACAG	230	0.53
6	CAAGCAG	115	0.26
7	CCAG	99	0.23
8	CAACAGCCGC	95	0.22
9	CAACAGCAGCAACAG	94	0.22
10	CAAGCAGCAACAG	85	0.19
11	CAACA	85	0.19
12	CAAGCAGCAGCAACAG	84	0.19
13	GCAGCAACAG	81	0.19
14	CCAGCAACAG	78	0.18
15	CCAGCAGCAACAG	75	0.17

B **109NI-5F (FAN1-/-)**

	Repeat flanking sequence	Reads	% of total
1	CAACAG	46150	85.11
2	C	1239	2.29
3		1172	2.16
4	CAACAGCAACAG	271	0.50
5	CAACAACAG	235	0.43
6	CCAG	128	0.24
7	CAAGCAG	122	0.23
8	CAACAGCCGC	118	0.22
9	CAACAGCAGCAACAG	116	0.21
10	CCAGCAACAG	116	0.21
11	GCAACAG	112	0.21
12	CAAGCAGCAACAG	100	0.18
13	CAAGCAACAG	100	0.18
14	GCAGCAGCAGCAACAG	100	0.18
15	GCAGCAGCAACAG	98	0.18

Table 4.2. Top 15 most frequent flanking sequences and read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ neuronal cell lines using data from long-read PacBio sequencing of library 600-iPSC-3. (A) *FAN1*$^{+/+}$ line. (B) *FAN1*$^{-/-}$ line. Flanking sequence: sequence between the 3' ends of RepeatDecoder restrictive and permissive profile repeat sequences. Expanded alleles are those with restrictive profile counts of 30 or more. The 15 most frequent flanking sequences are shown. Sequences highlighted yellow appear in both tables and share the same ranking. Sequences highlighted in grey appear in both tables but do not share the same ranking. Sequences in white only appear in one table.

The data generated in library 600-iPSC-3 could be analysed to gain further insight into the effect of flanking sequences, including to see whether changes in the profile of flanking sequences are associated with changes in cell age or CAG length. However, the analysis above demonstrates that 109NI iPSC expanded repeats can be sequenced at sufficient depth to gain insights relating to repeat expansion, stability and sequence and therefore satisfies the first aim of this chapter. I use these approaches to investigate these questions more comprehensively in the next experiment.

4.3.2 The effect of *FAN1* genotype and cell age in iPCS models of post-mitotic neurons on *HTT* CAG expanded allele repeat length, stability, and flanking sequence

4.3.2.1 Cell culture

To examine the effect of *FAN1 genotype* in stem cell models of HD, I used the same lines as those in chapter three. The *FAN1* wild-type line 11B11 and the isogenic homozygous *FAN1* knockout line 5F were cultured in triplicate by Jasmine Donaldson. A family tree showing the relationship of these lines to the 109NI parent line is shown in Figure 3.13. This experiment sought to establish rates of change in *HTT* CAG repeat lengths, somatic expansion rates, any changes in the sequence that occur between or within individual cultures and whether *FAN1 genotype* influenced any of those parameters in post-mitotic neurons.

iPSCs were first differentiated to neural precursor cells (NPCs). Following a 16-day neural induction of iPSCs to NPCs, cells were plated for terminal differentiation to forebrain neurons using a method established by Jasmine Donaldson (Donaldson 2019; McAllister et al. 2022), details of which are shown in section 2.2.2.2. While I did not stain for neuronal markers, using this method Jasmine Donaldson saw markers of MSN-like cells in these lines (Donaldson 2019). 5F and 11B11 represent individual clones of the parent line 109NI, which were cultured (see 2.1.1.2) in triplicate for each of four harvest time points – day 16, 37, 52 and 71 for a total of 24 wells cultured (Figure 4.4). Light micrographs of the cells were taken before they were harvested, pelleted and frozen at the specified time points (except day 16 – images were taken on day 18, two days after plating NPCs). Once all the cells had been harvested, pellets were defrosted in a water bath and the DNA was extracted (see 2.3.1).

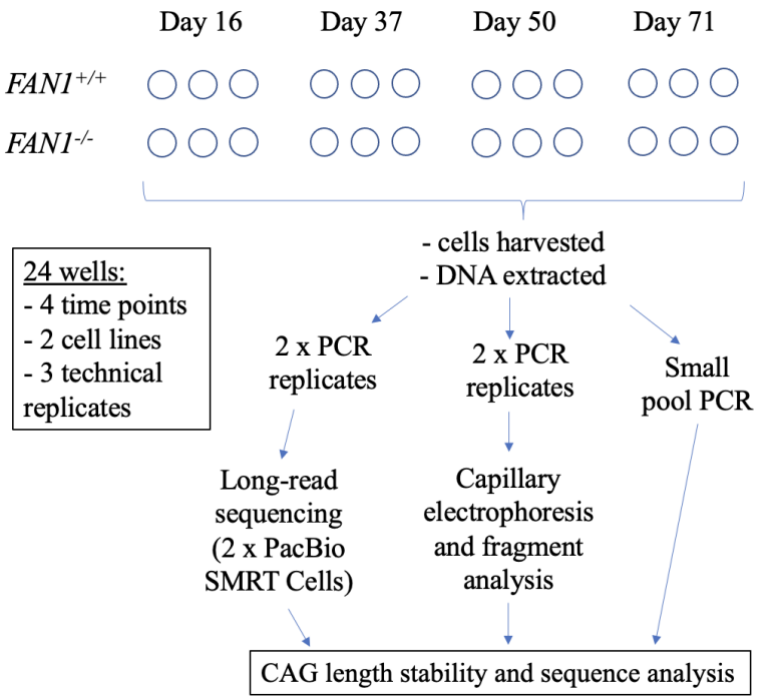


Figure 4.4. Experimental design of *FAN1* knockout experiment. Circles represent individual wells which were inoculated with neural precursor cells (NPCs) following a 16-day neural induction of iPSCs to NPCs. Day 16 represents zero days since plating for terminal differentiation to forebrain neurons.

4.3.2.2 Cell images

Cells were imaged at day 18, 37, 50 and 71 (2.1.8). Figure 4.5 shows a comparison of the cells over time and across cell lines. Neurons of both lines exhibit similar increases in branching and cell death over time. Dead cells appear as clumps of white dots, as indicated by the white arrow in the *FAN1*$^{+/+}$ day 71 image.

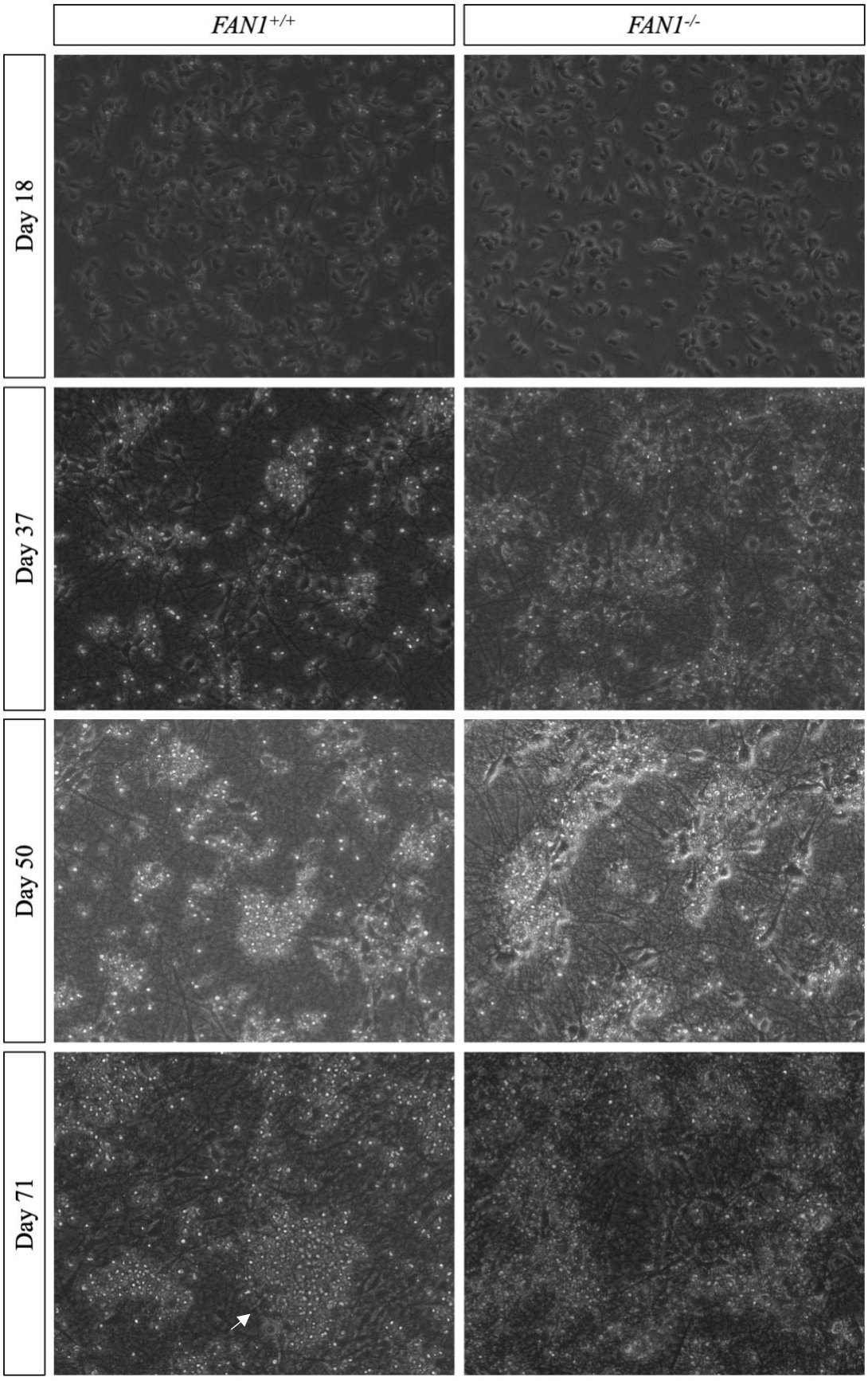


Figure 4.5. Maturation of 109NI *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ NPCs into terminally differentiated forebrain neurons. Bright field images taken at 10x magnification. Day 18 is two days after plating NPCs, which were harvested at day 16. The white arrow marks a cluster of dead cells.

4.3.2.3 CAG sizing by fragment analysis

Extracted DNA was amplified for fragment analysis (2.1.6) and checked by agarose gel electrophoresis (2.1.7). This was performed in duplicate on successive days. Bands were observed at approximately 150 bp and 450 bp – the expected product sizes for the WT and expanded allele for these primers and cell lines – for all 48 samples. No bands were observed in the water-only control. Amplified DNA was prepared for fragment analysis (2.1.6) and sent to The All Wales Medical Genomics Service, where the capillary electrophoresis of all samples was performed in the same run. Repeat sizes were extracted from the resulting data as described in section 2.4.

Figure 4.6 shows capillary electrophoresis traces analysed by fragment analysis from a set of representative samples for each cell line. Both cell lines exhibit progressive increases in modal CAG over time, with the *FAN1*$^{+/+}$ line undergoing an increase of 2.0 CAGs and the FAN$^{-/-}$ line undergoing an increase of 2.7 CAGs. The non-integer value of CAGs is due to the way fragment analysis calculates fragment length – by base pairs rather than by triplets – and the degree of error in the data: typically, +/- 1 base pair.

Progressive expansion over time is also reflected in the day 16-anchored expansion indices (see section 3.3.3.3 for an explanation of anchoring), with the FAN$^{+/+}$ line increasing from 1.20 at day 16 to 1.71 at day 71 (increase of 42.5%). As with modal CAG, the day 16-anchored expansion index increases more in the FAN$^{-/-}$ line, starting at 1.41 at day 16 and increasing to 2.51 at day 71 (increase of 78.0%).

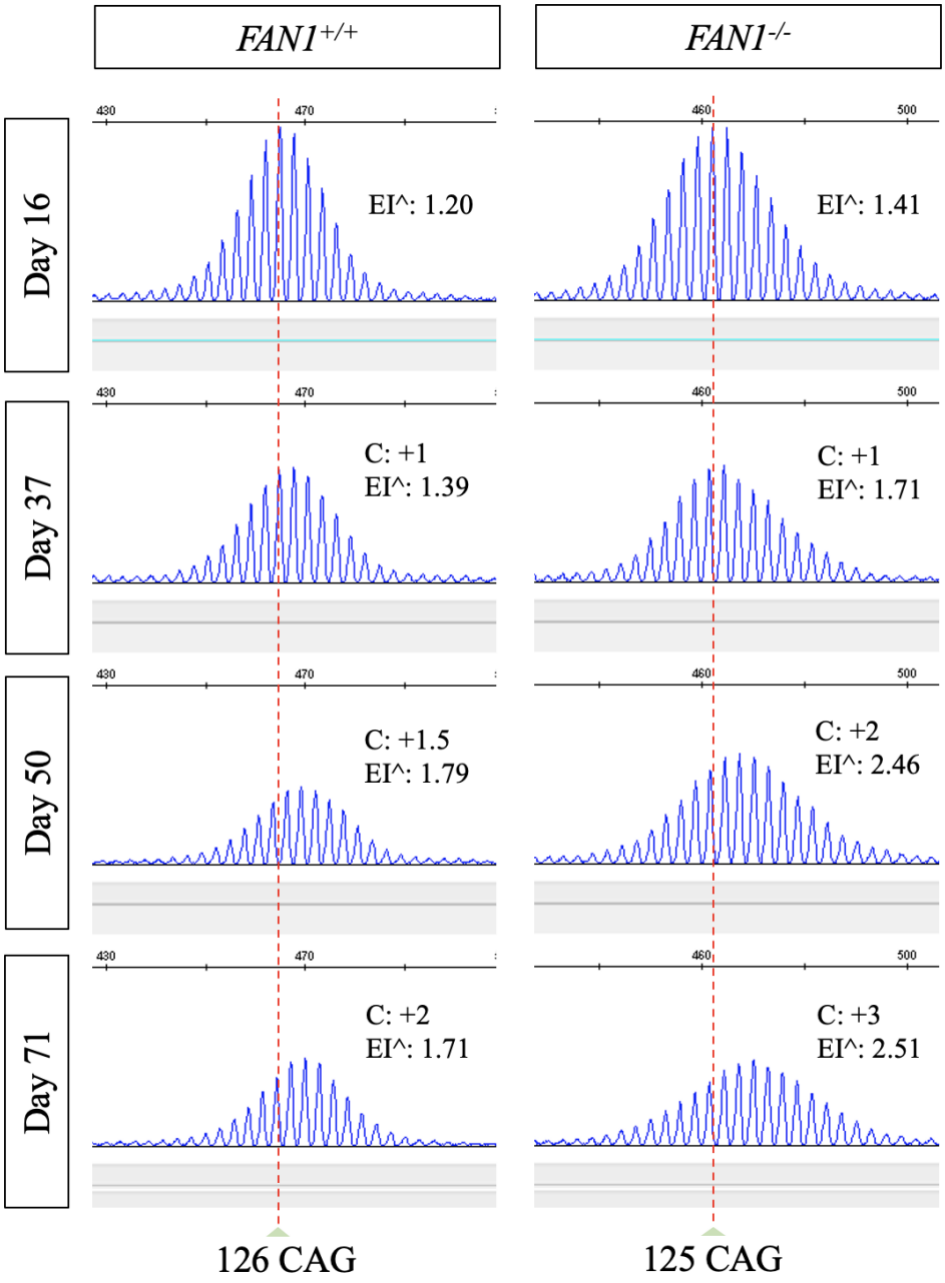


Figure 4.6. Representative electropherograms from fragment analysis of the expanded *HTT* CAG repeat in 109NI-*FAN1*$^{+/+}$ iPSCs compared to 109NI-*FAN1*$^{-/-}$ iPSCs across 4 time points. All samples from PCR 2 of replicate 2. Red lines indicate the modal CAG at day 16. C: change in modal CAG from day 16. EI^: day16-anchored expansion indices.

Summary data for all 48 samples analysed is shown in table 4.3. Expansion and instability indices are calculated as described in section 3.3.2.4.

To assess the amount of Spread in CAG repeat distribution data I decided to use a modified version of the instability index, the calculation of which is shown in Figure 3.16. Spread is defined here as the sum of the products of normalised read counts and the absolute change from the modal peak for CAG lengths with a read frequency greater than 10% of the modal CAG frequency.

A

PCR	Cell line	Day	Rep	WT allele: Modal CAG	WT allele: CAG change	WT allele: Expansion index	Expanded allele: Modal CAG	Expanded allele: CAG change	Expanded allele: Expansion index^	Expanded allele: Expansion index	Expanded allele: Instability index	Spread
1	*FAN1* +/+	16	1	17.3	0.0	0.00	125.7	0.0	1.06	1.06	-0.06	2.18
			2	17.0	0.0	0.16	125.7	0.0	1.09	1.09	0.06	2.11
			3	17.0	0.0	0.16	125.7	0.0	1.06	1.06	0.03	2.10
		37	1	17.7	0.3	0.29	127.0	1.3	1.40	0.91	-0.58	2.40
			2	17.7	0.7	0.38	127.0	1.3	1.46	0.94	-0.39	2.28
			3	17.7	0.7	0.00	127.0	1.3	1.38	0.89	-0.50	2.28
		50	1	17.7	0.3	1.00	128.7	3.0	1.70	0.72	-1.01	2.46
			2	19.7	2.7	0.00	128.0	2.3	2.07	1.05	-0.22	2.32
			3	17.7	0.7	0.79	127.0	1.3	1.70	1.14	-0.09	2.36
		71	1	19.7	2.3	0.00	128.0	2.3	2.46	1.37	0.20	2.55
			2	19.7	2.7	0.00	128.0	2.3	2.07	1.09	-0.15	2.33
			3	19.0	2.0	0.00	129.7	4.0	2.93	1.12	0.21	2.03
	FAN1 -/-	16	1	19.3	0.0	0.00	125.0	0.0	1.36	1.36	0.15	2.58
			2	18.7	0.0	0.22	125.0	0.0	1.34	1.34	0.05	2.62
			3	19.3	0.0	0.00	124.7	0.0	1.24	1.24	0.01	2.47
		37	1	19.3	0.0	0.00	125.7	0.7	1.57	1.28	-0.11	2.67
			2	18.7	0.0	0.00	125.7	0.7	1.78	1.47	0.11	2.83
			3	19.3	0.0	0.00	126.0	1.3	2.20	1.57	0.32	2.83
		50	1	18.7	-0.7	0.00	127.0	2.0	2.45	1.52	0.13	2.91
			2	18.7	0.0	0.20	128.0	3.0	2.62	1.42	-0.35	3.20
			3	18.7	-0.7	0.00	127.0	2.3	2.70	1.61	0.24	2.98
		71	1	19.3	0.0	0.00	131.0	6.0	2.59	0.74	-1.65	3.12
			2	19.3	0.7	0.00	127.0	2.0	3.35	2.29	1.05	3.53
			3	18.7	-0.7	0.00	127.0	2.3	3.42	2.20	0.94	3.46
Mean	-	-	-	18.6	0.5	0.13	126.9	1.65	1.96	1.27	-0.07	2.61

B

PCR	Cell line	Day	Rep	WT allele: Modal CAG	WT allele: CAG change	WT allele: Expansion index	Expanded allele: Modal CAG	Expanded allele: CAG change	Expanded allele: Expansion index^	Expanded allele: Expansion index	Expanded allele: Instability index	Spread
2	*FAN1* +/+	16	1	17.0	0.0	0.31	126.3	0.0	1.20	1.20	0.33	2.07
			2	17.0	0.0	0.30	126.0	0.0	1.20	1.20	0.28	2.13
			3	17.7	0.0	0.00	126.0	0.0	1.32	1.32	0.40	2.23
		37	1	17.7	0.7	0.34	127.0	0.7	1.11	0.86	-0.47	2.18
			2	17.7	0.7	0.22	127.0	1.0	1.39	1.00	-0.26	2.25
			3	17.7	0.0	0.13	127.0	1.0	1.38	0.99	-0.32	2.29
		50	1	19.7	2.7	0.00	128.7	2.3	1.48	0.73	-0.90	2.35
			2	19.3	2.3	0.00	127.7	1.7	1.79	1.06	-0.18	2.31
			3	17.3	-0.3	0.57	126.7	0.7	1.43	1.13	0.04	2.23
		71	1	17.7	0.7	1.16	128.0	1.7	2.28	1.45	0.38	2.52
			2	19.3	2.3	0.00	128.0	2.0	1.71	0.90	-0.39	2.19
			3	19.3	1.7	0.00	130.7	4.7	2.28	0.68	-0.73	2.09
	FAN1 -/-	16	1	19.3	0.0	0.00	125.0	0.0	1.39	1.39	0.27	2.52
			2	19.3	0.0	0.00	125.3	0.0	1.41	1.41	0.29	2.54
			3	19.3	0.0	0.00	125.3	0.0	1.35	1.35	0.03	2.66
		37	1	19.3	0.0	0.00	126.0	1.0	1.95	1.47	0.32	2.61
			2	19.3	0.0	0.00	126.0	0.7	1.71	1.41	0.19	2.62
			3	19.3	0.0	0.00	126.0	0.7	1.96	1.64	0.43	2.85
		50	1	19.3	0.0	0.00	127.0	2.0	2.54	1.59	0.29	2.89
			2	19.3	0.0	0.00	127.0	1.7	2.46	1.66	0.33	2.99
			3	19.3	0.0	0.00	127.0	1.7	2.44	1.63	0.38	2.87
		71	1	19.3	0.0	0.00	129.7	4.7	2.98	1.08	-0.52	2.68
			2	19.3	0.0	0.00	128.0	2.7	2.51	1.33	-0.30	2.97
			3	19.3	0.0	0.00	127.0	1.7	3.11	2.19	1.04	3.34
Mean	-	-	-	18.7	0.4	0.1	127.0	1.3	1.85	1.28	0.04	2.52

Table 4.3. Summary data from fragment analysis of the pure CAG repeat in *HTT* of *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells. (A) PCR replicate 1 values. (B) PCR replicate 2 values. Rep: Replicate. ^ Day 16-anchored.

Table 4.3 shows that WT allele CAG repeats in the FAN$^{+/+}$ line appear to expand over time, with a mean change of +1.9 CAGs across replicates (culture and PCR). PacBio

and Sanger sequencing data of these lines shows 20 uninterrupted CAGs is modal for the WT allele of the $FAN^{+/+}$ line and 19 CAGs for the $FAN^{-/-}$ line, whereas the mean modal CAG of the $FAN1^{+/+}$ by fragment analysis is 17.2. Inspection of the electrophoresis traces of this line showed fluorescent peaks 5 bp wide, consistent with overloaded PCR products and multi-modal peaks consistent with PCR artefacts. The WT allele should be stable in these lines with CAG changes of no more than +/-1 CAG expected – true of the equivalent PacBio data (Table 4.8). Fragment analysis data of the $FAN^{-/-}$ line does not appear to expand over time, with a mean modal change of 0 CAGs across all replicates.

Mean summary statistics of PCRs 1 and 2 are plotted in Figure 4.7. Expanded allele modal CAG increases progressively over time in both cell lines. Each modal CAG unit increase occurred in 19.6 days in $FAN1^{+/+}$ cells, compared to 17.2 days in $FAN1^{-/-}$ cells. The rate of increase in modal CAG over time was not significantly higher in $FAN1^{-/-}$ cells compared with $FAN1^{+/+}$ cells ($p = 0.91$; 2-way ANOVA). The *FAN1* genotype of the cells explained 0.82% of the variance in modal CAG observed.

Expanded allele day 16-anchored expansion index also increases progressively over time in both cell lines. The rate of increase in day 16-anchored expansion index over time was significantly higher in $FAN1^{-/-}$ cells compared with $FAN1^{+/+}$ cells ($p = 0.046$; 2-way ANOVA). The *FAN1* genotype of the cells explained 3.8% of the variance in day 16-anchored expansion index observed.

Expanded allele Spread also increases progressively over time in both cell lines. The rate of increase in expansion index over time was significantly higher in $FAN1^{-/-}$ cells compared with $FAN1^{+/+}$ cells ($p = 0.030$; 2-way ANOVA). The *FAN1* genotype of the cells explained 6.6% of the variance in Spread observed.

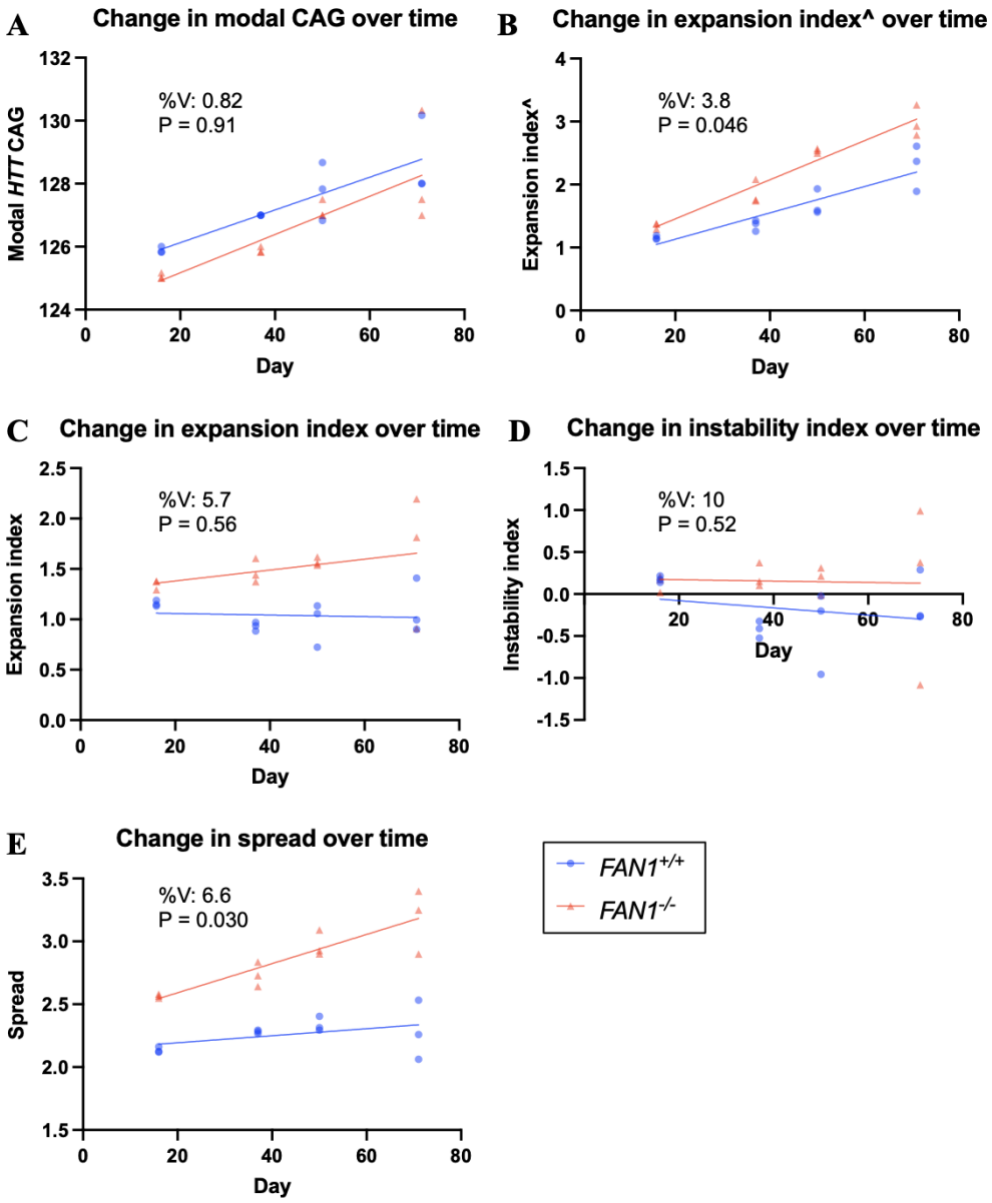


Figure 4.7. Change in modal CAG, day 16-anchored expansion index, expansion index, instability index and Spread over time for *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines using data from fragment analysis. Circles and triangles represent the mean of PCR replicates for 3 biological replicates per time point per cell line. Lines are simple linear regression lines of best fit. ^: day-16 anchored. %V: percentage of total variation explained by the interaction between harvest day and *FAN1 genotype* in 2-way ANOVA. P: p-value of the interaction between harvest day and *FAN1 genotype* in 2-way ANOVA.

Figures 4.7A, B and E show how the Modal CAG, day 16-anchored expansion index and Spread increase over time in both cell lines. Figure 4.7C and D show how unanchored expansion and instability indices change over time in both cell lines. The trend for expansion index in the *FAN1*$^{+/+}$ line is negative, while the trend for the FAN$^{-/-}$ line is positive. Despite this, the rate of increase in expansion index over time was not significantly higher in *FAN1*$^{-/-}$ cells compared with *FAN1*$^{+/+}$ cells (p = 0.56; 2-way ANOVA). The *FAN1* genotype of the cells explained 5.7% of the variance in

expansion index observed. This may be explained by the unusually high degree of variation in day 71 *FAN1*$^{-/-}$ cells.

Instability index, a measure of the distribution of peaks around the modal CAG, stays relatively stable over time in both lines and the rate of increase in instability index over time was not significantly higher in *FAN1*$^{-/-}$ cells compared with *FAN1*$^{+/+}$ cells ($p = 0.52$; 2-way ANOVA). The *FAN1* genotype of the cells explained 10% of the variance in instability index observed.

Overall, the modal CAG of FAN$^{-/-}$ cells is increasing over time and at a similar rate to the FAN $^{+/+}$ line, but the whole distribution is spreading at a higher rate than the FAN$^{-/-}$ line, which in turn is likely responsible for the higher expansion rates of index increase observed, as there is little change in the mean skew of the distributions (instability index). Deviations from the Normal distribution of peaks around the modal CAG length observed at day 16 emerge over time in individual clones (replicates), and are most pronounced in *FAN1*$^{-/-}$ at day 71.

4.3.2.4 PacBio library preparation

Libraries for SMRT sequencing were prepared in the same way as the 600 bp iPSC libraries in chapter three (3.3.3.1). PCR was conducted in duplicate on all 24 samples on successive days and a water control from the second day was included to give a total of 49 samples. Amplicons were pooled and prepared for sequencing using PacBio's Express V2 sequencing kit. Library preparation was checked by capillary electrophoresis (Figure 4.8). Expected product sizes are 359 for the WT allele and 689 for the expanded allele, for SMRTbell products with both adapters. The trace shows peaks at 328 bp, which corresponds to the WT peak, 510 bp, a peak which doesn't correspond to either allele and is likely to be a library preparation artefact that will be removed in the analysis pipeline, a peak at approximately 670 bp, which probably represents the expanded allele, a peak at 762 which could be library preparation artefacts and a broad smear from 800 – 6000 bp, which are also probably artefacts.

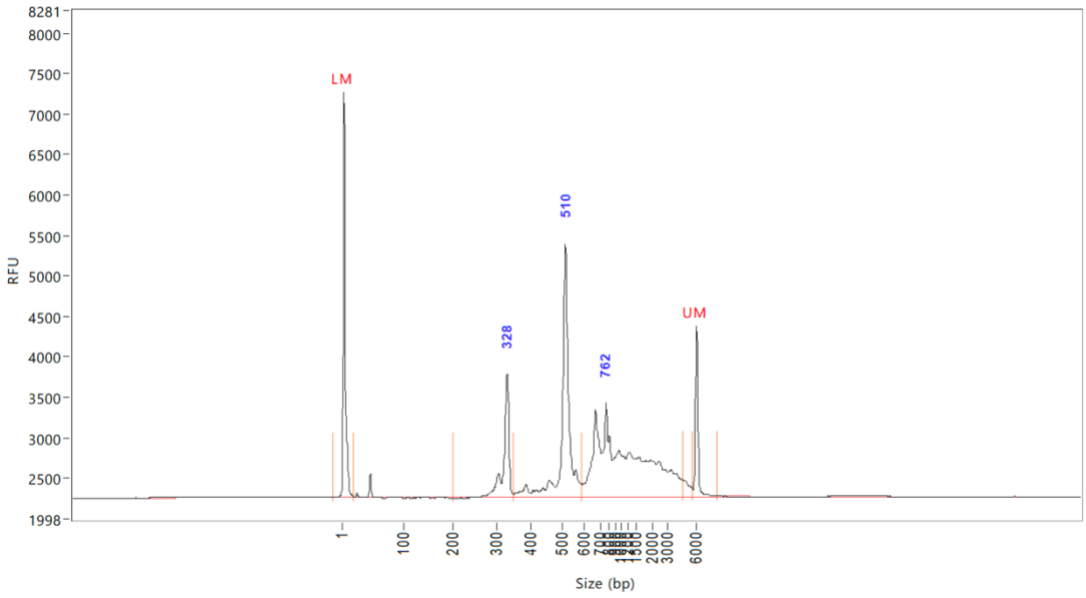


Figure 4.8. Capillary electrophoresis trace of pooled SMRTbell library 600-iPSC-4. LM: lower marker. UM: upper marker. bp: base pairs. RFU: relative fluorescence units.

The library was frozen and shipped to Exeter University where it was sequenced on two SMRTcells on successive days on a PacBio Sequel machine.

4.3.2.5 Sequencing data, quality control and filtering

413,512 reads were generated across the two SMRTcells of which 376,869 were good quality HiFi reads (2.1.4.1). Median HiFi read quality was Q42 (Figure 4.9A). Predicted read accuracy, Q, represents the average per-base quality score, which are derived from log-likelihood values computed by a Hidden Markov Model algorithm. See section 1.6.2 for a more detailed description of Q scores. The mean number of passes (subreads) per read was 22. The vast majority of predicted read accuracies are above Q20, with a mode of Q50 (maximum score).

Mean read length was 517 bp. Read lengths were distributed in two main peaks (Figure 4.9B), the first around 300 bp and the second at 500-700 bp. This is broadly consistent with the capillary electrophoresis trace in Figure 4.8, except the smear between 800-6000 bp is absent suggesting the smear comprises library preparation artefacts that did not comprise a CCS read or associate to a barcode in the first part of the sequencing analysis (consensus sequencing and demultiplexing).

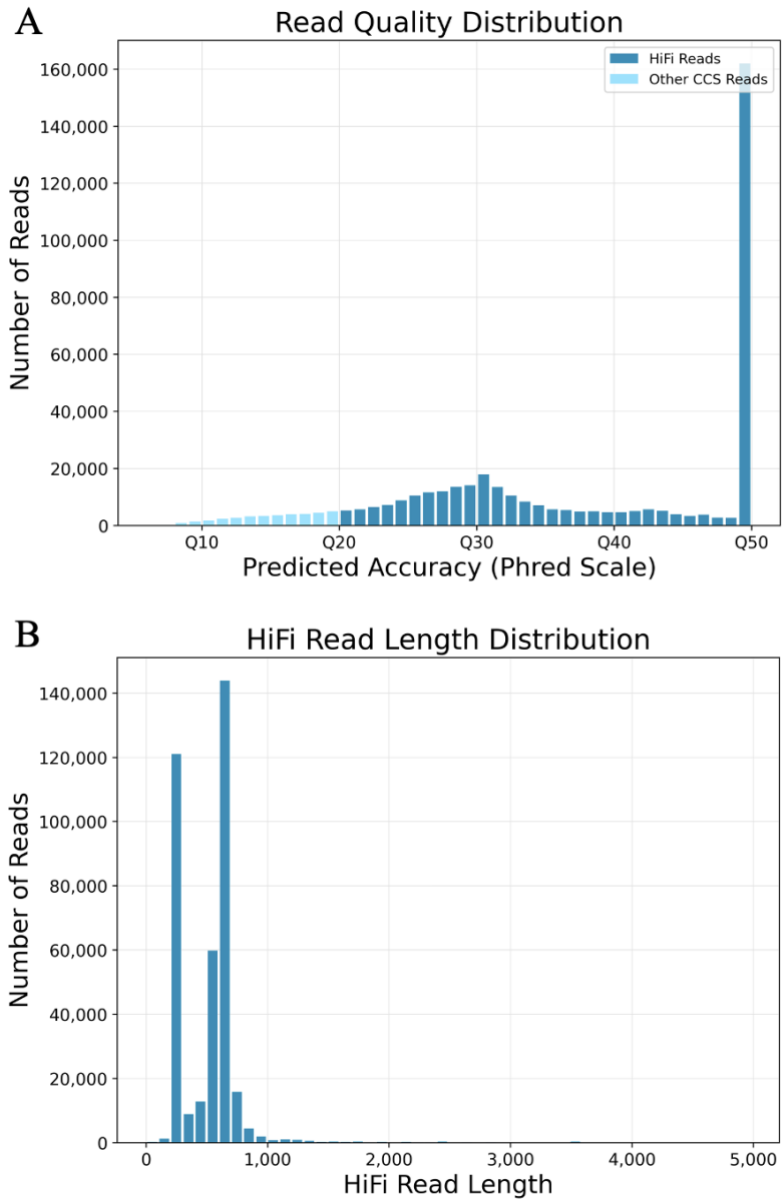


Figure 4.9. PacBio sequencing metrics for the 109NI *FAN1*$^{+/+}$ and $^{-/-}$ iPSC *HTT* CAG repeat library, 600-iPSC-4. (A) Read quality distribution plot. Q: Phred score. See section 1.6.2. for explanation of Phred scores. HiFi: reads with a Q score of 20 or more. CCS: circular consensus sequence. (B) Read length distribution plot.

Demultiplexing identified 49 unique barcodes associated with 361,332 reads. The minimum and maximum number of reads associated with a single non-control barcode was 3,187 and 22,216 reads respectively, with a mean of 5,884 (Table 4.4, Figure 4.10).

FASTQ files from the demultiplexed HiFi read set were downloaded from Smrtlink and run through a RD pipeline to generate CAG counts for each read before both the FASTQs and RD data was imported into the custom python analysis pipeline used in chapter 3 (see 3.3.1.2.). Reads are categorised by CAG length as in section (3.3.1.2.), i.e. < 7 CAGs: 'short', 7-29 CAGs: WT, >29 CAGs: expanded.

PCR	Cell line	Day	Replicate	Short retained	Short removed	WT retained	WT removed	Expanded retained	Expanded removed	Total
1	FAN1-/-	16	1	2	1,472	1,520	51	3,270	166	6,481
			2	1	2,279	1,231	32	3,639	77	7,259
			3	1	460	520	15	2,911	57	3,964
		37	1	1	567	1,400	44	4,531	168	6,711
			2	1	667	1,702	22	5,200	123	7,715
			3	0	516	669	35	4,220	177	5,617
		50	1	0	203	392	7	4,312	97	5,011
			2	5	170	7,247	66	2,862	82	10,432
			3	1	297	632	16	5,139	102	6,187
		71	1	0	165	527	26	2,546	175	3,439
			2	1	648	788	28	4,129	104	5,698
			3	2	340	651	19	5,477	112	6,601
	FAN1+/+	16	1	0	542	860	47	5,217	108	6,774
			2	2	806	3,525	199	3,187	164	7,883
			3	0	259	572	23	5,642	85	6,581
		37	1	0	235	499	40	3,641	93	4,508
			2	1	198	500	30	3,680	157	4,566
			3	0	230	461	36	4,331	111	5,169
		50	1	1	670	1,327	52	4,507	78	6,635
			2	0	524	1,244	69	3,804	125	5,766
			3	5	1,050	9,538	288	4,540	95	15,516
		71	1	1	758	731	48	5,539	119	7,196
			2	2	1,758	1,890	67	6,314	116	10,147
			3	1	224	1,346	122	2,131	353	4,177
2	FAN1-/-	16	1	4	377	1,385	112	2,869	355	5,102
			2	1	702	853	46	1,539	46	3,187
			3	0	401	2,880	114	2,534	72	6,001
		37	1	4	687	3,233	56	2,558	76	6,614
			2	5	726	7,976	212	5,000	85	14,004
			3	1	1,547	3,274	48	3,885	143	8,898
		50	1	1	1,982	4,235	225	4,549	181	11,173
			2	1	593	1,786	162	2,523	234	5,299
			3	1	1,012	2,052	22	4,182	77	7,346
		71	1	5	1,243	3,737	57	3,161	146	8,349
			2	1	2,808	2,349	21	2,911	84	8,174
			3	6	2,247	4,430	64	3,218	140	10,105
	FAN1+/+	16	1	3	2,974	6,561	289	12,124	265	22,216
			2	3	2,013	3,710	46	3,707	68	9,547
			3	6	2,315	2,795	27	3,969	72	9,184
		37	1	8	2,036	2,713	30	3,488	77	8,352
			2	3	897	3,213	30	3,161	52	7,356
			3	2	1,013	2,860	203	3,172	145	7,395
		50	1	1	1,263	1,767	96	3,470	243	6,840
			2	0	2,035	1,851	55	3,741	60	7,742
			3	1	855	2,472	99	2,475	72	5,974
		71	1	3	1,329	2,745	134	2,665	90	6,966
			2	2	2,255	2,077	74	4,393	101	8,902
			3	3	795	2,074	28	3,020	58	5,978
-	-	-	H2O	0	43	400	10	137	5	595
Total				93	49,186	113,200	3,642	189,220	5,991	361,332

Table 4.4. Read counts of all barcode-paired samples in PacBio library 600-iPSC-4, categorised by RepeatDecoder restrctive profile CAG length and filtered status. Short: <7 CAGs. WT: 7-29 CAGs. Expanded: 30 or more CAGs.

Read filtering is summarised in Figure 2.1. Table 4.4 shows that most 'short' reads were removed by filtering (0.189% retained), and that most of both WT and expanded reads were retained with 96.9% surviving filtering in both alleles. The water control sample has 595 reads associated with it, 67.0% of which are of WT length and retained and 23% are expanded, suggesting there is a low level of cross-contamination between samples.

In summary analyses, results from individual replicates were weighted equally to avoid over-representation of results from samples with high read counts and under-representation of those with low read-counts. For example, in Figure 4.11, percentages of expanded allele reads have been averaged from all the samples per condition.

Figure 4.10 represents the data in Table 4.4 graphically. There is a wide range of the percentage of reads in each CAG length category across samples, with the difference between PCR 1 and PCR 2 samples particularly striking. This is further highlighted in Figure 4.11 columns 1 and 2. There are a greater number of reads overall in the PCR 2 samples (PCR 1: 160,033, PCR 2: 200,704). Read counts of PCR 1 and 2 were not normally distributed in a Shapiro-Wilk test at a 5% significance level. A Mann-Whitney U test was performed and found PCR 1 and 2 read counts were not equal at a 5% significance level ($U = 170$, $p = 0.008$). Total expanded allele read counts were higher in PCR 1 samples (PCR 1: 103,813, PCR 2: 91,256). Expanded allele read counts of PCR 1 were normally distributed in a Shapiro-Wilk test at a 5% significance level but counts of PCR 2 were not. A Mann-Whitney U test was performed and found expanded allele PCR 1 and 2 read counts were not equal at a 5% significance level ($U = 167$, $p = 0.006$).

Table 4.5 and Figure 4.11 shows the extent of the difference in the proportion of expanded allele reads between PCR 1 and 2 samples, with 69.4% of PCR 1 samples and only 45.7% of PCR 2 samples categorised as expanded. The differences in expanded read percentage within the other experimental variables are all less than 10% (Day: 4.7%, cell line: 0.5%, Replicate: 8.6%, chip: 1.9%).

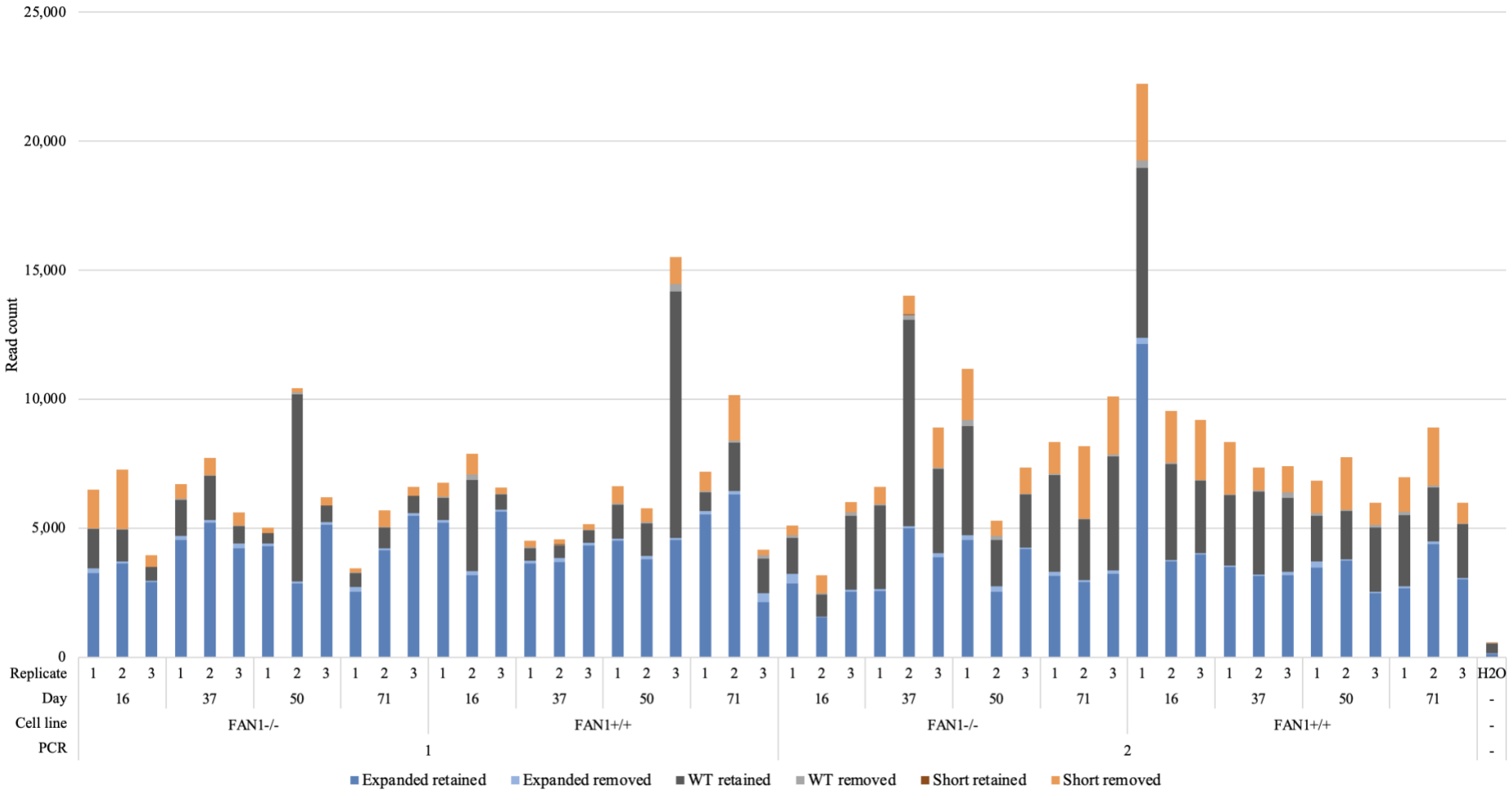


Figure 4.10. Read counts associated with all barcode-paired samples coloured by CAG length and filtered status using data from long-read PacBio sequencing of library 600-iPSC-4. RepeatDecoder restrictive profile count categories; Expanded: 30 or more CAGs, WT: 7-29 CAGs, Short: < 7 CAGs. Retained: survived filtering. Removed: filtered out. See figure 3.6 for filtering details. Replicate: biological replicate. Day: harvest day.

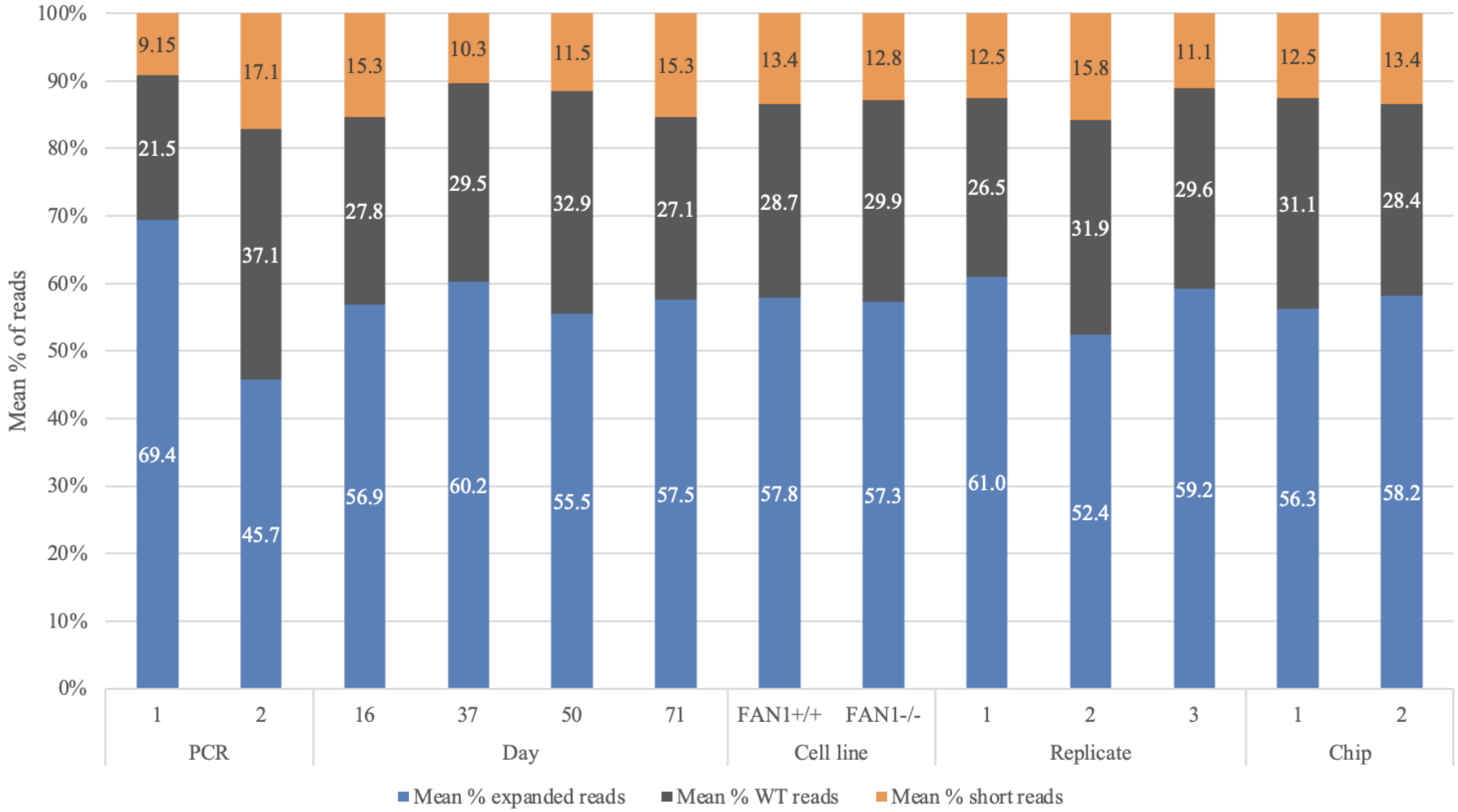


Figure 4.11 Mean normalised read counts in library 600-iPSC-4 by CAG length category, experimental variable and condition. Unfiltered PacBio reads counted by RepeatDecoder restrictive profile. Expanded: 30 or more CAGs WT: 7-29 CAGs. Short: < 7 CAGs. Replicate: biological replicate. Day: harvest day.

Experimental variable	Condition	Mean read count	Mean % short	Mean % WT	Mean % expanded
PCR	1	6668	9.15	21.5	69.4
	2	8363	17.1	37.1	45.7
Day	16	7848	15.3	27.8	56.9
	37	7242	10.3	29.5	60.2
	50	7827	11.5	32.9	55.5
	71	7144	15.3	27.1	57.5
Cell line	*FAN1*+/+	7974	13.4	28.7	57.8
	FAN1-/-	7057	12.8	29.9	57.3
Replicate	1	7648	12.5	26.5	61.0
	2	7730	15.8	31.9	52.4
	3	7168	11.1	29.6	59.2
Chip	1	2568	12.5	31.1	56.3
	2	4947	13.4	28.4	58.2

Table 4.5. Mean read counts and percentage of reads by CAG length category, experimental variable, and condition. Unfiltered PacBio HiFi reads counted by RepeatDecoder restrictive profile. Short: < 7 CAGs. WT: 7-29 CAGs. Expanded: 30 or more CAGs Replicate: biological replicate. Day: harvest day.

Table 4.6 shows the number of reads surviving each stage of filtering by CAG length category. A flow diagram of how reads are filtered is shown in Figure 2.1. 54% of unfiltered reads were classified as 'expanded', 32% 'WT' and 14% 'short'. 99.8% of 'short' reads were removed by the first filter (flanking sequence match). 62.5% of filtered reads were classified as 'expanded', 37.4% 'WT' and 0.031% 'short'.

	Reads			
Filter applied	**short**	**WT**	**expanded**	**Total**
None	49,279	116,842	195,211	361,332
Flanking sequences	105	114,193	193,486	307,784
Chimera	104	114,149	193,338	307,591
Permissive vs restrictive	93	113,200	189,220	302,513

Table 4.6. Reads at each stage of filtering by CAG size classification. Reads less than 7 CAGs, as counted by RepeatDecoder restrictive profile, are classified as 'short', 7-29 'WT' and more than 29 'expanded'. Filters were applied cumulatively, and in the order listed. Flanking sequences: read contains matches to two 12 bp CAG repeat flanking sequences. Chimera: reads do not contain 4 consecutive 'CAG's and 4 consecutive 'CTG's. Permissive vs restrictive: RepeatDecoder permissive count minus RepeatDecoder restrictive count is less than 20. See Figure 3.9 for explanation of counting profiles.

Table 4.7 shows the number of expanded allele reads surviving each filter in each cell line. Reads from knockout line samples represent approximately 47% of all reads at each stage of filtering. This shows that there is a roughly even representation of reads in each cell line and that those reads removed by filtering are evenly distributed between the cell lines.

	Reads			
Filter	*FAN1*$^{+/+}$	*FAN1*$^{-/-}$	Total	% *FAN1*$^{-/-}$
None	103,281	91,788	195,069	47.054
flanked	102,381	90,965	193,346	47.048
non-palindromic	102,288	90,910	193,198	47.056
lt20	100,268	88,815	189,083	46.971

Table 4.7. Expanded allele read filtering by cell line. Filters are applied cumulatively, and in the order listed. flanked: read contains matches to *HTT* flanking sequences. Non-palindromic: reads do not contain 4 consecutive 'CAG's and 4 consecutive 'CTG's anywhere in the sequence. lt20: RepeatDecoder permissive count minus RepeatDecoder restrictive count is less than 20.

Table 4.7 shows the number of reads at each stage of filtering. The proportion of *FAN1*$^{-/-}$ reads stays within 0.1% of the starting percentage after each filter.

4.3.2.6 Analysis of PacBio sequencing data

4.3.2.6.1 Changes in modal CAG and measures of expansion and instability over time

Figure 4.12 shows read frequency distribution plots from a set of representative samples for each cell line (PCR replicate 1, biological replicate 1 in all). Both cell lines exhibit increases in modal CAG over time, with the *FAN1*$^{+/+}$ line undergoing an increase of 1 CAG and the FAN$^{-/-}$ line undergoing an increase of 2 CAGs. Unlike fragment analysis, only integer values for CAG count are given by RD as the algorithm estimates the closes whole number of CAGs. Also unlike fragment analysis, the distributions are not perfectly smooth reflecting the single molecule nature of the technique. Fragment analysis signals are based on fluorescently labelled bulk PCR products and as such are typically based on many more of copies of DNA than current single molecule long-read sequencing approaches generate. As a result, modal peaks are shifted away from the centre of the distribution more often than in fragment analysis data. This effect is most pronounced in samples with the fewest reads. Despite this, Figure 4.12 shows that the CAG length distributions increase progressively over time in both lines.

This progressive expansion over time is also reflected in the day 16-anchored expansion indices (except for *FAN1*$^{+/+}$ day 37), with the FAN$^{+/+}$ line increasing from 1.56 at day 16 to 2.49 at day 71 (increase of 59.6%). As with modal CAG, the day 16-anchored expansion index increases more in the FAN$^{-/-}$ line, starting at 1.18 at day 16 and increasing to 3.18 at day 71 (increase of 169%).

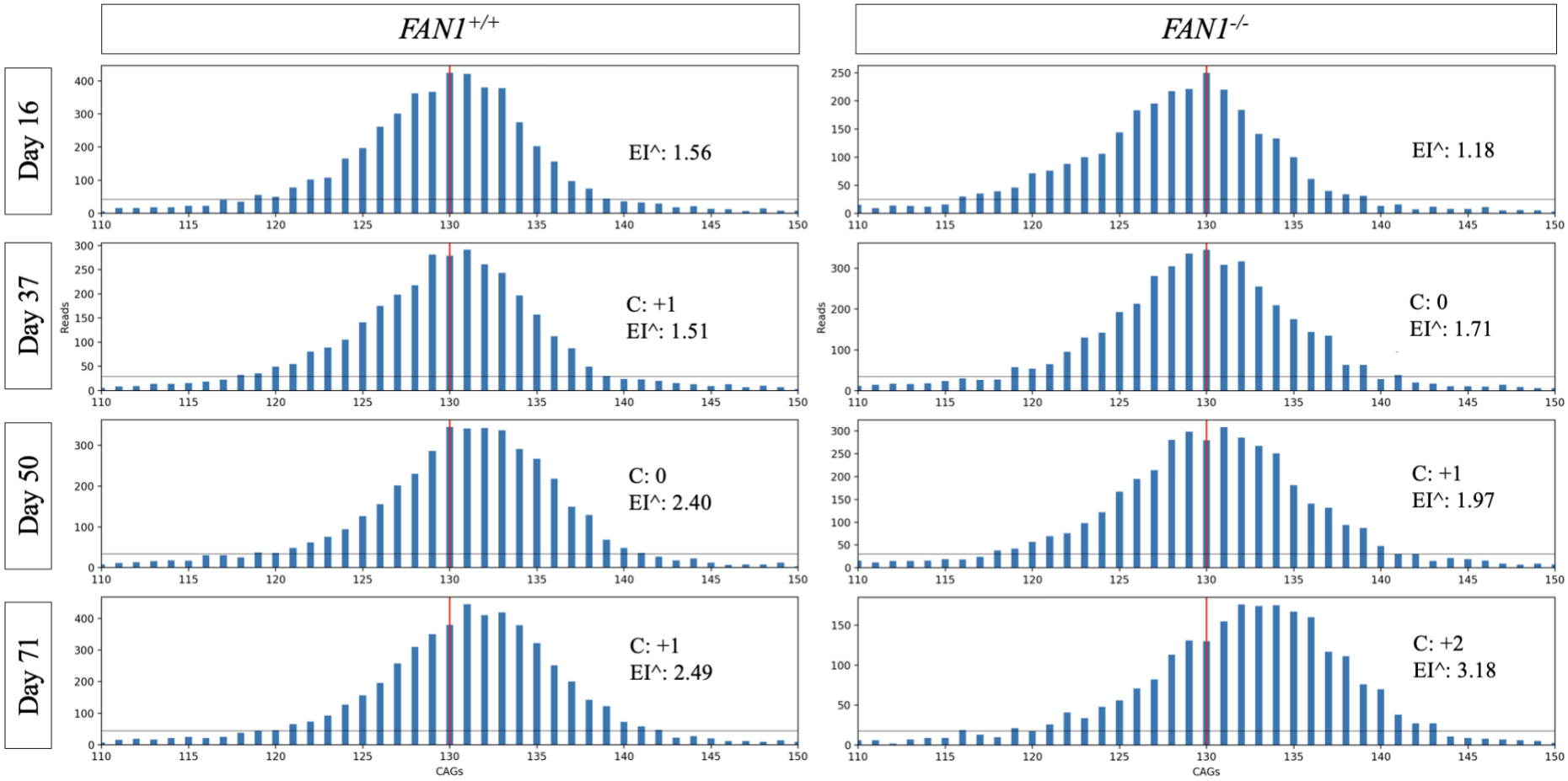


Figure 4.12. Illustration of CAG length distribution plots of PacBio sequencing data of the *HTT* repeat locus in *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ 109NI iPSCs by harvest day. C^: Day 16-anchored change in modal CAG. EI^: Day 16-anchored expansion index. All data is from PCR1, replicate 1. Red lines indicate the modal CAG at day 16. Horizonal black/grey lines represent the 10% of the modal CAG frequency.

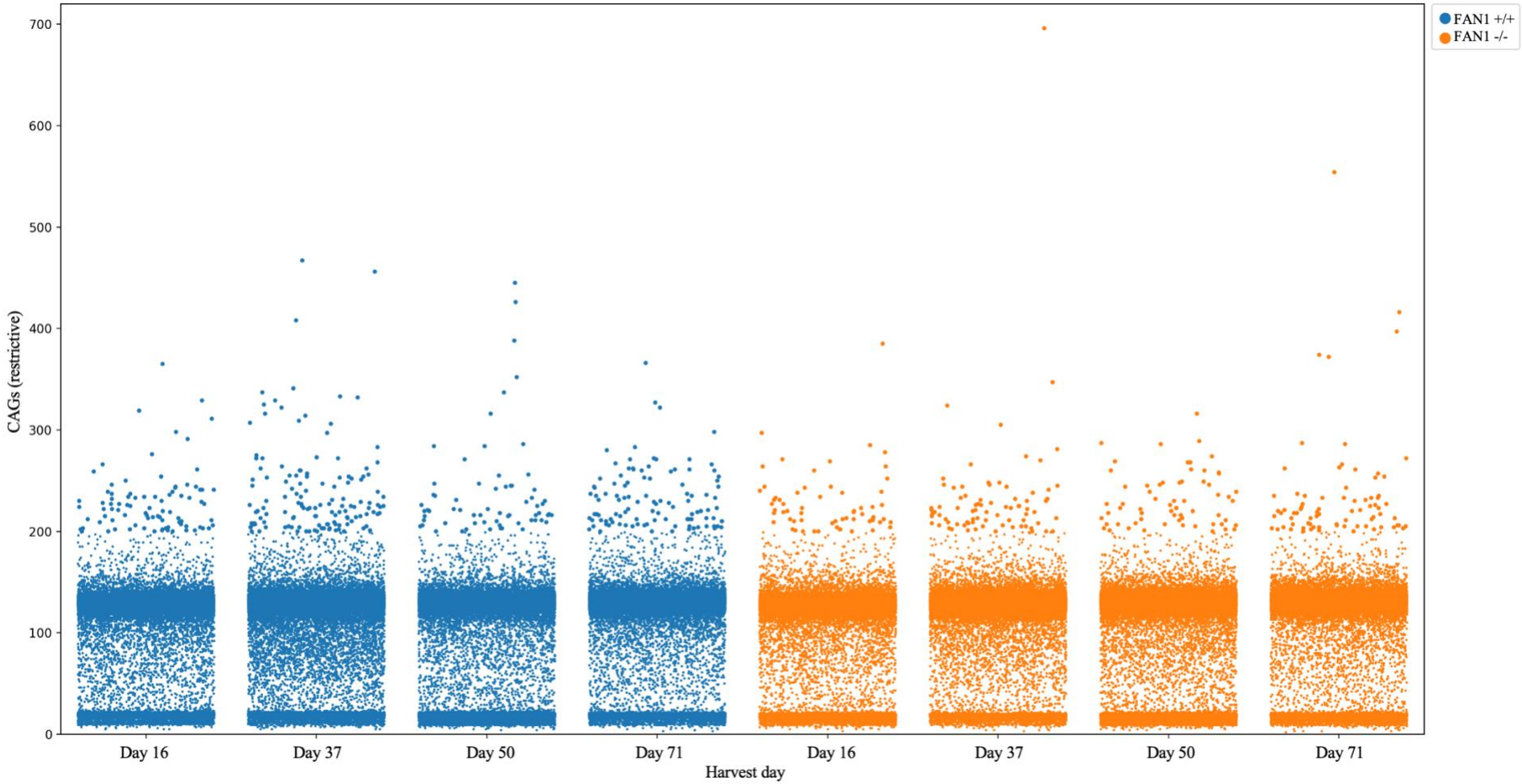


Figure 4.13. CAG length distribution of all filtered PacBio reads of the *HTT* repeat locus in *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ 109NI iPSCs by cell line and harvest day. RepeatDecoder restrictive profile counts. Each dot represents a single read. Dots above 200 CAGs are enlarged for visual clarity.

Figure 4.13 shows the distribution of restrictive CAG counts in all filtered reads from the current experiment, categorised by cell line and harvest day. The *HTT* WT allele is clearly visible in all conditions as a cluster of reads at around 20 CAGs. The primary expanded allele cluster is centred on approximately 130 CAGs in all conditions. Above that, reads appear at diminishing frequency up to 480 CAGs in the *FAN1*$^{+/+}$ line and 696 CAGs in the FAN$^{-/-}$ line.

The CAG distributions shown Figure 4.13 are broadly consistent between cell line and harvest day conditions, although subtle differences are visible. For example, while all samples have a cloud of reads between the WT and expanded allele clusters, *FAN1*$^{+/+}$ day 37 has a particularly dense read count here despite having fewer expanded reads in total compared to the equivalent *FAN1*$^{-/-}$ condition.

Table 4.8 shows CAG repeat distribution summary statistics for PCR 1 and 2. Where modal CAG of the WT allele varied between 17-20 CAGs in the fragment analysis data, PacBio varies between 19 and 20. Day 16 WT allele modal CAG counts agree with Sanger sequencing data for these cell lines at 20 and 19 CAGs for the *FAN1*$^{+/+}$ line and the *FAN1*$^{-/-}$ lines respectively. The *FAN1*$^{+/+}$ line showed a reduction of 1 CAG for nearly all replicates from day 50 and 71 time points, however this is within error guidelines for this repeat length for PCR-based CAG counting methods (Losekoot et al. 2013) and the read frequencies of 19 and 20 CAGs were close in 10 out of the samples which showed a reduction. This is reflected in the relatively high WT allele expansion indices observed compared to samples which showed no reduction.

Where fragment analysis data showed no overall change in the modal CAG of the WT allele in the *FAN1*$^{-/-}$ cells, PacBio data also showed no change. The WT expansion indices are all less than 0.1, which is relatively low compared to the expanded allele (all above 1.0), which is expected from observing sharp WT allele peaks in all *FAN1*$^{-/-}$ samples.

A

				WT allele			Expanded allele					
PCR	Cell line	Day	Rep	Modal CAG	CAG change	Expansion index	Modal CAG	CAG change	Expansion index^	Expansion index	Instability index	Spread
1	*FAN1* +/+	16	1	20	0	0.00	130	0	1.56	1.56	-0.30	3.43
			2	20	0	0.00	130	0	1.73	1.73	-0.01	3.47
			3	20	0	0.06	131	0	0.91	0.91	-1.77	3.58
		37	1	20	0	0.00	131	1	1.51	1.15	-1.41	3.71
			2	20	0	0.07	131	1	1.34	1.01	-1.63	3.65
			3	20	0	0.07	131	0	1.08	1.08	-1.57	3.72
		50	1	19	-1	0.30	130	0	2.40	2.40	1.02	3.77
			2	19	-1	0.28	132	2	2.36	1.52	-0.54	3.59
			3	20	0	0.05	132	1	1.33	1.02	-1.75	3.79
		71	1	19	-1	0.39	131	1	2.49	2.00	0.29	3.71
			2	19	-1	0.31	131	1	2.29	1.81	0.03	3.60
			3	19	-1	0.14	134	3	2.56	1.33	-1.00	3.65
	FAN1 -/-	16	1	19	0	0.07	130	0	1.18	1.18	-1.72	4.09
			2	19	0	0.06	129	0	1.62	1.62	-0.55	3.79
			3	19	0	0.00	129	0	1.45	1.45	-0.87	3.76
		37	1	19	0	0.07	130	0	1.71	1.71	-0.30	3.72
			2	19	0	0.00	131	2	2.02	1.31	-1.51	4.12
			3	19	0	0.00	129	0	2.22	2.22	0.46	3.98
		50	1	19	0	0.06	131	1	1.97	1.57	-0.88	4.01
			2	19	0	0.00	130	1	2.95	2.43	0.68	4.17
			3	19	0	0.00	130	1	2.82	2.29	0.72	3.86
		71	1	19	0	0.00	132	2	3.18	2.17	0.11	4.24
			2	19	0	0.00	133	4	3.07	1.56	-1.28	4.40
			3	19	0	0.00	130	1	3.17	2.63	0.96	4.31
Mean	-	-	-	19.3	-0.21	0.08	130.8	0.92	2.04	1.65	-0.53	3.84

B

				WT allele			Expanded allele					
PCR	Cell line	Day	Rep	Modal CAG	CAG change	Expansion index	Modal CAG	CAG change	Expansion index^	Expansion index	Instability index	Spread
2	*FAN1* +/+	16	1	20	0	0.05	131	0	1.01	1.01	-1.82	3.84
			2	20	0	0.05	130	0	1.41	1.41	-0.68	3.49
			3	20	0	0.00	131	0	1.11	1.11	-1.61	3.83
		37	1	20	0	0.05	130	-1	1.43	1.91	-0.02	3.84
			2	20	0	0.06	130	0	1.74	1.74	-0.18	3.67
			3	20	0	0.00	131	0	1.40	1.40	-1.05	3.85
		50	1	19	-1	0.30	132	1	1.69	1.29	-0.86	3.44
			2	19	-1	0.32	131	1	2.34	1.85	0.20	3.50
			3	19	-1	0.45	133	2	1.06	0.59	-2.86	4.05
		71	1	19	-1	0.30	134	3	1.38	0.69	-3.34	4.71
			2	19	-1	0.30	133	3	1.82	0.91	-2.03	3.85
			3	19	-1	0.11	135	4	2.11	0.92	-2.01	3.86
	FAN1 -/-	16	1	19	0	0.00	129	0	1.62	1.62	-0.58	3.82
			2	19	0	0.00	130	0	1.17	1.17	-1.83	4.17
			3	19	0	0.06	128	0	1.91	1.91	-0.24	4.05
		37	1	19	0	0.00	131	2	1.97	1.26	-1.37	3.88
			2	19	0	0.05	131	1	1.70	1.35	-1.69	4.38
			3	19	0	0.00	130	2	2.78	1.87	-0.21	3.94
		50	1	19	0	0.06	130	1	2.79	2.29	0.23	4.36
			2	19	0	0.00	131	1	2.41	1.96	-0.25	4.16
			3	19	0	0.00	133	5	3.24	1.46	-1.32	4.24
		71	1	19	0	0.06	133	4	3.41	1.67	-0.74	4.09
			2	19	0	0.05	133	3	2.46	1.40	-1.65	4.46
			3	19	0	0.00	132	4	3.23	1.68	-1.34	4.71
Mean	-	-	-	19.3	-0.25	0.09	131.3	1.50	1.97	1.44	-1.14	4.01

Table 4.8. Summary statistics of PacBio sequencing data of the *HTT* CAG repeat in 109NI iPSCs. (A) PCR replicate 1 values. (B) PCR replicate 2 values. WT: wild type. Rep: biological replicate. ^ Day 16-anchored.

Data from Table 4.8 is plotted in Figure 4.14. Figures 4.14A, B and C show how, for the expanded allele, the Modal CAG, day 16-anchored expansion index and Spread

increase progressively over time in both cell lines. The rate of increase in modal CAG over time was not significantly higher in *FAN1*$^{-/-}$ cells compared with *FAN1*$^{+/+}$ cells ($p = 0.75$; 2-way ANOVA). The *FAN1* genotype of the cells explained 1.9% of the variance in modal CAG observed.

There was a trend towards a faster rate of increase of day 16-anchored expansion index over time in the *FAN1*$^{-/-}$ cells (Fig 4.14B) although this did not reach statistical significance ($p = 0.23$; 2-way ANOVA). The *FAN1* genotype of the cells explained 5.0% of the variance in day 16-anchored expansion index observed.

The rate of increase in Spread over time was not significantly higher in *FAN1*$^{-/-}$ cells compared with *FAN1*$^{+/+}$ cells ($p = 0.7$; 2-way ANOVA). The *FAN1* genotype of the cells explained 2.3% of the variance in modal CAG observed.

Figure 4.14C and D show how unanchored expansion and instability indices change over time in both cell lines. For unanchored expansion index there is a positive gradient in the FAN$^{-/-}$ line and a negative gradient in the *FAN1*$^{+/+}$ line but the rate of increase in expansion index over time was not significantly higher in *FAN1*$^{-/-}$ cells compared with *FAN1*$^{+/+}$ cells ($p = 0.60$; 2-way ANOVA). The *FAN1* genotype of the cells explained 6.1% of the variance in modal CAG observed.

Instability index showed a high degree of variation at all time points. Mean data points of the two cell lines consistently overlap each other. There is a positive gradient in the FAN$^{-/-}$ line and a negative gradient in the *FAN1*$^{+/+}$ line but the rate of increase in instability index over time was not significantly higher in *FAN1*$^{-/-}$ cells compared with *FAN1*$^{+/+}$ cells ($p = 0.82$; 2-way ANOVA). The *FAN1* genotype of the cells explained 4.4% of the variance in modal CAG observed.

Overall, while the broad trends in the data reflect those observed in the fragment analysis data (i.e. no difference in rate of modal CAG change between lines, rate of Spread and expansion indices increases more over time in FAN$^{-/-}$ line), there is more noise, meaning none of the differences in rates of change observed between lines is statistically significant. More data or a reduction in noise is needed to confirm these possible interactions in the PacBio data.

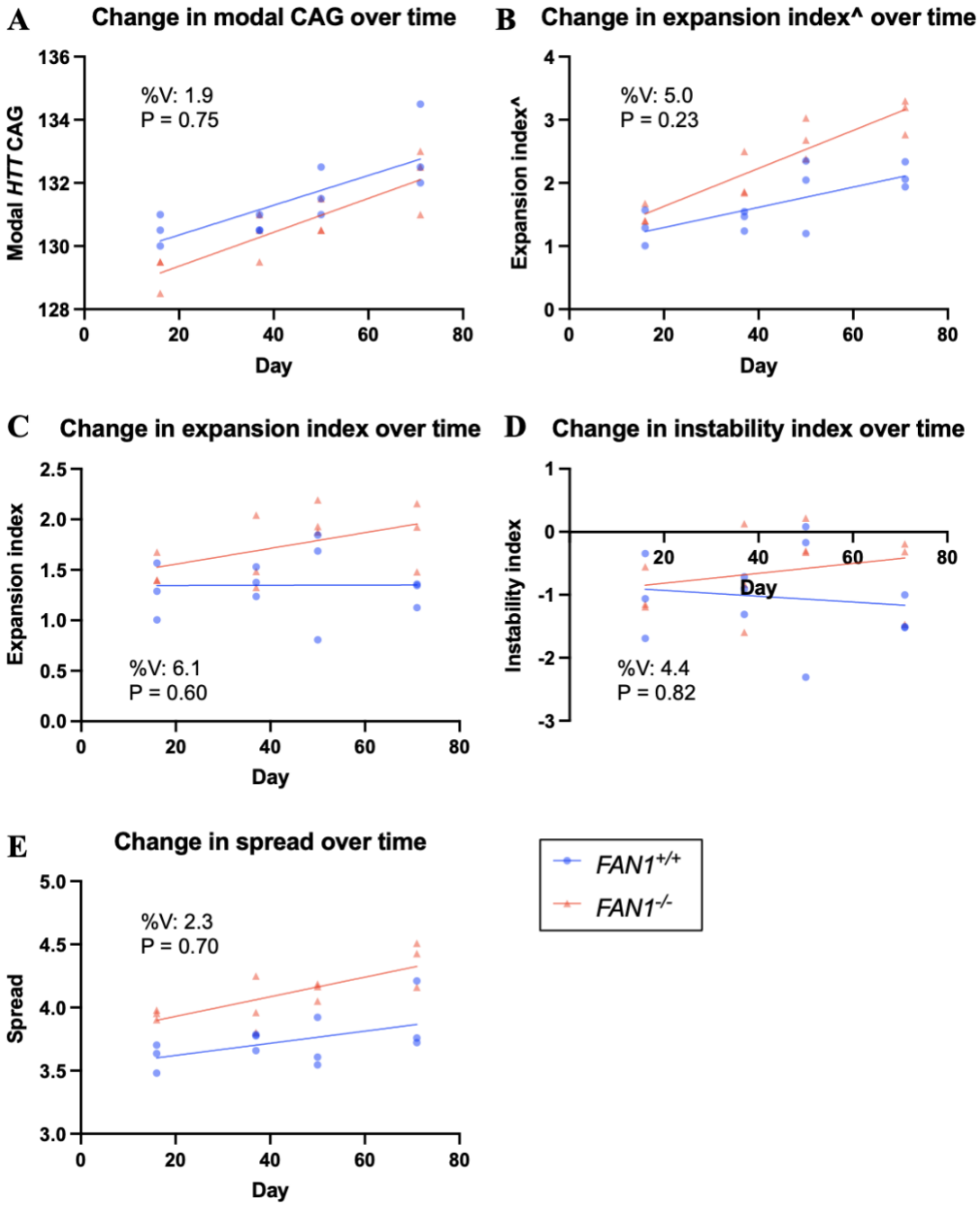


Figure 4.14. Change in modal CAG, day 16-anchored expansion index, expansion index, instability index and Spread over time for the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines using data from long-read PacBio sequencing. RepeatDecoder restrictive profile CAG counts. Circles and triangles represent the mean of PCR replicates for 3 biological replicates per time point per cell line. Lines are simple linear regression lines of best fit. ^: day-16 anchored. %V: percentage of total variation explained by the interaction between harvest day and *FAN1 genotype* in 2-way ANOVA. P: p-value of the interaction between harvest day and *FAN1 genotype* in 2-way ANOVA.

4.3.2.6.2 The distribution of expanded *HTT* CAG repeat lengths

One of the key advantages of long-read PacBio sequencing data in assessing expanded CAG repeats is that it gives you the sequence of very long repeats and the ability to

A

PCR	Cell line	Day	Rep	Modal CAG	Expanded allele reads	Reads < modal	Modal reads	Reads modal +1 to modal +30	Reads > modal +30	Reads < modal (%)	Modal reads (%)	Reads modal +1 to modal +30 (%)	Reads > modal +30 (%)
1	*FAN1* +/+	16	1	130	5,217	2,471	425	2,263	58	47.4	8.15	43.4	1.11
			2	130	3,187	1,611	241	1,288	47	50.5	7.56	40.4	1.47
			3	131	5,642	3,368	497	1,750	27	59.7	8.81	31.0	0.48
		37	1	131	3,641	2,011	291	1,308	31	55.2	7.99	35.9	0.85
			2	131	3,680	2,061	322	1,248	49	56.0	8.75	33.9	1.33
			3	131	4,331	2,497	357	1,438	39	57.7	8.24	33.2	0.90
		50	1	130	4,507	1,741	345	2,377	44	38.6	7.65	52.7	0.98
			2	132	3,804	1,931	313	1,523	37	50.8	8.23	40.0	0.97
			3	132	4,540	2,820	361	1,324	35	62.1	7.95	29.2	0.77
		71	1	131	5,539	2,444	445	2,587	63	44.1	8.03	46.7	1.14
			2	131	6,314	2,863	499	2,874	78	45.3	7.90	45.5	1.24
			3	134	2,131	1,089	171	821	50	51.1	8.02	38.5	2.35
	FAN1 -/-	16	1	130	3,270	1,912	250	1,069	39	58.5	7.65	32.7	1.19
			2	129	3,639	1,822	265	1,520	32	50.1	7.28	41.8	0.88
			3	129	2,911	1,541	231	1,112	27	52.9	7.94	38.2	0.93
		37	1	130	4,531	2,269	345	1,876	41	50.1	7.61	41.4	0.90
			2	131	5,200	2,964	380	1,811	45	57.0	7.31	34.8	0.87
			3	129	4,220	1,850	309	2,006	55	43.8	7.32	47.5	1.30
		50	1	131	4,312	2,278	308	1,689	37	52.8	7.14	39.2	0.86
			2	130	2,862	1,276	197	1,363	26	44.6	6.88	47.6	0.91
			3	130	5,139	2,102	401	2,604	32	40.9	7.80	50.7	0.62
		71	1	132	2,546	1,092	176	1,244	34	42.9	6.91	48.9	1.34
			2	133	4,129	2,225	290	1,566	48	53.9	7.02	37.9	1.16
			3	130	5,477	2,280	376	2,766	55	41.6	6.87	50.5	1.00
Mean	-	-	-	130.8	4,199	2,105	325	1,726	43	50.3	7.71	40.9	1.06

B

PCR	Cell line	Day	Rep	Modal CAG	Expanded allele reads	Reads < modal	Modal reads	Reads modal +1 to modal +30	Reads > modal +30	Reads < modal (%)	Modal reads (%)	Reads modal +1 to modal +30 (%)	Reads > modal +30 (%)
2	*FAN1* +/+	16	1	131	3,172	1,832	249	1,042	49	57.8	7.85	32.8	1.54
			2	130	4,549	2,327	378	1,783	61	51.2	8.31	39.2	1.34
			3	131	2,523	1,441	196	855	31	57.1	7.77	33.9	1.23
		37	1	130	1,539	748	115	644	32	48.6	7.47	41.8	2.08
			2	130	2,534	1,237	200	1,069	28	48.8	7.89	42.2	1.10
			3	131	12,124	6,910	822	4,138	254	57.0	6.78	34.1	2.10
		50	1	132	5,000	2,634	416	1,906	44	52.7	8.32	38.1	0.88
			2	131	3,741	1,674	310	1,718	39	44.7	8.29	45.9	1.04
			3	133	2,475	1,686	182	568	39	68.1	7.35	22.9	1.58
		71	1	134	2,665	1,845	175	601	44	69.2	6.57	22.6	1.65
			2	133	4,393	2,608	368	1,373	44	59.4	8.38	31.3	1.00
			3	135	3,020	1,911	238	832	39	63.3	7.88	27.5	1.29
	FAN1 -/-	16	1	129	3,885	1,952	290	1,587	56	50.2	7.46	40.8	1.44
			2	130	3,488	2,086	244	1,116	42	59.8	7.00	32.0	1.20
			3	128	3,161	1,480	226	1,402	53	46.8	7.15	44.4	1.68
		37	1	131	4,182	2,359	333	1,428	62	56.4	7.96	34.1	1.48
			2	131	3,470	2,015	231	1,178	46	58.1	6.66	33.9	1.33
			3	130	2,869	1,455	198	1,178	38	50.7	6.90	41.1	1.32
		50	1	130	3,707	1,853	236	1,542	76	50.0	6.37	41.6	2.05
			2	131	3,969	1,994	272	1,655	48	50.2	6.85	41.7	1.21
			3	133	2,558	1,439	190	902	27	56.3	7.43	35.3	1.06
		71	1	133	3,161	1,690	218	1,210	43	53.5	6.90	38.3	1.36
			2	133	2,911	1,699	196	981	35	58.4	6.73	33.7	1.20
			3	132	3,218	1,779	203	1,197	39	55.3	6.31	37.2	1.21
Mean	-	-	-	131.3	3,680	2,027	270	1,329	53	55.1	7.36	36.1	1.39

Table 4.9. Modal CAG length, read counts and normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines by CAG size category using data from long-read PacBio sequencing. (A) PCR 1 samples. (B) PCR 2 samples. RepeatDecoder restrictive profile counts. Day: Harvest day. Rep: culture replicate.

Experimental variable	Condition	Mean modal CAG	Mean expanded reads	Mean % of reads < modal	Mean % of modal reads	Mean % of reads modal +1 to modal +30	Mean % of reads > modal +30	Mean % of reads > modal
Cell line	*FAN1*+/+	131.5	4178	54.0	7.92	36.8	1.27	38.1
	FAN1-/-	130.6	3701	51.4	7.14	40.2	1.19	41.4
Harvest day	16	129.8	3720	53.5	7.74	37.6	1.21	38.8
	37	130.5	4360	53.3	7.57	37.8	1.30	39.1
	50	131.3	3885	51.0	7.52	40.4	1.08	41.5
	71	132.6	3792	53.2	7.29	38.2	1.33	39.5
Replicate	1	130.9	3805	51.7	7.50	39.4	1.30	40.7
	2	131.0	3867	52.4	7.57	38.9	1.14	40.0
	3	131.2	4146	54.0	7.53	37.2	1.24	38.4
PCR	1	130.8	4199	50.3	7.71	40.9	1.06	42.0
	2	131.3	3680	55.1	7.36	36.1	1.39	37.5
Chip	1	-	1309*	56.5	7.44	34.9	1.14	36.1
	2	-	2630*	50.8	7.57	40.4	1.28	41.7

Table 4.10. Mean modal CAG length, mean read counts and mean normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines by CAG size category and experimental condition using data from long-read PacBio sequencing. RepeatDecoder restrictive profile counts. *Mean from all 48 samples.

count them, unlike PCR-electrophoresis methods for which the number of molecules is correlated to the signal only, making it impossible to distinguish between target DNA and background noise in regions of low signal levels. Sequencing allows for a much more detailed and quantitative investigation of these areas, which may play an important role in disease pathogenesis, therefore I went on to examine the distribution of *HTT* CAG repeat lengths in the PacBio data.

Table 4.9 shows the distribution of CAG repeat lengths across all 48 non-control samples sequenced. Category cut-offs were chosen to highlight the clusters of expanded allele reads I was most interested in, specifically the expansions and the very large expansions, being represented by the 'modal +1 to modal +30' and 'greater than modal +30' categories respectively. The modal reads category acts to separate the expanded and contracted alleles and as a reference point, and the 'less than modal CAG' category represents the shorter alleles in the analysis.

Table 4.10 is a summary of the data in Table 4.9, grouping the samples into the variables present in the current experiment, and showing the mean percentages of reads by CAG length category of all the samples per condition, which are plotted in Figure 4.15.

As seen previously in Table 4.5, the greatest variability in the mean number of reads per sample comes from the chip (sequencing run), with chip 2 counts 101% greater than chip 1. The conditions of the remaining variables all have mean read counts within 600 reads of each other. The range of mean percentages of shorter reads is 50.3-56.5% in all experimental conditions. For modal reads the range is 7.14-7.92% in all conditions. Sequencing run has the largest effect on the proportion of mean percentage of reads with a greater than modal CAG repeat (difference between chip 1 and 2 is 5.6%), followed by PCR (difference 4.5%) and cell line (3.3%). Harvest day and biological replicate have the smallest effect with ranges of 2.7% and 2.3% respectively. Of the repeats greater than the modal CAG +30, PCR has the largest effect (difference of 0.33%), followed by harvest day (0.25%), replicate (0.16%), chip (0.14%) and, finally, cell line (0.08%).

Despite PCR 1 having 4.5% more reads greater than the modal CAG overall compared to PCR 2, PCR 2 has the greater proportion of reads greater than the modal CAG +30 (1.39% compared to 1.06%). And, despite the *FAN1*$^{-/-}$ line having 3.3% more expanded reads overall compared to the *FAN1*$^{+/+}$ line, the *FAN1*$^{+/+}$ line has the greater proportion of reads greater than the modal CAG +30 (1.27% compared to 1.19%), although the difference is smaller than that seen between PCR1 and 2.

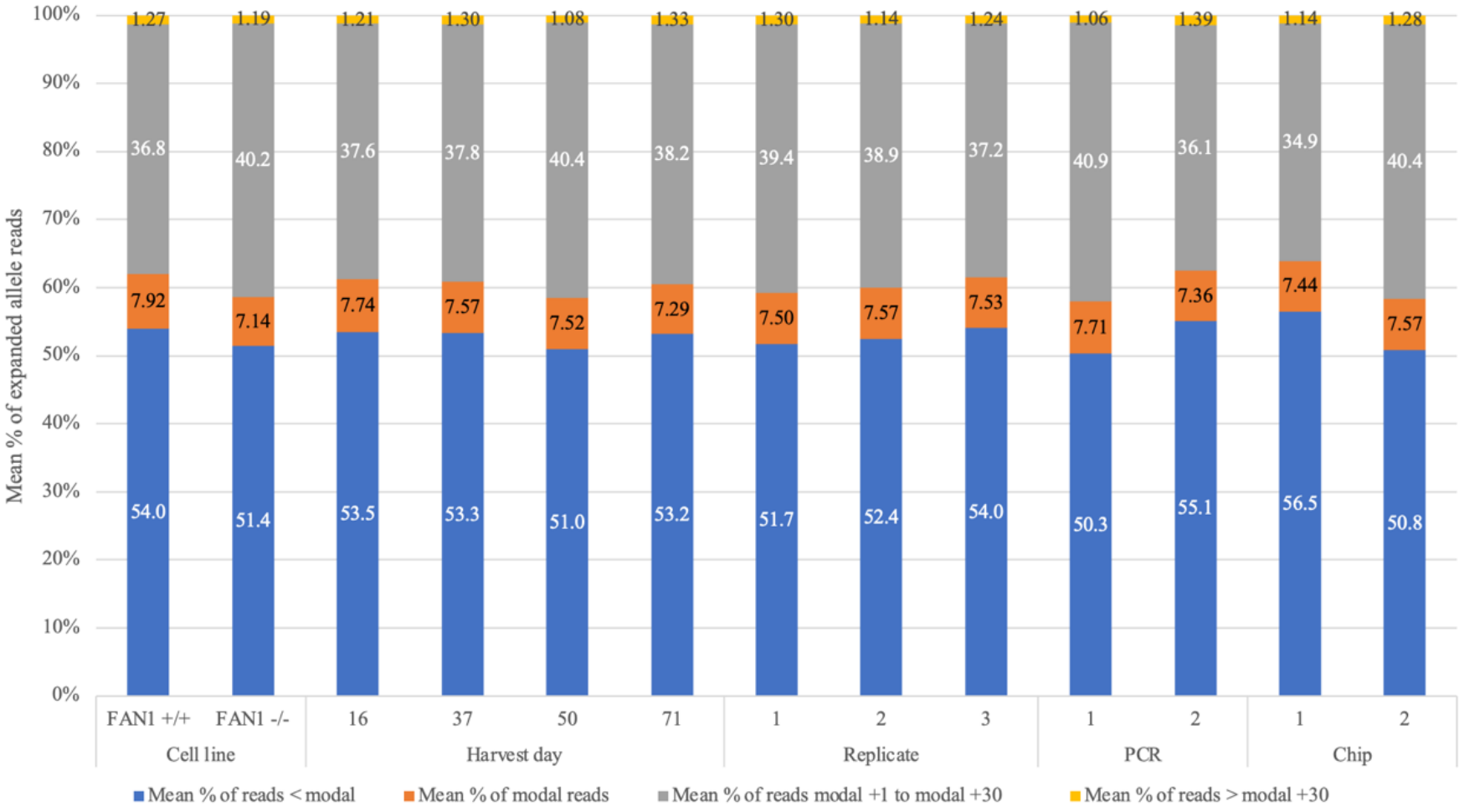


Figure 4.15. Mean normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines by CAG size category and experimental condition using data from long-read PacBio sequencing. RepeatDecoder restrictive profile CAG counts. Expanded read: 30 or more CAGs.

Table 4.11 shows averaged CAG repeat length distributions of samples by cell line and harvest day. These data are plotted in Figure 4.16. In the *FAN1*$^{+/+}$ line, the proportion of reads greater than the modal CAG increases progressively from day 16 to day 50 (38.0 to 39.2%) but then decreases at day 71 to 36.8%. The same trend is seen in the *FAN1*$^{-/-}$ line with CAG increasing from 39.5% at day 16 to 43.8% at day 50, although the day 71 decrease is smaller than the *FAN1*$^{+/+}$ line (1.5 vs 2.4%). The overall change in the proportion of reads longer than the modal CAG from day 16 to 71 is a decrease of 1.2% in the *FAN1*$^{+/+}$ line and an increase of 2.8% in the *FAN1*$^{-/-}$ line. The overall change in the proportion of reads longer than the modal CAG +30 from day 16 to 71 is an increase of 0.25% in the *FAN1*$^{+/+}$ line and a decrease of 0.01% in the *FAN1*$^{-/-}$ line.

Cell line	Day	Mean modal CAG	Mean expanded reads	Mean % of reads < modal	Mean % of modal reads	Mean % of reads modal +1 to modal +30	Mean % of reads > modal +30	Mean % of reads > modal
FAN1 +/+	16	130.5	3392	53.9	8.07	36.8	1.20	38.0
	37	130.7	4048	53.9	7.86	36.9	1.39	38.3
	50	131.7	4079	52.8	7.97	38.2	1.04	39.2
	71	133.0	4642	55.4	7.80	35.4	1.44	36.8
FAN1 -/-	16	129.2	3758	53.1	7.41	38.3	1.22	39.5
	37	130.3	4011	52.7	7.29	38.8	1.20	40.0
	50	130.8	3574	49.1	7.08	42.7	1.12	43.8
	71	132.2	4010	50.9	6.79	41.1	1.21	42.3
Mean	-	131.0	3,939	52.7	7.53	38.5	1.23	39.7

Table 4.11. Mean modal CAG, mean reads and mean normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines by CAG size category, cell line and harvest day using data from long-read PacBio sequencing. RepeatDecoder restrictive profile CAG counts. Expanded read: 30 or more CAGs. Day: harvest day.

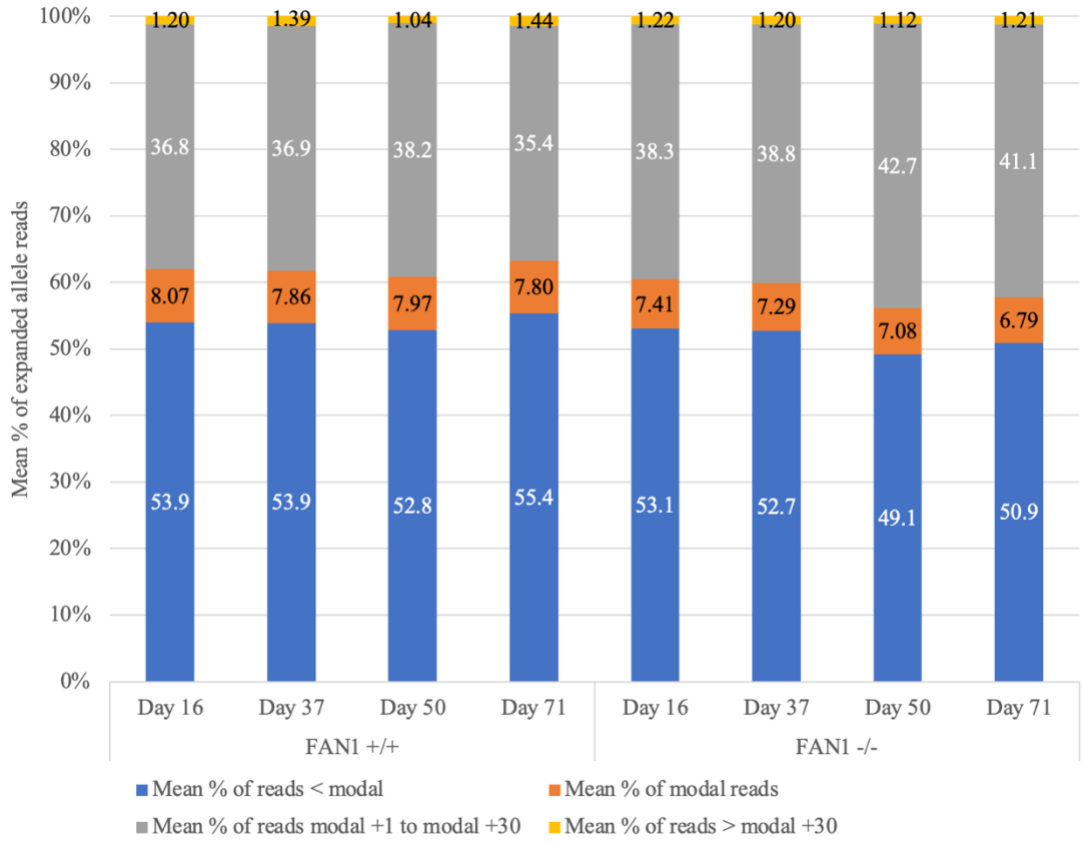


Figure 4.16. Mean normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines by CAG size category, cell line and harvest day using data from long-read PacBio sequencing. RepeatDecoder restrictive profile CAG counts. Expanded read: 30 or more CAGs. Day: harvest day.

Table 4.12 shows averaged CAG repeat length distributions of samples by cell line and culture replicate. This data is plotted in Figure 4.17. The range of the mean normalised reads longer than the modal CAG is 7.9% in the *FAN1*$^{+/+}$ line and 5.2% in the *FAN1*$^{-/-}$ line, both of which are larger than the ranges seen for harvest day (2.4 and 4.3% for the *FAN1*$^{+/+}$ and $^{-/-}$ lines respectively). The ranges of mean proportion of reads greater than the modal CAG +30 for the *FAN1*$^{+/+}$ line is greater across harvest days than replicates (0.44 vs 0.15%), but for the *FAN1*$^{-/-}$ line is greater across replicates than harvest days (0.24 vs 0.10%).

Cell line	Rep	Mean modal CAG	Mean expanded reads	Mean % of reads < modal	Mean % of modal reads	Mean % of reads modal +1 to modal +30	Mean % of reads > modal +30	Mean % of reads > modal
FAN1 +/+	1	131.1	3910	51.7	7.75	39.3	1.28	40.5
	2	131.0	4025	50.8	8.16	39.8	1.19	41.0
	3	132.3	4598	59.5	7.85	31.3	1.34	32.6
FAN1 -/-	1	130.8	3699	51.8	7.25	39.6	1.33	41.0
	2	130.8	3709	54.0	6.97	37.9	1.09	39.0
	3	130.1	3694	48.5	7.21	43.1	1.14	44.2
Mean	-	131.0	3939.2	52.7	7.53	38.5	1.23	39.7

Table 4.12. Mean modal CAG, mean reads and mean normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines by CAG size category, cell line and culture replicate using data from long-read PacBio sequencing. RepeatDecoder restrictive profile CAG counts. Expanded read: 30 or more CAGs. Day: harvest day.

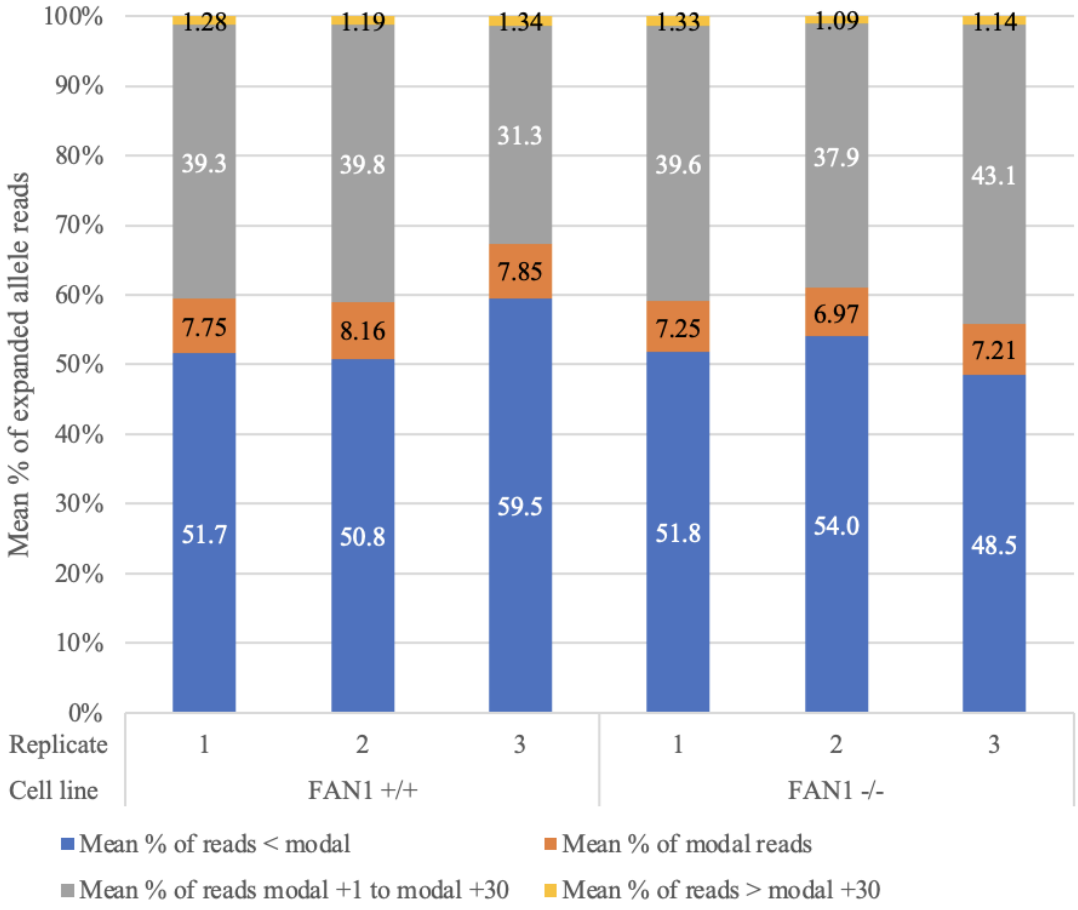


Figure 4.17 Mean normalised read counts of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cell lines by CAG length category, cell line and biological replicate. PacBio Hifi reads of the *HTT* CAG repeat of 109NI iPSCs. RepeatDecoder restrictive profile CAG counts. Expanded: 30 or more CAGs.

4.3.2.7 Comparison of PacBio and fragment analysis CAG sizing

HTT CAG repeat sizing was conducted in parallel by fragment analysis and PacBio sequencing primarily to validate the PacBio data against a known reliable repeat sizing method, but also to give the data more context and to see how the two platforms

compare on identical samples. To this end, Table 4.13 shows a summary comparison of all 48 samples in the current experiment. In both instances, the length of the uninterrupted CAG tract has been calculated.

Comparison	**Modal CAG**		**Expansion Index^**	
Allele	WT	Expanded	WT	Expanded
N	48	48	48	48
Mean FA	18.6	127	0.130	1.90
Mean PB	19.3	131	0.022	2.00
SD FA	0.900	1.42	0.260	0.652
SD PB	0.449	1.47	0.0285	0.701
Normal	No	No	No	No
r_s	-0.751	0.653	N/A	0.836
p-value	$7.9x10^{-10}$	$4.9x10^{-7}$	N/A	$1.4x10^{-13}$

Table 4.13. Comparison of fragment analysis and PacBio-RepeatDecoder calls of the *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neurons. PacBio library 600-iPSC-4. ^ day 16-anchored. WT: wild type. N represents the number of samples analysed. FA: fragment analysis, PB: PacBio-RepeatDecoder. Normal: 'Yes' indicates that neither data set deviates from a Normal distribution in a Shapiro-Wilk test at a 5% significance level. r_s is the Spearman's rank correlation coefficient. p-values derived from a 2-tailed t-test of the correlation coefficient. The mean number PacBio reads with the modal CAG was 5555 for the WT allele, and 175 for the expanded allele.

Mean modal CAG is greater in PacBio than fragment analysis in both the WT and expanded alleles, with a ratio 1.038 in the WT and 1.031 in the expanded allele. This suggests there may be a linear relationship in the relative sizes given over CAG length between the two platforms and therefore a systematic difference in the CAG counts they produce. While the WT modal CAG values are highly significantly negatively correlated, the range of values (plotted in Figure 4.18A) is 1 for the PacBio and 2.3 for the genescan, with almost half of the data centred on a single point. In this scenario, a platform error of +/- 1 CAG can explain negative correlations observed. The expanded allele modal CAG (plotted in Figure 4.18B), by contrast, has a range of 7 for the PacBio and 6.4 for the fragment analysis and the data is more evenly spread, thus the highly significant positive correlation observed ($r_s = 0.653$, $p = 4.9x10^{-7}$) is much more reliable.

16 out of 48 of the WT allele expansion indices are 0 for both the PacBio and fragment analysis (plotted in Figure 4.18C), meaning it was not possible to perform a spearman's rank correlation on this data. PacBio and fragment analysis day 16-anchored expansion indices for the expanded alleles (plotted in Figure 4.18D) were highly significantly positively correlated ($r_s = 0.836$, $P = 1.\ 4x10^{-13}$).

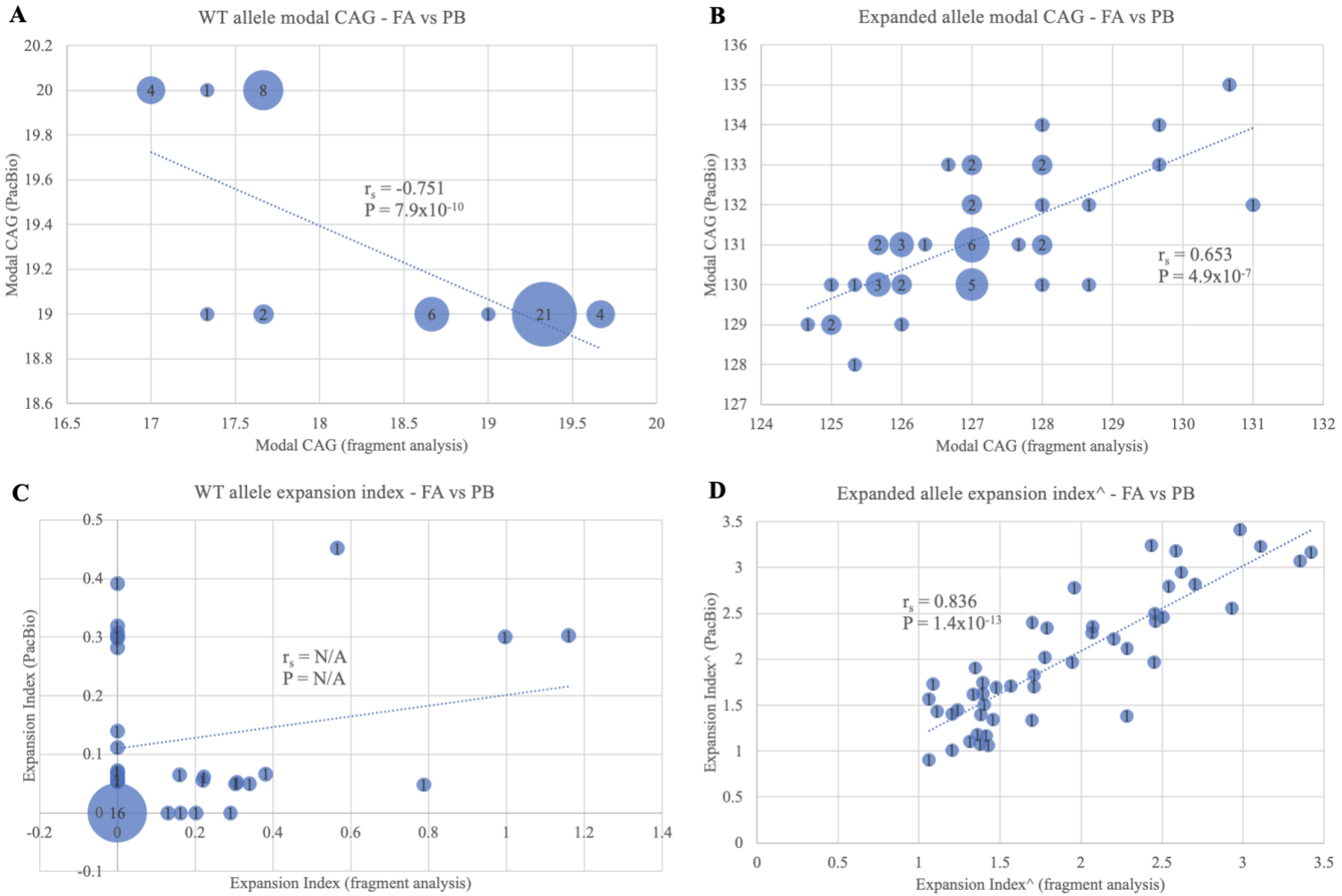


Figure 4.18. Comparison of fragment analysis-Autogenescan and PacBio-RepeatDecoder calls of the *HTT* repeat locus in 600-iPSC-4 library samples. Modal CAG points are jittered randomly between -0.25 and +0.25 on both the x and y axes. rs: Spearman's rank correlation coefficient. P: p-value. Dashed line: line of best fit.

4.3.3 CAG repeat flanking sequence alteration analysis

Changes in the expanded *HTT* CAG repeat flanking sequences were determined as described in section 3.3.1.3.3, whereby the sequence between the 3' end of the restrictive and permissive profile repeat sequences is given as the flanking sequence 3' to the uninterrupted *HTT* CAG repeat. Canonically, this has the sequence "CAACAG" and represents the transition between the uninterrupted CAG repeat and the polyproline repeat, or $(CCGCCA)_x$-$(CCG)_y$-$(CCT)_z$ sequence. Sequencing data from Ciosi et al. 2019, shows that more than 95% of HD patients expanded alleles exhibit a single CAACAG flanking sequence. The same study identified several alterations to this sequence, of which, duplication of CAACAG sequence was the most common (~1-3% of expanded alleles), followed by loss of the CAA (~1% of expanded alleles). CAACAACAG was observed once.

Cells grown for the current experiment were derived from a patient with an expanded allele with the canonical flanking sequence. Table 4.14 shows the 15 most common flanking sequences observed in all expanded allele reads of the *FAN1*$^{+/+}$ and $^{-/-}$ cell lines. 86% of the reads of both cell lines have a typical flanking sequence. Loss of the CAA flanking sequence is the next most common sequence observed and is seen at a rate of 1.82% in the *FAN1*$^{+/+}$ line and 1.80% in the $^{-/-}$ line. The next most frequently observed change in the flanking sequence in both lines is "C", which is associated with no single consistent downstream sequence (Appendix 1B). A gain of one "CAACAG" with respect to the canonical flanking sequence is the next most common sequence at 0.6% in both lines. "CAACAACAG" is the next most common and present at roughly 0.5% in both lines. The remaining flanking sequences appear at less than 0.3% each. 13 out of the 15 most common flanks are shared across cell lines. Of these, none of the rates observed differ by more than 0.1% between the two cell lines.

Only the 15 most frequent flanking sequences are shown in Table 4.14. In all, 1,567 unique flanking sequences were identified in 100,268 *FAN1*$^{+/+}$ reads, the majority of which 1,160 (77.6%), were observed once. 1,495 unique flanking sequences were identified in 88,815 *FAN1*$^{-/-}$ reads, 1,088 (72.8%) of which were observed once.

A ***FAN1*$^{+/+}$**

	Flanking sequence	Reads	% of total
1	CAACAG	86422	86.19
2	Loss of CAA	1826	1.82
3	C	1824	1.82
4	CAACAGCAACAG	600	0.60
5	CAACAACAG	525	0.52
6	CAACAGCAGCAACAG	234	0.23
7	CAAGCAGCAACAG	229	0.23
8	CAAGCAG	223	0.22
9	GCAACAG	190	0.19
10	CCAGCAACAG	187	0.19
11	CAAGCAACAG	181	0.18
12	CAAGCAGCAGCAACAG	174	0.17
13	CAAGCAGCAGCAGCAACAG	164	0.16
14	CAACAGCCGC	163	0.16
15	CAACA	157	0.16

B ***FAN1*$^{-/-}$**

	Flanking sequence	Reads	% of total
1	CAACAG	76470	86.10
2	Loss of CAA	1603	1.80
3	C	1554	1.75
4	CAACAGCAACAG	539	0.61
5	CAACAACAG	428	0.48
6	CAAGCAG	233	0.26
7	CCAGCAACAG	208	0.23
8	CAACAGCAGCAACAG	201	0.23
9	CAAGCAGCAACAG	186	0.21
10	CAAGCAGCAGCAACAG	185	0.21
11	CAACAGCCGC	159	0.18
12	CAAGCAGCAGCAGCAACAG	155	0.17
13	CAAGCAACAG	150	0.17
14	CCAGCAGCAACAG	141	0.16
15	CCAG	140	0.16

Table 4.14. Read counts and normalised read counts of flanking sequences immediately downstream of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neurons using data from long-read PacBio sequencing. (A) *FAN1*$^{+/+}$ line. (B) *FAN1*$^{-/-}$ line. Flanking sequence: sequence between the 3' ends of RepeatDecoder restrictive and permissive profile repeat sequences. Expanded alleles are those with restrictive profile counts of 30 or more. The 15 most frequent flanking sequences are shown. Sequences highlighted yellow appear in both tables and share the same ranking. Sequences highlighted in grey appear in both tables but do not share the same ranking.

Nearly all the flanking sequences observed once take the format $(CAG)_{0\text{-}17}$CAACAG with 1 to 5 substitutions/indels. Note that 19 CAGs is the maximum flanking sequence length as the third read filter (Figure 2.1) removes all reads with a permissive minus restrictive count of more than 19.

Read ID	CAGs	Condition	Sequence	Sequence Q-scores
4194375	128	FAN1+/+ d71 r2 p1	**CAGCAG**CAACAGCCGCCACCG	**[31, 27, 50, 54, 65, 65]**[57, 50, 93, 63, 53, 56][43, 76, 75, 45, 90, 58, 55, 76, 67]
4194376	142	FAN1−/− d50 r1 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194378	132	FAN1−/− d50 r1 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194408	133	FAN1+/+ d50 r1 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194411	130	FAN1−/− d71 r3 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194413	131	FAN1+/+ d37 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194426	126	FAN1−/− d50 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[71, 56, 93, 66, 91, 93]**[84, 93, 93, 24, 28, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194432	158	FAN1+/+ d37 r3 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194434	126	FAN1−/− d37 r3 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194441	130	FAN1+/+ d37 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 90, 93, 93]**[87, 32, 65, 57, 89, 61][38, 93, 93, 10, 90, 87, 75, 93, 93]
4194446	136	FAN1−/− d71 r1 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194448	129	FAN1+/+ d37 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194450	132	FAN1+/+ d50 r1 p2	**CAGCAG**CAACAGCCGCCACCG	**[75, 72, 66, 73, 75, 47]**[67, 40, 70, 86, 82, 76][62, 79, 70, 80, 63, 66, 62, 83, 77]
4194469	136	FAN1+/+ d71 r1 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194470	132	FAN1+/+ d37 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 43, 69, 93]
4194471	129	FAN1+/+ d71 r2 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 51][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194480	120	FAN1−/− d37 r2 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194489	132	FAN1+/+ d16 r1 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194494	134	FAN1−/− d71 r2 p1	**CAGCAG**CAACAGCCGCCACTG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194498	120	FAN1−/− d16 r2 p1	**CAGCAG**CAACAGCCGCCACCG	**[81, 43, 86, 93, 93, 53]**[78, 65, 93, 90, 93, 93][77, 93, 93, 70, 93, 91, 66, 85, 56]
4194499	132	FAN1−/− d16 r2 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 77, 93, 93]
4194504	135	FAN1+/+ d37 r2 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 91, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194515	127	FAN1+/+ d71 r3 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194517	106	FAN1+/+ d37 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 65, 93, 93]
4194545	126	FAN1+/+ d50 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194549	86	FAN1−/− d71 r3 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194560	130	FAN1−/− d71 r1 p1	**CAGCAG**CAACAGCCGCCACCG	**[82, 57, 93, 93, 93, 93]**[93, 62, 93, 88, 74, 93][62, 93, 93, 61, 93, 93, 56, 93, 62]
4194575	130	FAN1+/+ d16 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4194577	128	FAN1+/+ d37 r3 p2	**CAGCAG**CAACAGCCGCCACCG	**[42, 45, 41, 58, 54, 69]**[55, 66, 61, 48, 57, 67][58, 49, 47, 62, 67, 77, 56, 75, 71]
4194587	135	FAN1−/− d71 r1 p1	**CAGCAG**CAACAGCCGCCACCG	**[93, 83, 93, 93, 93, 93]**[93, 39, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]

Table 4.15. Flanking sequence windows and per-base Phred quality scores of 30 randomly selected reads with a "CAACAG" flanking sequence immediately downstream of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neurons using data from long-read PacBio sequencing. CAGs counted by RepeatDecoder restrictive profile. Flanking sequence, coloured blue, is the region between the 3' ends of the restrictive and permissive profiles. Q-score: Phred quality score. Q-scores range from 0 to 93 (see table 1.1). Condition takes the format: cell line, d: harvest day, r: biological replicate, p: PCR replicate.

To assess the changes in the flanking sequence I examined some of the reads associated with the canonical and other sequences in more detail, looking at the per-base quality scores associated with each to establish the likelihood that these sequences are artefacts of sequencing.

Scores are Phred quality scores (Q-scores), representing the log-likelihood of a basecall being correct. 0 is the minimum and 93 is the maximum per-base quality score. A score of 20 equates to 99% predicated accuracy, i.e., 1 error per 100 bases. See section 1.6.2 for more information about Q-scores.

Table 4.15 shows flanking sequence windows of 30 randomly selected reads with the flanking sequence "CAACAG". All but one read matches the canonical sequence over the whole window, read 4194494, which has a penultimate substitution but maximum quality scores from start to finish. Most reads have consistently high scores, although there are several examples with slightly lower scores, e.g. read 4194375, and several have occasional lower scores, e.g. read 4194560, however all scores observed are greater than 20.

Appendix 1A shows flanking sequence windows of 30 randomly selected reads a with no CAACAG sequence. 66% of reads match the canonical sequence minus the "CAACAG" over the whole window. If these were canonical except for an A to G substitution error at position 3, the scores at position 3 should be low. Indeed, 2 of the position 3 scores are lower than 15 (reads 4326074 and 4784851), and could conceivably be sequencing errors, however, the rest are all above 35 and 66% are 93 (maximum score) and therefore very unlikely to be sequencing errors. Of the 10 sequences which deviate from the canonical minus "CAACAG" structure, most could be a canonical read with one or two substitutions or indels, e.g., "CAGCAGCGACAGCCG" (3 occurrences) would be canonical with an A at position 8.

Appendix 1B shows flanking sequence windows of 30 randomly selected reads with the flanking sequence "C". Most of the sequences observed are one substitution/indel away from a perfect canonical sequence, although there is no consistently false position/error mode. Quality scores occasionally dip at positions where reversing such substitutions/indels would restore the canonical sequence but considering that a Phred

score of 20 is has a predicted accuracy of 99%, only 10 are likely to be sequencing errors.

Appendix 1C shows flanking sequence windows of 30 randomly selected reads with the flanking sequence "CAACAGCAACAG". Most reads have the perfect canonical sequence plus a "CAACAG". Again, if these were the result of a consistent sequencing error of the penultimate pure CAG, position 9 should have lower quality scores. However, this is not the case, with most having a maximum score at this position. Occurring at a rate of 0.61% overall suggests these alterations are occurring prior to sequencing.

All 30 randomly selected reads with "CAACAACAG" match perfectly the canonical sequence plus an additional CAA prior to the "CAACAG" across the entire flanking sequence window (see Appendix 1D). Overall, quality scores are very high: 4 reads have one or more scores below 10 within the CAACAACAG.

Most of the 30 randomly selected reads with a "CAAGCAG" match perfectly the canonical sequence plus a G insertion after the "CAA" across the entire flanking sequence window (see Appendix 1E). Of those that don't, two appear to have a CCG deletion immediately before the CCA. The remainder have a mix of sequence changes to the right of the CAAGCAG. 10 of the 30 reads have quality scores lower than 10 within the "CAAGCAG", with 6 occurring at the G after the CAA. There are 12 scores lower than 20 at this position. It is likely that some of these are sequencing errors.

Overall, there are many substitutions/indels within or immediately after the flanking sequence, some of which are likely to be PacBio sequencing errors but most of which are not. The latter may represent sequence alterations in cells but are more likely to have arisen during the PCR required prior to sequencing.

Coding	Flanking sequence	Overall %	
Canonical	CAACAG	86.1	
Loss	""	1.81	
Gain Includes any flanking sequence with no frameshift indels with a read frequency > 0.1%.	CAACAGCAACAG	0.602	Sum total: 2.03%
	CAACAACAG	0.504	
	CAACAGCAGCAACAG	0.230	
	CAACAGCAGCAGCAACAG	0.136	
	CGGCAACAG	0.125	
	CGGCAGCAACAG	0.114	
	CGGCAGCAGCAACAG	0.113	
	CAACAGCAGCAGCAGCAACAG	0.102	
	CGGCAGCAGCAGCAGCAACAG	0.101	
Other	Everything else	10.1	

Table 4.16. Flanking sequence category coding. Flanking sequence: the string of bases that occur after the pure CAG repeat and before the proline repeat. Detected by subtracting the RepeatDecoder restrictive profile sequence from the permissive profile sequence. Overall percentage is from all expanded allele reads.

To investigate whether there was any systematic difference in types of flanking sequence observed, they were categorised by alteration type for subsequent analysis. Table 4.16 shows the coding used and overall frequency observed in expanded allele reads. A 0.1% frequency threshold was applied to the "Gain" category as it included the most common alterations observed in HD patients while limiting the number that have likely arisen by chance. In addition, none of the common alterations observed in patients are frameshift-inducing indels, so these were included in the 'Other' category, including those reads with the flanking sequence "C".

While it is unclear when or how these alterations have occurred, it is interesting that the most common alterations in HD patients (CAACAG duplication, CAA loss, CAACAACAG) are also the most common alterations in individual reads and are at a considerably higher rate than the rest. Of the other alterations observed by exome sequencing, $CAC(CAG)_3CAA$ and CAACAACAA are observed in the data but at less than 0.005% each.

To establish whether altered interruption structure is associated with altered CAG length, I examined the frequency of reads categorised by flanking sequence, CAG repeat length and cell line (Table 4.17). I decided to remove reads with fewer CAGs than the modal CAG from this analysis as it is difficult to distinguish between true contractions that occurred in cells and PCR stutter/counting artefacts and there is lots of variation between samples in the proportion of these reads. I categorised the sequences with respect to the canonical CAACAG sequence: loss of the CAA was coded "Loss", certain flanking sequences longer than CAACAG were coded "Gain",

and everything else was coded as "Other" (see Table 4.16). The data are plotted in Figure 4.19.

Cell line	Flanking sequence category	Reads	% of modal reads	% of reads modal +1 to modal +30	% of reads > modal +30	% of reads > modal
FAN1+/+	Canonical	42059	17.0	80.6	2.46	83.0
	Loss	838	12.2	81.7	6.09	87.8
	Gain	685	22.8	71.1	6.13	77.2
	Other	2926	17.5	77.9	4.61	82.5
FAN1-/-	Canonical	39049	14.6	83.2	2.20	85.4
	Loss	788	10.4	85.4	4.19	89.6
	Gain	637	18.4	77.9	3.77	81.6
	Other	2929	15.7	80.2	4.13	84.3

Table 4.17. Normalised read counts of categorised flanking sequences immediately downstream of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells by CAG length category using data from long-read PacBio sequencing. See Table 4.16 for flanking sequence category definitions.

The proportion of *FAN1*$^{+/+}$ reads greater than the modal CAG was significantly higher for reads coded "loss" compared to canonical (Chi-Square = 13.6, p = 2.3×10^{-4}), whereas the proportion of *FAN1*$^{+/+}$ reads greater than the modal CAG was significantly lower compared for reads coded "gain" compared to canonical (Chi-Square = 15.9, p = 6.5×10^{-5}). The proportion of "other" *FAN1*$^{+/+}$ reads greater than the modal CAG was not significantly different to canonical reads (Chi-Square = 0.57, p = 0.449).

The proportion of *FAN1*$^{-/-}$ reads greater than the modal CAG was significantly higher for reads coded "loss" compared to canonical (Chi-Square = 11.0, p = 9.1×10^{-4}), whereas the proportion of *FAN1*$^{-/-}$ reads greater than the modal CAG was significantly lower compared for reads coded "gain" compared to canonical (Chi-Square = 7.06, p = 7.9×10^{-3}). The proportion of "other" *FAN1*$^{-/-}$ reads greater than the modal CAG was not significantly different to canonical reads (Chi-Square = 2.60, p = 0.11).

The proportion of *FAN1*$^{+/+}$ reads (all codings) greater than the modal CAG is significantly lower than the proportion of *FAN1*$^{-/-}$ reads (all codings) greater than the modal CAG (Chi-Square = 93.3, p = 0), however the proportion of *FAN1*$^{+/+}$ reads (all codings) greater than the modal CAG +30 is significantly higher than the proportion of *FAN1*$^{-/-}$ reads (all codings) greater than the modal CAG +30 (Chi-Square = 9.6, p = 1.9×10^{-3}).

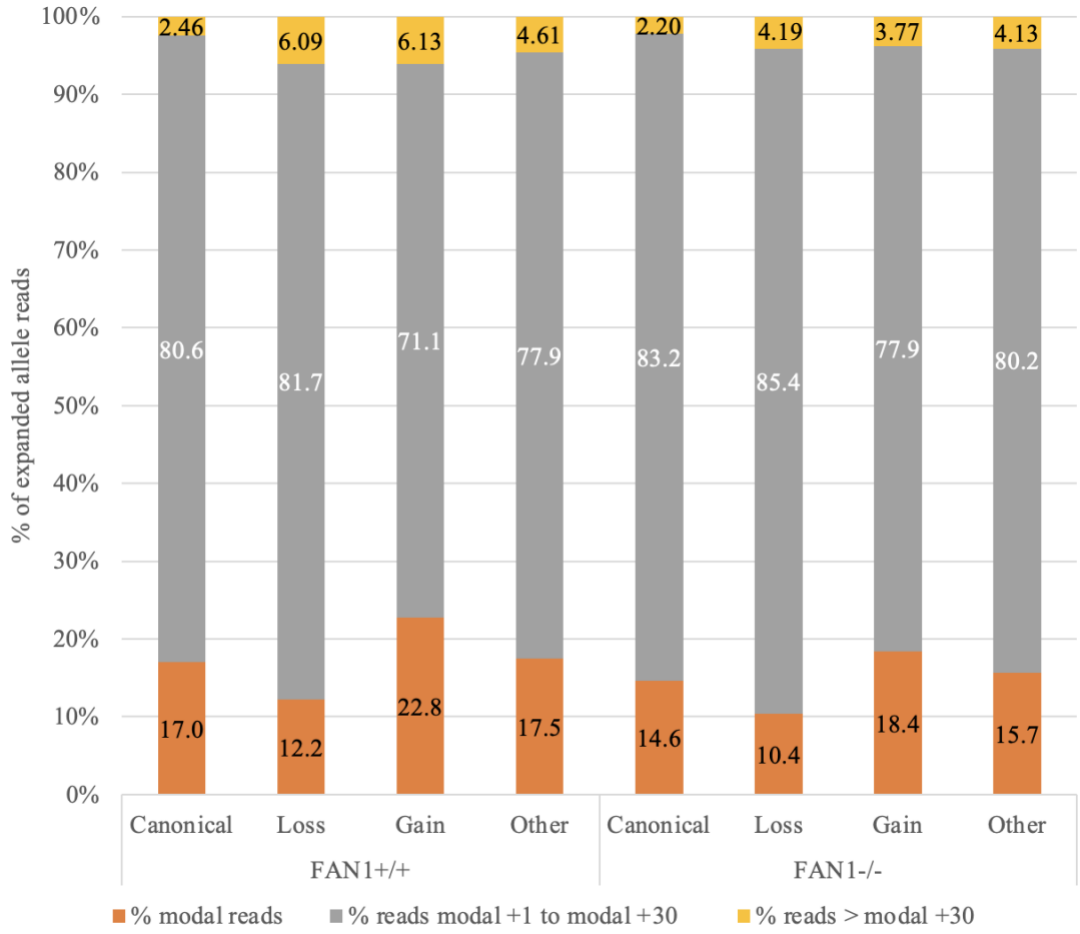


Figure 4.19. Normalised read counts of categorised flanking sequences immediately downstream of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells by CAG length category using data from long-read PacBio sequencing. RepeatDecoder restrictive profile CAG counts. See Table 4.16 for flanking sequence category definitions.

			Flanking sequence category (%)				
Increase in CAG	Bin width	Read count	Canonical	Non-canonical	Loss	Gain	Other
0-4	5	58198	91.1	8.90	1.47	1.52	5.90
5-9	5	20472	90.8	9.21	2.22	1.16	5.83
10-14	5	4968	86.2	13.8	2.54	1.53	9.72
15-19	5	2092	84.1	15.9	1.96	1.24	12.7
20-24	5	1083	83.0	17.0	3.42	1.11	12.5
25-29	5	697	83.5	16.5	3.59	2.15	10.8
30-34	5	479	81.8	18.2	3.13	3.76	11.3
35-39	5	333	83.2	16.8	2.10	2.10	12.6
40-44	5	239	84.1	15.9	2.51	2.51	10.9
45-49	5	207	87.4	12.6	3.86	2.42	6.28
50-54	5	147	76.9	23.1	3.40	4.08	15.6
55-59	5	145	83.4	16.6	2.76	4.14	9.66
60-64	5	93	87.1	12.9	3.23	3.23	6.45
65-69	5	100	86.0	14.0	3.00	1.00	10.0
70-74	5	100	81.0	19.0	2.00	2.00	15.0
75-79	5	73	75.3	24.7	5.48	4.11	15.1
80-84	5	72	80.6	19.4	2.78	4.17	12.5
85-89	5	59	83.1	16.9	0.00	1.69	15.3
90-94	5	51	86.3	13.7	3.92	0.00	9.80
95-99	5	45	77.8	22.2	8.89	2.22	11.1
100-104	5	45	82.2	17.8	11.1	2.22	4.44
105-111	7	34	67.6	32.4	11.8	11.8	8.82
112-116	5	31	90.3	9.7	6.45	0.00	3.23
117-124	8	31	87.1	12.9	9.68	0.00	3.23
125-132	8	30	83.3	16.7	3.33	6.67	6.67
133-148	16	30	80.0	20.0	0.00	0.00	20.0
149-191	43	30	66.7	33.3	3.33	3.33	26.7
192-566	375	27	74.1	25.9	14.8	0.00	11.1

Table 4.18 Normalised read counts of categorised flanking sequences immediately downstream of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells by increase in CAG from the modal CAG using data from long-read PacBio sequencing. RepeatDecoder restrictive profile counts. Bins are 5 CAGs wide up to 100-104, at which point bin width was increased to maintain a minumum read count of 30, except for the final bin, which has 27 reads.

To further examine the results presented in Figure 4.19, I investigated in more detail the reads longer than the modal CAG to establish whether alterations in read flanking sequences are statistically associated with increased CAG length. Table 4.18 shows the change in the proportion of flanking sequence read frequencies with the increase in *HTT* CAG from the modal CAG. The data were modelled by binomial regression using the GLM function in the Stats package in R (see 2.8.2). A 'success' outcome (x) for the four models were Non-canonical, Loss, Gain, and Other, respectively, while a 'fail' outcome (y) is Canonical in all models. Change in the log-odds of a success outcome for an increase of one CAG and p-values are shown in Figure 4.20. The modelling assumes all reads are independent observations.

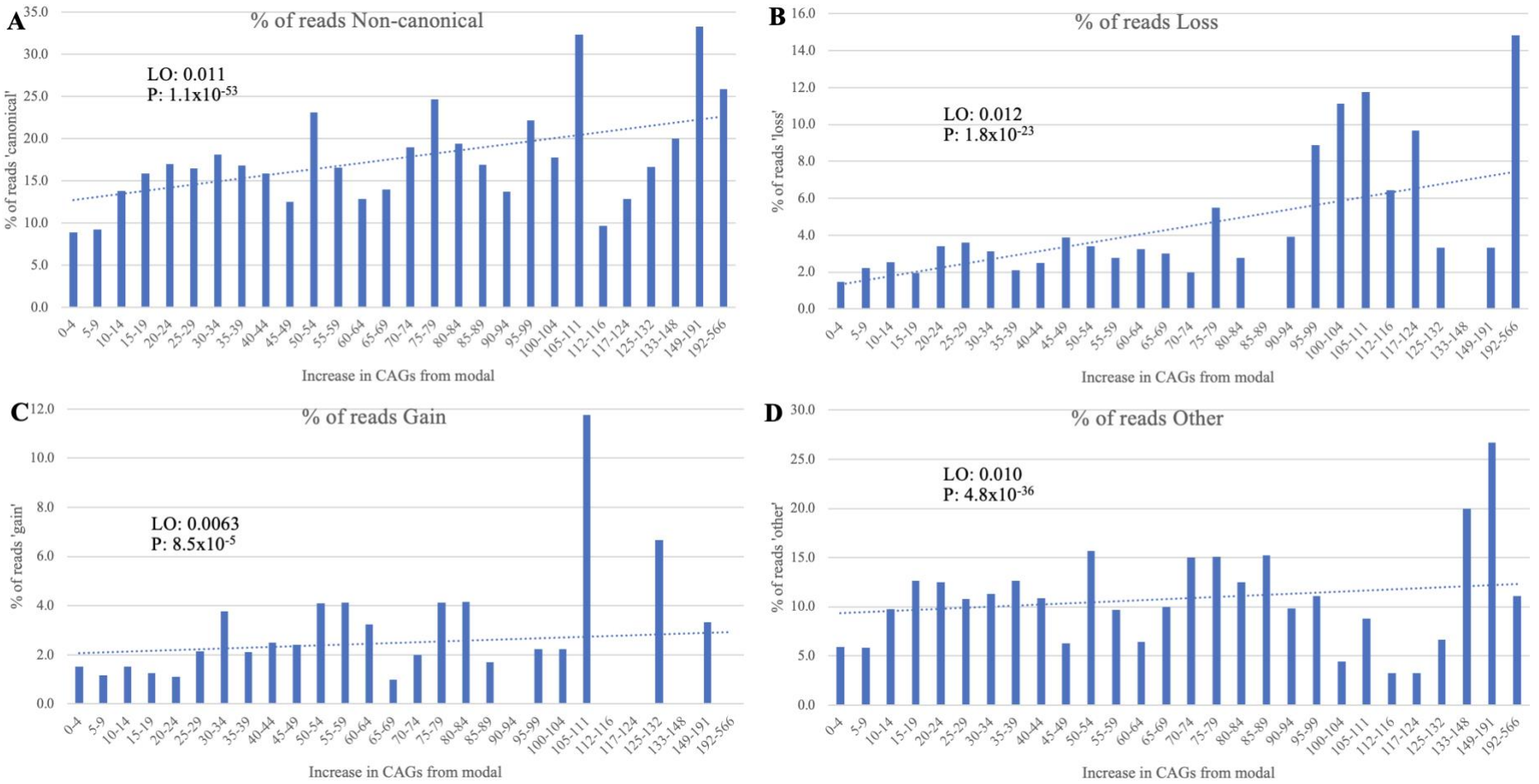


Figure 4.20. Normalised read counts of categorised flanking sequences immediately downstream of the expanded *HTT* CAG repeat in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells by increase in CAG from the modal CAG using data from long-read PacBio sequencing. (A) Non-canonical flanking sequences. (B) Loss flanking sequences. (C) Gain flanking sequences. (D) Other flanking sequences. RD restrictive profile counts. Bins are 5 CAGs wide up to 100-104, where bin width was increased to maintain read count ≥ 30, except for the final bin, which has 27 reads. Binomial regression stats: LO: change in the log odds of a 'success' outcome, P: p-value. Trend lines are lines of best fit. See table 4.23 for flanking sequence category definitions.

The probability of a read having a Non-canonical flanking sequence increases with increasing CAG length. For each increase in CAG the odds of a read possessing a Non-canonical flanking sequence increases by a factor of 1.011. Of the sub-categories of Non-canonical flanking sequence, the probability of a read having a Loss, Gain and Other flanking sequence all increase with increasing CAG length. For each increase in CAG the odds of a read possessing a Loss, Gain or Other flanking sequence increases by a factor of 1.012, 1.006 and 1.010 respectively. Log-odds coefficients were statistically significant in all four models. The p-value for Non-canonical is extremely low at $1.1x10^{-53}$. Loss and Other p-values are also extremely low at $1.8x10^{-23}$ and $4.8x10^{-36}$ respectively. The p-value of Gain is considerably higher than the rest but still significant at $8.5x10^{-5}$.

4.3.4 Validation of repeat length changes using small-pool PCR

Small pool PCR (SP-PCR) is a useful tool for assessing the degree of repeat expansion in a mosaic population of DNA molecules, like short tandem repeats in repeat expansion disorders, as it can detect rare large somatic expansions (Ciosi et al. 2021). SP-PCR involves diluting samples to just one or two molecules per PCR reaction and, in doing so, overcomes one of the main limitations of bulk-PCR approaches, i.e., that of the preferential amplification of shorter repeats. However, as a southern blot is required to visualise PCR products, the method is very labour intensive and has a low throughput. Also, as with bulk-PCR methods, contractions are still confounded by PCR slippage.

I was constrained in how much small pool sequencing could be conducted by the pandemic, which delayed both appropriate training and experimentation. However, I did conduct some SP-PCR on the samples analysed by long-read PacBio sequencing and fragment analysis in section 4.3.2, as in samples from the *FAN1* knockout experiment illustrated in Figure 4.4. Figure 4.21 shows a sample of the images taken of the two full southern blot membranes run for each cell line. Day 16 samples were used for both $FAN1^{+/+}$ and $^{-/-}$ lines. I had hoped to assay later time points as well but ran out of time. SP-PCR was conducted with 25 picograms of template DNA per reaction. Expansions (alleles > 900 bp) are visible in the $FAN1^{-/-}$ line but not the $FAN1^{+/+}$ line. One possible contraction (alleles < 800 bp) is visible in the $FAN1^{-/-}$ line. These results are reflected in the overall expansion counts. 2 people independently counted the number of expanded and contracted alleles in both membranes. The mean

of those counts is presented in Table 4.19. A mean number of expansions observed of 2.5 in the *FAN1*$^{+/+}$ line and 9 in the *FAN1*$^{-/-}$ line, representing 2.3% and 4.8% of the total number of alleles respectively. The difference in the number of expansions was not significant (Chi-square: 1.077, p = 0.299), however, this is based on a relatively small number of observations. The trend of the *FAN1*$^{-/-}$ line having more expansions overall is mirrored by PacBio sequencing data of these cell lines, where the proportion of reads with a CAG repeat longer than the modal CAG was significantly higher in the *FAN1*$^{-/-}$ line (see 4.3.3. and further discussion in 5.2). No contractions were counted in either cell line.

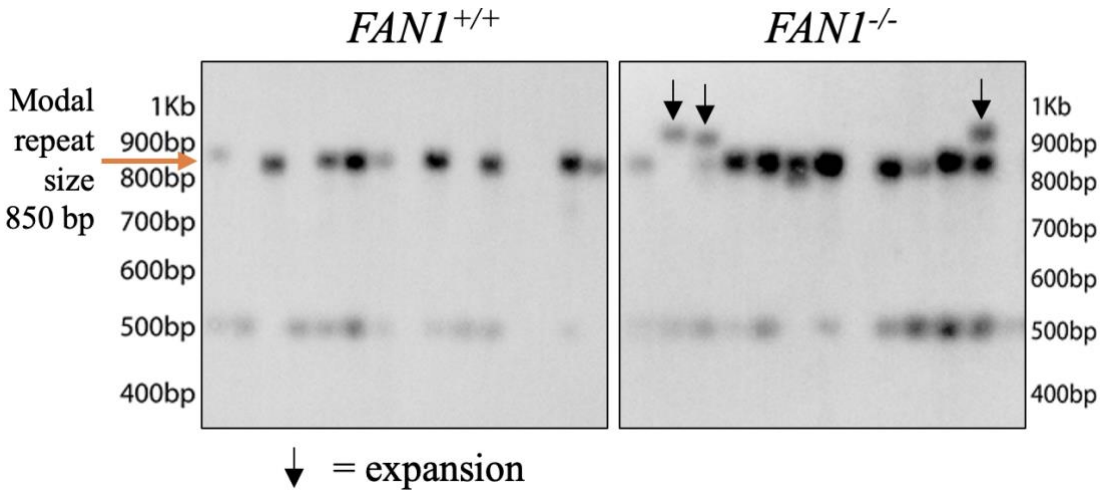


Figure 4.21. Illustration of small pool PCR southern blot of *HTT* CAG repeats in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells. Panel on the left shows 14 reactions of sample *FAN1*$^{+/+}$ day 16 cell replicate 1. Panel on the right shows 12 reactions of sample *FAN1*$^{-/-}$ t day 16 cell replicate 2. 25 picograms of template DNA was used in each reaction. An expansion is defined as any allele greater than 900 bp.

	Membrane totals	
	FAN1+/+	*FAN1*-/-
Number of alleles	108	186
Mean expansions	2.5	9
Mean contractions	0	0
% expansions	2.3	4.8
% contractions	0.0	0.0

Table 4.19. Summary of counts of expanded and contracted alleles in small pool PCR membranes. Membrane totals are sum totals from 2 membranes. An expansion is defined as any allele greater than 900 bp. A contraction is defined as any allele less than 800 bp. Mean expansions and contractions: counts from two people were averaged.

The modal peak was taken to be 850 bp for both cell lines, corresponding to a CAG length of approximately 145 CAGs. Expansions at 900 and 950 bp represent CAG allele lengths of around 163 and 180 CAGs respectively. A mean of 4 expansions of

100 bp were observed out of 186 alleles in the *FAN1*$^{-/-}$ line where 0 were observed out of 96 alleles in the *FAN1*$^{+/+}$ line. For the expansions >100 bp, a Chi-square test could not be made because there were no observations in the *FAN1*$^{+/+}$ line.

4.4 Discussion

4.4.1 Technical aspects of long-read sequencing

Long read sequencing is at present the only way to reliably determine both the length and sequence of repeats longer than 270 bp or 90 CAGs (Ciosi et al. 2021), such as those in some repeat expansion disorders, including the neuronal cell models presented here and many mouse models of HD (Kaye et al. 2021). To date it is not known whether these models undergo sequence changes in addition to CAG length expansions over time. Data presented in this chapter indicates that long-read PacBio sequencing can accurately determine the length and sequence of expanded repeats of 130 CAGs and above, which may enable insight into the dynamics of repeat expansion in HD and other REDs.

The first aim was to determine whether the long repeats present in the cellular model of the expanded *HTT* CAG repeat could be sequenced at sufficient depth to assess repeat expansion, stability, and sequence variation. The 600 bp library 600-iPSC-3 run on a single PacBio 1M SMRT cell generated approximately 98,000 reads, equating to ~8000 reads per sample with 12 samples. This read depth was high enough to produce reproducible measures of modal CAG length between PCR duplicates, detect differences in the change modal CAG length and expansion index over time, and observe differences in the rate of very large expansions between cell lines. Flanking sequence analysis of the same data revealed that the predominant flanking sequence in these lines, CAACAG, is the same as that determined by Sanger sequencing and that this does not change with cell age or *FAN1 genotype*. Within samples, approximately 15% reads have altered flanking sequences. The frequency of common alterations, including the loss the CAACAG and duplication of the CAACAG, changes little with PCR duplicate, cell age and *FAN1 genotype*. Over 90% of flanking sequence alterations are observed fewer than ten times.

Data from the 600-iPSC-3 library demonstrates that long-read PacBio sequencing can be used to perform novel experiments that can assess both repeat expansion stability

and sequence of the *HTT* CAG repeat. The potential and limitations of this technique are discussed further below.

PCR-electrophoresis based methods have been used extensively to estimate repeat lengths in both clinical and research settings and can be reliable and accurate. While they can provide a semi-quantitative assessment of the degree of somatic mosaicism and somatic expansion in cell populations they provide no information about sequence variation and are not sufficiently sensitive to detect alleles with a frequency of <10% (Massey et al. 2018).

As demonstrated in the previous chapter, long-read PacBio sequencing is well correlated with fragment analysis for measures of modal CAG length and, for longitudinal samples sequenced at high depth, measures of somatic expansion. In addition, long-read sequencing overcomes several limitations of fragment analysis in that it gives the entire repeat sequence and the flanking sequence. Highly accurate (> 99.9%) reads averaging 15 kb can be achieved according to recent data on PacBio's website (https://www.pacb.com/smrt-science/smrt-sequencing/ accessed 26th Jan 2022), allowing for very long flanking sequences to be captured. Rare alleles, including large repeat expansions can be detected due to the higher sensitivity and ability to count individual reads. This may prove important for future research if expansions in individual tissues and cells are demonstrated to be important in HD onset and progression (Swami et al. 2009; Donaldson et al. 2021).

While it is difficult to directly compare the sizing accuracy of fragment analysis and sequencing, sequencing pipelines permit stringent minimum quality thresholds to be set, which ensure only high-quality reads are present in analyses. Applying this and sequence-based filters to the reads allows for a high level of confidence that the reads observed in the data are of the repeat locus, rather than an artefact of PCR or library preparation. In addition, reads can be visually inspected and sequences graphically represented (Höijer et al. 2018), further adding to the level of confidence in the data. By contrast, guidelines for the clinical use of PCR-electrophoresis methods indicate that sizing is most accurate for shorter repeats, with error limits of repeats below 40 CAGs set at +/- 1 repeats, while for those above 39 repeats error limits were set at +/- 3 repeats (Losekoot et al. 2013). Long-read sequencing is likely to be more accurate at sizing longer repeats; however, this remains to be confirmed experimentally.

Measures of CAG size by fragment analysis cannot detect polymorphisms within the repeat and are prone to erroneous repeat counts in samples containing them. Sequencing the repeat allows determination of the uninterrupted CAG length and the flanking sequences, and thereby overcomes this limitation. This is of benefit to research into repeat expansion diseases as these are key sources of onset determining variation and may shed light on the mechanisms of repeat expansions (Ciosi et al. 2019; GeM-HD Consortium 2019; Wright et al. 2019). Sequencing data will be needed to improve the accuracy of diagnosis in patients with alleles of intermediate-length and improve age-at-onset prediction models which may in turn increase the power of clinical trials (Wright et al. 2020). Flanking sequences may also become important for future treatments as knowledge of flanking SNPs may allow for allele-specific *HTT*-lowering therapies.

Systematic differences in CAG sizing were observed between fragment analysis and PacBio data. The length of the uninterrupted CAG was 3.1 and 3.8% longer in PacBio data for the WT and expanded alleles respectively. This may be due to the sequence of the repeat as Losekoot et al. mention that the CAG repeat has altered mobility compared to heterogenous DNA in electrophoresis and that repeats of known size should ideally be used as standards (Losekoot et al. 2013).

Fragment analysis data of the neuronal cell *FAN1* knock-out experiment in section 4.3.2.3 showed a significant difference in the rate of change in day 16-anchored expansion index and Spread in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells. While similar trends were observed in the equivalent PacBio data, the differences were not significant. Significant differences in cell lines were seen in library 600-iPSC-3 (Figure 4.2A and B), so this may be a result of a small effect sizes in the data (the p-values were only just significant in fragment analysis) and more noise in the PacBio data, with modal CAG and day 16-expansion index having a higher expanded allele standard deviation (Table 4.13).

Increased variation in the PacBio data could be a result of the greater number of PCR cycles used in the library preparation. For fragment analysis there are 35 rounds of amplification and for PacBio undergoing 44 rounds. More PCR cycles results in more stutter and more chances to make replication mistakes. The same polymerase, LA Taq, was used in the preparation of fragment analysis and PacBio libraries. LA Taq is a proofreading polymerase which substantially improves the fidelity of amplification,

however error rates are still ~1 in 24,000 bases on heterogeneous DNA (NEB website: https://www.neb.com/tools-and-resources/feature-articles/polymerase-fidelity-what-is-it-and-what-does-it-mean-for-your-pcr [accessed 17 Jan 2022], Takara Bio website: https://www.takarabio.com/products/pcr/gc-rich-pcr/la-taq-dna-polymerase-with-gc-buffers [accessed 17 Jan 2022]). If a single DNA molecule is replicated 35 times this equates to 35 in 24,000 bases or 1 error in 686 bases. If a single DNA molecule is replicated 44 times this equates to 44 in 24,000 bases or 1 error in 545 bases. This means that there is a 26% higher chance of random polymerase errors from PCR cycles alone in PacBio compared with fragment analysis amplicons, which may explain some, but not all, of the differences in variation observed between the two methods. This higher error rate may amount for the large numbers of alleles that appear to fall between the expanded and WT alleles, as RD's restrictive profile is sensitive to non-CAGs sequences within the repeat.

While LA Taq is specifically engineered to perform well on high GC content DNA (website claims up to 73% GC content: https://www.takarabio.com/products/pcr/gc-rich-pcr/la-taq-dna-polymerase-with-gc-buffers?catalog=RR02AG accessed 26th Jan 2022), it is not designed to perform well on repetitive DNA. Amplification of STRs is known to be problematic, generating frameshift products known as 'stutter' possibly due to the complex secondary structures they are prone to form (Daunay et al. 2019). Amplification-free sequencing would overcome the issues with PCR mentioned here and give a clearer picture of sequence changes occurring in the cell.

Quality scores of the 600 bp amplicon sequencing data are high (Q40-42, which equates to ~99.99% accuracy) due to high number of passes per CCS read enabled by the short amplicon length. Despite this, changes in the sequence can be introduced prior to sequencing, during PCR or in the cell, and it is not possible to determine the source of these changes without further experiments. Indeed, the high sequencing quality scores suggest that most sequence alterations occur in the cell or during PCR. Using an amplification-free sequencing approach would help to establish the source of these alterations.

In the absence of PCR-free data, studies looking at error rates in PacBio data can help determine the reliability of commonly observed alterations in our data. The overall error rate of PacBio CCS sequencing data depends on the insert size and sample quality but data published by Weirather et al. showed an overall error rate of 1.72%

over 1.5kb, with insertions, deletions and substitutions comprising 0.087%, 0.34% and 1.30% respectively (Weirather et al. 2017). PacBio's own data suggests average read concordance is 99.8%, with 92.0% of discordances occurring in homopolymers (Wenger et al. 2019), however, it should be noted that this data is from unamplified DNA. Foox et al. sequenced multiple repeat types and showed that PacBio has an overall error rate of approximately 1% on simple repeats, with a similar rate for repeats with 25-70% GC content (CAG repeat is approximately 66% GC) (Foox et al. 2021). Assuming a substitution rate of 1%, and assuming the error rates for all possible substitutions are equal in PacBio data, an A to G substitution occurring at a specific base position should occur at a rate of 0.33% by chance alone. The fact that absence of the CAACAG is seen at a rate of 1.81% overall suggests they may be occurring in the cell, however error rates may be different in the context of a repeat.

Many of the single base changes that occur within the repeat sequence result in erroneous read length counts by RepeatDecoder. This is likely to be partly responsible for the high number of unique flanking sequences observed, and combined with PCR stutter, the high number of expanded reads with many fewer CAGs than the modal CAG (Figure 4.13).

While PCR was observed to have little effect on mean modal CAG and expansion indices in both fragment analysis and PacBio data (Table 4.3 and Table 4.8), the effect on normalised read counts in the PacBio data was much more striking. Table 4.5 shows that the mean percentage of expanded reads was 69.4% for PCR1 and 45.7% for PCR2. A similar but smaller effect is seen when considering the expanded reads only, with the mean percentage of reads greater than the modal CAG at 42.0% for PCR1 and 37.5% for PCR2 (Table 4.10). This shows firstly that modal CAG and expansion index are robust to differences in the proportions of reads above or below the modal peak. PCR itself is unlikely to be the source of this variation as there were no differences visible in the first-round products of PCR replicates. A more likely source of variation is separate handling of the replicates on different days, including multiple paramagnetic bead-based purifications in volumes less than 10 µl per sample.

The high degree of variation in the mean percentage of reads with a longer CAG than the modal CAG observed in PCR replicate, chip, i.e., sequencing run (Figure 4.15) and cell culture replicate (Figure 4.17) may explain the lack of clear trends emerging relating to *FAN1 genotype* and harvest day.

Some contamination was detected in the PacBio data, with a total of 537 reads surviving filtering in the water-only control. These are more likely to have come from cross-contamination, i.e., DNA molecules being transferred between samples, rather than human contamination, i.e., DNA molecules being transferred from laboratory users to samples, as 137 of these were classified as expanded alleles. The rate of their appearance is however relatively low, especially for the expanded allele. If we assume an equal rate of contaminating reads appears in the other 48 samples, around 17% of WT alleles may be contaminants and 3.5% of expanded alleles. A rate of 3.5% is unlikely to explain the lack of biological trends emerging relating to *FAN1 genotype* and harvest day.

4.4.2 Biological inferences from long-read sequencing

The second aim of this chapter was to examine the effect of *FAN1 genotype* and length of time in culture, i.e., developmental stage, in our neuronal cell model on the expanded *HTT* CAG repeat length, stability, and sequence. PacBio sequencing data from the 600-iPSC-3 library show that modal CAG and passage 4-anchored expansion index increase at a significantly higher rate in $FAN1^{+/+}$ cells compared to $FAN1^{-/-}$ cells (Figure 4.2). Data from 600-iPSC-3 are significantly positively correlated to equivalent fragment analysis data in both modal CAG and expansion index for the expanded allele, suggesting the effect may also be significant. In the following experiment, PacBio data from the 600-iPSC-4 library show that modal CAG and day 16-anchored expansion index do not increase at a significantly higher rate in $FAN1^{+/+}$ cells compared to $FAN1^{-/-}$ cells. Indeed, the modal CAG and anchored expansion index increase faster in $FAN1^{-/-}$ cells than the $FAN1^{+/+}$ cells. While this is not significant, there is a trend toward significance for anchored expansion index, which is in fact significant in the equivalent fragment analysis data.

Variation is seen between experiments for several reasons. In the initial experiment, samples with known high instability were chosen to provide the highest likelihood to detect large expansions when optimising PacBio sequencing and to establish whether there was a higher incidence of these in $FAN1^{-/-}$ samples. The effect of *FAN1 genotype* on the change in modal and expansion index over time was not initially considered for this experiment. The $FAN1^{+/+}$ sample chosen was very unstable and the $FAN1^{+/+}$ and $^{-/-}$ clones were not cultured in parallel. Furthermore, the initial experiment was with iPSC pellets that were taken at each passage, and the second experiment was

conducted in neurons- so there will also be variation and differences in expansion rates due to this: neuronal expansion rates are usually slower (McAllister et al. 2022). Also, 11B11 and 5F are both daughter clones of 109N1, but they were generated in different CRISPRs and from different stocks of 109N1. A more accurate comparison would have been to use unedited clones from the *FAN1* CRISPR. These were not available at the time but that is now routine practise in our lab.

The second experiment was better controlled as both clones were cultured in triplicate and in parallel, though only one clone per genotype was used. The *FAN1*$^{+/+}$ clone selected for this experiment has been seen to exhibit higher than usual expansion rates in both iPSCs and neurons in our lab. This may explain why modal CAG increases at a similar rate to the *FAN1*$^{-/-}$ line and may also explain why the difference in anchored expansion index was only just significant in fragment analysis data and not significant in PacBio data. The experiment should be repeated with additional 109N1 clones to reliably establish the effect of *FAN1*.

While the total proportion of expanded reads is higher in the FAN$^{-/-}$ line (41.4% compared to 38.1% *FAN1*$^{+/+}$), the proportion of reads greater than modal +30 is higher in the *FAN1*$^{+/+}$ line (1.27% compared to 1.19%). These relatively small differences are less than those observed in other variables in the experiment, for instance those due to the PCR replicate. Repeating the experiment with more replicates or on samples with a larger effect size would make it easier to detect biological differences. Alternatively, repeating the analysis with day-16 anchored values for modal CAG may show a larger effect size.

Table 4.7 shows that the percentage of reads in each cell line stays roughly the same before and after filtering, which indicates that there is no systematic bias between cell lines in the data.

There is a trend to the percentage of reads longer than the modal CAG increasing from day 16 to day 51, but then decreasing at day 71 in both cell lines. Again, more data would be needed to confirm this trend as the differences are relatively small compared to the differences in PCR replicate and sequencing run. Repeating the analysis with day-16 anchored values for modal CAG may show a larger effect size. It could be that longer reads are deleterious to cells and so more of them die, especially when older

and more stressed in culture. Quantifying the level of cell death using flow cytometry would have been informative in respect to this question.

While there was no immunocytochemistry conducted on these cells due to lack of available training time caused by pandemic restrictions, these cell lines have previously been shown to express markers for mature, post-mitotic MSNs using the same differentiation protocol (Donaldson 2019). Also, cell imaging shown in Figure 4.5 suggests these are neuronal cells, so I am confident that the intended cell type was present in our cultures. Changes to media colour in some wells at the later time points and the observation of dividing cells suggest other cells were present. This may explain why the variation in modal CAGs observed between replicates increases in both PacBio and fragment analysis data.

The third aim of this chapter was to examine whether reads with altered flanking sequences have altered repeat lengths compared to reads with non-altered flanking sequences in the expanded *HTT* CAG repeat of cell models. As mentioned previously in this section, many flanking sequence alterations are observed in the reads analysed in this chapter, however the majority of these occur at very low frequency and are likely to artefacts of PCR or sequencing. Others, however, consistently occur at relatively high frequency, such as the loss of the canonical CAACAG and its duplication. A frequency threshold of 0.1% was applied to alterations in our data, with reads above this threshold considered plausibly occurring in cells. Changes in the flanking sequences are significantly associated with changes in the distributions of repeat lengths and are closely reflected in both cell lines, as seen in Figure 4.19 and Table 4.17. Furthermore, the proportion of reads with a CAACAG repeat is significantly negatively associated with increasing CAG length. The proportion of reads with the loss of a CAACAG, a gain of sequence or ‘other’ sequence alteration is significantly positively associated with increasing CAG length. While the p-values of these significant effects are very low, the modelling assumes that each read represents an independent observation. While it could be argued that each read is an independent observation of a population of reads, all reads are derived from a limited number of related samples. More biological replicates, alternative statistical approaches and higher read depth experiments will be needed to confirm these tentative findings.

A study of the flanking sequences of expanded *HTT* CAG repeats in ~1000 patients found that the canonical glutamine-encoding flanking sequence CAACAG occurs in 95% of alleles, however it is not reported whether alterations were observed in the reads of these individuals or to what level (Lee et al. 2019). While the canonical sequence 3' to the CAG repeat was only observed at a rate of 86% in the reads of library 600-iPSC-4, the true number is likely to be higher than this due to the high rate of sequence changes introduced by PCR and sequencing (discussed above). The most common non-canonical allele observed in the study by Lee et al. was loss of the CAA (2.5%), followed by CAACAG duplication (1.75%), while other alleles were observed at much lower frequency: $(CAA)_3CAG$ (0.21%), $CAC(CAG)_3(CAACAG)_2$ (0.21%) $CAC(CAG)_3CAACAG$ (0.1%). Loss of the CAA is also the most common alteration observed the reads of my data at 1.8%, followed by CAACAG duplication at 0.60%. Of the other alleles seen in the Lee et al. study, $(CAA)_3CAG$ and $CAC(CAG)_3CAACAG$ were observed in my data but at a rate of 0.0048% and 0.0026% respectively. $CAC(CAG)_3(CAACAG)_2$ was not observed in my data. Other *HTT* sequencing studies have observed a CAACAACAG alteration at a rate of 0.18% (Ciosi et al. 2019) and 0.5% (McAllister et al. 2022), although the latter figure is from a sample containing patients at extreme residual age-at-onset so will have had a higher representation of non-canonical sequences. This sequence was seen at 0.50% in our data. These data show how the most common alterations to the glutamine encoding flanking sequence in HD patients are also the most common alterations in individual reads. This holds true for alleles observed at a frequency of 0.5% and above and lends credibility to the suggestion that alterations seen in reads are occurring in cells rather than just during PCR/library prep. Amplification-free sequencing at similar read depth to those in the experiments presented here could confirm this and could provide insight into the mechanisms by which these sequences are altered.

While I was not able to model these data due to time constraints, using linear models could be a powerful way to further examine the effects of *FAN1* genotype and cell age on read CAG length and flanking sequence in this data.

The fourth aim of this chapter was to determine whether the large CAG repeat expansions observed in long-read PacBio sequencing of iPSC models are also observed in SP-PCR. Table 4.20 shows a comparison of PacBio and SP-PCR data. 50 bp is the equivalent of 16.6 CAGs, while 100 bp is the equivalent of 33.3 CAGs. The

percentage of shorter expansions is higher than longer expansions for both lines in both methods. While the percentage of shorter expansions is higher in the *FAN1*$^{-/-}$ line in the PacBio data, the difference in the shorter expansions between lines is greater in the SP-PCR data, although the difference is not significant. The percentage of longer expansions is similar in *FAN1*$^{+/+}$ and $^{-/-}$ cells in the PacBio data. While the number of longer expansions in the *FAN1*$^{-/-}$ cells was higher in SP-PCR a test of difference was not possible.

	PacBio		SP-PCR	
Cell line	**% 16.6+ CAG expansions**	**% 33.3+ CAG expansions**	**% 50+ bp expansions**	**% 100+ bp expansions**
FAN1+/+	2.6	1.1	2.3	0
FAN1-/-	3.0	1.0	4.8	2.2

Table 4.20. Mean percentage of expansions of equivalent size in PacBio and SP-PCR data in *FAN1*$^{+/+}$ and $^{-/-}$ neuronal cells. For both methods, 'expansions' is defined as the number of reads/alleles greater than or equal to the modal CAG plus a specified threshold.

SP-PCR seems to show the difference between the *FAN1*$^{+/+}$ and $^{-/-}$ cells more clearly than PacBio – this is likely due to the dilution of alleles. However, differences are not statistically significant, and more data SP-PCR data is needed to confirm this. The number of alleles observed in the small pool data was 108 and 186 for the $^{+/+}$ and $^{-/-}$ lines respectively. This is several orders of magnitude lower than the reads in the PacBio data and may yet explain the variability observed between cell lines. Despite this, the large CAG expansions seen in PacBio data are seen in SP-PCR.

Chapter 5 : General discussion

5.1 Summary of findings

The overarching aim of this thesis was to evaluate the accuracy of measures of the *HTT* CAG repeat derived from long-read PacBio sequencing compared to existing methods in samples derived from HD patients and iPSCs. I started by sequencing a 3000 bp region including the *HTT* CAG repeat in LBCs and PBMCs derived from HD patients as I had access to other sequencing data against which to validate PacBio. Repeat counts correlated well with those of ultra-high depth repeat-spanning MiSeq reads using both ScaleHD, a repeat counting method designed for MiSeq-length reads, and RD, a novel repeat counting algorithm. The maximum difference observed between repeat counts on the two sequencing platforms, regardless of counting algorithm, was 1 CAG.

A comparison of PBMC and LBC sequencing data showed that while modal CAG lengths of the WT and expanded alleles were significantly and strongly correlated for both MiSeq and PacBio data, expansion indices were only significantly positively correlated in MiSeq data of the WT allele. This is likely to be explained a lack of expansion in the data, meaning it is essentially the degree of noise in each sample that's being correlated. The WT allele shows little expansion in HD (Swami et al. 2009) and *HTT* CAG repeats show little expansion in unaffected tissues except the liver (Kennedy et al. 2003).

I then sequenced DNA extracted from several iPSC cell pellets with ~130 CAGs and found that counts of modal CAG were correlated with fragment analysis data of identical samples but that expansion indices were not. However, because there was a low representation of the expanded allele amplicons amongst all amplicons, I decided to sequence a shorter amplicon length at 600 bp and introduce paramagnetic bead-based size-selection steps to physically enrich the expanded allele. This increased the read depth of the expanded allele approximately 7-fold and increased the median read quality by more than an order of magnitude but failed to generate expansion indices that were significantly positively correlated with fragment analysis data, which may be because the alleles showed relatively little expansion, being from cross-sectional samples. PacBio-RD CAG counts were consistently ~4-5% longer than their fragment

analysis equivalents which could be due to the anomalous migration of CAG repeat-containing DNA in electrophoresis (Losekoot et al. 2013). Losekoot et al. recommend running with a ladder of known repeat sizes to mitigate this issue. Despite this, PacBio-RD expansion indices were significantly positively correlated with those of fragment analysis in data from longitudinal samples, probably due to the higher levels of somatic expansion observed which is the result of using longitudinal samples.

Approximately 2500 expanded allele reads (1.3%) of library 600-iPSC-4 had uninterrupted CAG lengths of 160 CAGs or longer, while 679 (0.4%) had a CAG length longer than 200 CAGs. Due to the high penalties for mismatches and indels, RD's restrictive profile represents a stringent counting algorithm, which, in combination with the filters applied in my analysis pipeline (Figure 2.1), makes it extremely unlikely for library preparation artefacts or non-*HTT* CAG sequences to survive analysis and is further evidence that PacBio reads can reliably span CAG repeats in excess of 200 CAGs (Ciosi et al. 2021). The longest repeat observed in my data was 696 CAGs, which although only one of two reads longer than 500 repeats, is clearly a continuous repeat upon manual inspection of the FASTA sequence. Hafford-Tear et al. sequence repeats up to 1500 CAGs, suggesting this is well within the maximum capability of PacBio sequencing platform (Hafford-Tear et al. 2019). My data shows that PacBio sequencing can be used to sequence the expanded alleles of repeat diseases with very large (> 500) repeats, which is important in diseases such as Fragile X syndrome, where stability of the repeat is sequence-dependent (Pollard et al. 2004).

The other main objective of this thesis was to conduct experiments that examine the effect of *FAN1* genotype and cell maturity on repeat length, instability, and sequence variation in 109-HD iPSC samples. PacBio data from iPSC cell pellets showed that the rate of increase in modal CAG and passage-4 anchored expansion index over time was significantly higher in $FAN1^{+/+}$ cells compared to $FAN1^{-/-}$ cells. No difference was observed in PacBio data of neuronal cell lines for modal CAG or day-16 anchored expansion index, however day-16 anchored expansion index increased significantly faster over time in the $FAN1^{-/-}$ in fragment analysis data and there was a trend towards a faster rate in the $FAN1^{-/-}$ cells in the PacBio data. CAG read length distributions, including the rate of expansions and rare large expansions, showed greater variability

for experimental replicates, especially library preparation and sequencing run, than for genotype, or harvest day (discussed further in section 5.2).

Flanking sequence variation was observed within samples, with the most common sequence alterations observed, i.e. loss of the CAA immediately downstream of the pure CAG repeat and its duplication, mirrored by the most common flanking sequence alterations observed in HD patients (Ciosi et al. 2019; Lee et al. 2019; Wright et al. 2019). CAA interruptions are also observed in other CAG repeat diseases, including SCA2 (Charles et al. 2007) and SCA17 (Gao et al. 2008), where presence of the CAA is associated with reduced repeat expansion, as in HD. CAA interruptions have been shown to stabilise the CAG repeat in biophysical studies (Rolfsmeier and Lahue 2000; Dorsman et al. 2002) and increase the fidelity of DNA polymerases (Dorsman et al. 2002). CAG repeats have been shown to form hairpin structures and their stability depends on the flanking sequence (Gacy et al. 1995). Another *in vitro* study shows that CAT or AGG interruptions of the CAG repeat reduce the propensity of slipped-strand DNA formation (Pearson et al. 1998), while data from yeast show that interruptions that are centrally located within a repeat tract are less prone to expansion, which is consistent with the hypothesis that interruptions inhibit expansions by reducing hairpin stability (Rolfsmeier and Lahue 2000). From an evolutionary perspective, the stabilising effect of CAA interruptions could explain why it is present in nearly all normal *HTT* alleles and most HD alleles, acting as a repeat expansion inhibitor, which could confer a selective advantage. Variants other than those involving the loss or gain of CAAs are observed in HD expanded repeats but at a much lower frequency, i.e. 0.31% vs 4.42% (Lee et al. 2019). The fact that the same mutations are observed in expanded *HTT* repeats within and between samples suggests common mechanisms operating at the DNA level, for instance, error biases of polymerases or DNA repair machinery. Other repeat expansion diseases have different repeat stabilising interruptions, expansion biases and tissues-dependence (Khristich and Mirkin 2020), suggesting that the cellular context plays an important role in determining replication/repair error biases.

The most common flanking sequence changes were observed at similar rates in both *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ cells, suggesting that the mechanism by which *FAN1* stabilises the repeat is independent of the mechanism that governs flanking sequence alterations.

As discussed above, presence of a CAA within the CAG repeat is associated with the inhibition of repeat expansions in biophysical studies compared to its absence. Furthermore, loss of the CAA is associated with earlier AMO in HD, while gain of a CAACAG is associated with later AMO (Lee et al. 2019). If somatic expansion is a key driver of HD pathology, as suggested by previous findings (Swami et al. 2009), then loss of a CAA interruption should be associated with an increase in the rate of expanded repeats in cells and vice versa for CAACAG duplication. The ideal experiment would use matched cell lines where the flanking sequence mutations had been precisely introduced into the parent lines, however, flanking sequence variation in my data allows for natural experiments looking at the association of the rate of flanking sequence alterations with CAG expansion length alterations in reads. Loss of the CAA was associated with a significant increase in the proportion of expanded (CAG repeat greater than modal CAG) reads compared to the canonical CAACAG in both cell lines. Gain of flanking sequence was associated with a significant decrease in the proportion of expanded reads compared to the canonical sequence in both cell lines. The proportion of reads greater than the modal CAG for the 'Other' flanking sequence group was not significantly different from the canonical sequence. These results suggest that alterations in flanking sequences in individual neurons are associated with alterations in CAG length and the direction of effect for the Loss and Gain groups are consistent with previous literature on the subject.

Furthermore, the proportion of expanded reads with the canonical flanking sequence category was inversely associated with expansion size, while the proportion of expanded reads with a 'loss', 'gain' or 'other' flanking sequence category was positively associated with expansion size. While the direction of effect is that which you would expect based on previous literature and the above results for the loss of CAA, gain of sequence confounds expectations, although the effect size is the smallest of the four, with an increase in the proportion of reads equivalent to ~1% over 567 CAGs, compared to ~7% for loss of CAA and ~6% for 'other' sequences. The inverse association observed in the canonical reads reflects the positive associations in the other groups. Some of the associations observed, particularly in the "other" group may be driven by the inverse association between read length and predicted read accuracy in PacBio sequencing, as seen between the 3000 bp and 600 bp libraries in my data.

Also, limited number of reads at higher read lengths may be a factor, as, I suspect, is the assumption in the modelling that each read is an independent observation.

5.2 Reproducibility of repeat counts

The data in Chapter 4 show that while other factors affected the proportion of reads of different lengths that were observed, the library preparation (PCR) and sequencing run (chip) had the most marked effect (Figure 4.15 and Table 4.10). This may in part be down to the relatively high expansion in the FAN$^{+/+}$ cells discussed in chapter 4. It could also in part be due to the selective loss of cells with longer repeats at later time points. However, the high degree of variation between library preparations and sequencing runs is more certainly a contributing factor. Regarding, library preparation, using different master mixes/template DNA aliquots between PCR 1 and 2, would have introduced some variation – this could be removed by using one master mix/template pot for both reactions. Bead purifications were conducted at low volumes (~10ul). The exact bead:sample ratio is important in determining the size-selection threshold (https://www.broadinstitute.org/genome-sequencing/broadillumina-genome-analyzer-boot-camp accessed 13/02/2022). Diluting the DNA prior to purification would reduce the amount of pipetting error at this stage and potentially reduce the amount of variation in CAG lengths observed between library preparations. Sequencing libraries were loaded at different concentrations on Chips 1 and 2 as we had not optimised loading of our samples on Exeter's machine, meaning the first chip was underloaded. PacBio is reported to over-represent shorter reads (Ciosi et al. 2021), which could plausibly be influenced by the library concentration. Repeating the experiment with the optimal loading conditions used for the second run may reduce sequencing run variation.

Removing reads with fewer repeats than the modal CAG from the analysis may also reduce noise in the experiment as contractions that occurred in the cell can't be distinguished from those induced by library preparation/CAG counting artefacts – PCR stutter is biased towards deletions (contractions) (Veitch et al. 2007), and the stringency of RD's restrictive profile means it is also biased towards contractions. Indeed, small-pool PCR data indicates that contractions may be very rare in these cell lines, with no contractions counted among the ~300 expanded alleles observed (compared to the PacBio data where contractions consistently comprised more than 50% of expanded allele reads).

PacBio and small-pool PCR data are much more consistent with respect to expansions. Large CAG repeat expansions (>30 CAGs) seen in PacBio data were also observed in SP-PCR membranes. The proportion of expanded reads was significantly higher in PacBio data of the *FAN1*$^{-/-}$ cell line (Chi-Square = 93.3, p = 0), although Chi-Square assumptions may be violated (as reads may not qualify as independent observations) and this needs to be modelled with all experimental variables. SP-PCR showed the same effect of the *FAN1* genotype on the rate of expansions. While this effect was not statistically significant, it was based on only 300 alleles from a single time point. Given more time, it would have been useful to run multiple membranes at all time points for better power to detect differences in genotype, for a more comprehensive comparison and to see, using another method, if cell maturity/harvest day influenced the rate of expansions.

Amplification-free analyses would overcome the preferential amplification of shorter repeats, contraction biases, and other errors introduced during library preparation as the cellular DNA is assayed directly. Use of these techniques are discussed in more depth in section 5.8. More replicates (library preparation, cell culture and sequencing runs) would increase the power to detect biological differences, as would eliminating the low-level contamination between samples observed.

5.3 The advantages of long-read sequencing

To date, long-read DNA sequencing is the only technique that gives reliable information about the size and sequence of repeats of 90 CAGs or longer (Ciosi et al. 2021). This is becoming increasingly important for the investigation of novel animal and cell models as repeat expansion phenotypes are rarely observed in models with repeats below this size ((The HD iPSC Consortium et al. 2012). The findings presented here suggest that repeat expansion may be associated with alterations to the flanking sequences of individual DNA molecules within pathogenically relevant somatic cells. Amplification-free long-read sequencing has the potential to eliminate the ambiguity of the origin of sequence alterations and can provide the methylation status of DNA (Giesselmann et al. 2019). Evidence suggests methylation status modifies the repeat expansion diseases frontotemporal dementia and amyotrophic lateral sclerosis (Xi et al. 2013; Russ et al. 2015). Furthermore, broader patterns of alterations revealed by whole-genome sequencing, including complex chromosomal rearrangements in cancer, are easier to detect with long-read NGS methods as longer reads make

mapping across repetitive DNA possible (Nattestad et al. 2018; Nesic et al. 2018). Understanding mutational signatures may also shed light on repeat disorders, as is the case in cancer (Gold 2017; Díaz-Gay and Alexandrov 2021), where codon mutation biases in different cancer types point to different underlying mechanisms.

5.4 Clinical implications

Current methods routinely used to size repeats in clinical diagnostics fail to capture the polymorphic sequence variation immediately downstream of the pure *HTT* CAG repeat. This leads to erroneous repeat sizes in approximately 1% of patients, in some cases by as many as 7 repeats (Wright et al. 2019). Accurate sizing of the uninterrupted repeat is particularly important in individuals with repeat lengths at the borders of intermediate, reduced penetrance (36-39 CAGs) and full penetrance alleles as this has implications for family members (Wright et al. 2020). Sequencing platforms which generate reads spanning typical expanded allele repeat lengths (36-55 CAGs) repeat are best suited to do this as they can provide accurate CAG sizing and detect non-canonical sequences. Ciosi et al. demonstrated this using Illumina's MiSeq, which can be deployed at high throughput at relatively low cost (Ciosi et al. 2021). In rare cases of *HTT* repeats longer than 90 CAGs, as seen in some juvenile HD cases, MiSeq sequencing will establish the presence of a very long repeat, but not its size. It would be necessary to follow this up with one of the long-read NGS methods to accurately determine the full length and sequence in such alleles.

It is increasingly recognised that the repeat sequences immediately downstream of the CAG repeat have a modifying effect on HD onset. This has been considered spurious due to miscounting of the pure CAG repeat due to assuming a canonical repeat (GeM-HD Consortium 2019) but recent evidence suggests that the actual sequences have a small but significant effect on phenotype (McAllister et al. 2022). Although CAG length and flanking sequence measures have significant effects on HD at the population level, they are insufficiently accurate at the level of an individual patient to be able to use them in clinical practice, for example in prognostication. However, adding them as covariates in analysis of research studies and clinical trials will add power to these analyses and, as we understand more about HD pathogenesis and its modification, they might eventually have a clinical use.

5.5 Implications for other repeat expansion diseases

HD is an intensely studied repeat expansion disease, with 6,894 publications on the subject in the last 40 years (https://pubmed.ncbi.nlm.nih.gov search term: (huntington's disease[Title]) AND (("1982/02/13"[Date - Publication] : "3000"[Date - Publication])), accessed 13/02/22), and serves as a model disease for other repeat expansion diseases (REDs). REDs are a family of related disorders, which often have neurodegeneration as part of their disease spectrum but different parts of the CNS/PNS degenerate and there is a broad spectrum of phenotypes. More than 40 have been discovered to date, with that number growing year-on-year (Chintalaphani et al. 2021). Some REDs have pathogenic repeat sizes like or shorter than HD, but only manifest with much longer repeats such as myotonic dystrophy 1 (50-10,000 CTGs) and *C9ORF72* ALS/FTD (24-4000 GGGCCs). Long-read sequencing is an important tool in accurately sizing and sequencing repetitive DNA, especially in the case of REDs with very long repeats, as, at the time of writing, no other technique will give this information.

Breakthroughs in our understanding in HD can lead to breakthroughs in understanding of other repeat diseases and vice versa. For example, GWAS studies have identified SNPs around *FAN1* being associated with altered AMO in multiple CAG-repeat diseases, including HD (Bettencourt et al. 2016; GeM-HD Consortium 2019). This prompted Zhao and Usdin to test the effect of knocking out *FAN1* on repeat expansion in a mouse model of Fragile X syndrome. They found that *FAN1* had a protective effect against repeat expansions in the brain and some other somatic cells but not germline cells (Zhao and Usdin 2018). Evidence that *FAN1* is important in preventing repeat expansion in one RED suggests it may prevent repeat expansion in others, especially ones where *FAN1* variants are associated with AMO. Similar experiments have since been conducted in models of HD, which have also shown an expansion preventing effect of *FAN1* (Goold et al. 2019; McAllister et al. 2022).

5.6 Limitations of current work

Variation between subclones derived from the same clonal line was observed between the two experiments presented in chapter 4 and is as seen previously (Donaldson 2019). This highlights the stochastic nature of repeat expansion in these cell models and presents a challenge to their analysis and utility. Functional variability between

pluripotent stem cell clones is thought to arise from genetic and epigenetic variation (Cahan and Daley 2013), however exome sequencing of a number of 109NI subclones revealed few single nucleotide variants between them (McAllister et al. 2022). The emergence of the duplication of chromosome 1 in both parent and daughter 109NI lines underlines the need for stringent controls including karyotyping at multiple passages, as was conducted on the iPSCs analysed in this project. Differences between the *FAN1*$^{+/+}$ and $^{-/-}$ lines were observed in iPSC and neuronal cells, albeit in the opposite direction: *FAN1*$^{+/+}$ cells showed a significantly higher rate of increase in modal CAG and expansion index compared to *FAN1*$^{-/-}$ cells in iPSCs, whereas *FAN1*$^{-/-}$ neurons showed a significantly higher rate of expansion index increase (fragment analysis only but trending towards significance in PacBio data) compared to *FAN1*$^{+/+}$ neurons (discussed further in section 4.4). However, Donaldson argues that more than 3 subclones of each line should be used when trying to establish differences in repeat expansion between genotypes (Donaldson 2019).

In addition to the variability between iPSC clones, there is also heterogeneity between neuronal cultures from the same clone, as demonstrated by modal CAG, expansion index values across the 3 culture replicates shown in Figure 4.7 and in the proportion of reads at different CAG lengths shown in Figure 4.17. Neuronal cultures tend to be a mixed population of different neuronal subtypes and the challenges of obtaining pure populations of MSNs are reported elsewhere (Le Cann et al. 2021). Observations of changes to media colour in some of our cultures and dividing cells suggests that they contained a mixture of differentiated, partially differentiated, and dividing cells, including some neurons and other neuronal lineage cells. Ideally, we would have used markers of MSNs to estimate their prevalence (The HD iPSC Consortium et al. 2012) or purified the cells using FACs prior to the experiment (Basu et al. 2010). The presence of non-MSN cells may be diluting the repeat expansion phenotype in this experiment - repeat expansion is marked in HD patients' striatum (Kennedy et al. 2003) and MSNs comprise 95% of total striatal cells in human. Furthermore, MSNs are known to be particularly susceptible in HD (Vonsattel et al. 2008). The selective loss of the longest repeat expansions may be further diluting the repeat expansion phenotype.

iPSCs are easier to modify genetically and higher throughput than animals and expansion over shorter-time scales makes them attractive research models. However,

so far, no iPSC models with typical pathogenic expanded allele repeat lengths (36-55) have demonstrated a reliable repeat expansion phenotype in culture (The HD iPSC Consortium et al. 2012; Xu et al. 2017), which raises the question of how comparable these cells are to adult-onset HD. Longer repeats like those of the HD-109 lines are only seen in juvenile HD and in some cells of adult HD brain, although most are still 40-50 CAGs (Swami et al. 2009). iPSCs are typically grown over a very short period so how relevant they are to adult neurodegeneration is debatable, however they can be used model repeat expansions, which is a defined phenotype. Repeat expansion is observed over 4 weeks with this model and allows for controlled experiments to be conducted on a shorter timescale compared to animal models.

Bulk-PCR methods including the amplicon sequencing approach used in this project have multiple limitations, including the preferential amplification of shorter repeats - resulting in the under-representation of expansions - and PCR slippage. If somatic expansion drives neuronal degeneration, cells with large expansions may be removed from cell culture populations, making large expansions even harder to detect by bulk-PCR methods. In Small-pool PCR (SP-PCR), the preferential amplification of shorter repeats is limited as the template DNA is diluted to one or two template molecules per reaction. SP-PCR has been shown to capture repeat length gains of up to 1000 CAGs in human HD striatal tissue (Kennedy et al. 2003). A comparison of SP-PCR to bulk PCR method on the *HTT* CAG repeat showed that large expansions are better detected by SP-PCR (Ciosi et al. 2021). However, SP-PCR is labour intensive and relatively low throughput compared to bulk-PCR approaches, meaning I was only able to generate a small amount of data in the limited time I had. Given more time I would like to have conducted more SP-PCR assays, especially on samples from later time points, to better capture changes in the rate of large repeat expansions over time in these cell lines and differences between genotypes, if present.

5.7 Future work

Another active area of development is in the use of amplification-free sequencing approaches. Methods of physically enriching target loci include the use of CRISPR/Cas9 using guide RNAs complementary to the target, which has been applied to both PacBio (Tsai et al. 2017) and ONT sequencing (Giesselmann et al. 2019), and the use of real-time recognition and rejection of off-target DNA, which is possible with ONT's Read Until API (Payne et al. 2021). Directly sequencing unamplified

DNA allows most of the limitations of PCR to be overcome and in theory provides a much more accurate reflection of the population of cells assayed. At the time of writing amplification-free techniques have been successfully applied to the *HTT* CAG repeat but are limited to several hundreds of reads per run and require several micrograms of sample DNA (Tsai et al. 2017; Höijer et al. 2018). Until amplification-free sequencing can routinely yield several thousand reads per sample, small pool PCR sequencing may provide a viable alternative for capturing the size and sequence of rare repeat expansions (Ciosi et al. 2021).

In the data presented in Chapter 4, flanking sequences immediately downstream of the expanded *HTT* CAG repeat tract in neuronal cells showed a high degree of variability within samples. Approximately 15% of reads in all samples were found to have a non-CAACAG flanking sequence. A relatively simple extension to this work would involve establishing this rate in WT alleles of the same samples to act as a baseline and, further, to check the rate of flanking sequence alterations in the patient alleles of earlier libraries. Given more time I would like to have used more sophisticated statistical approaches to interrogate flanking sequence-CAG length associations, for example modelling the CAG length and flanking sequencing from all reads over time and genotype with the other experimental variables as covariates in mixed linear regression.

Single-cell sequencing methods would allow the profiling of specific cell types and may provide insight into questions relating to the pathogenic repeat size threshold. One study edited the *HTT* CAG lengths of human embryonic stem cells, to generate a range of pathogenic repeat sizes and differentiated them to different cell types, including neurons, and identified multiple CAG length-dependent and cell type specific transcriptional and proteomic phenotypes (Ooi et al. 2019). Single-cell sequencing would enable cell-specific RNA expression changes to be observed in the context of molecular phenotypes and greatly improve the specificity of assays of the type conducted by Ooi et al.

Singe-cell sequencing of both RNA and DNA simultaneously would give the exact CAG repeat length and show whether this is reflected by the RNA sequence in the same cell. Some studies suggest the RNA itself may be a pathogenic agent in REDs (Hsu et al. 2011; Lawlor et al. 2011). A single-cell sequencing approach would also reveal RNA expression levels and allow the investigation of the effects of *HTT* CAG

length on the transcriptome and vice versa. Furthermore, it could be used to reveal the effects of different repeat lengths and known genetic modifiers on individual cells in model systems. While this approach remains extremely technically challenging at the time of writing, it has been demonstrated (Macaulay et al. 2015; Macaulay et al. 2016) and is a highly active area of research.

There is mounting evidence that *FAN1* is protective against repeat expansion in cells. Goold et al. show that increased *FAN1* expression is significantly associated with delayed AMO and that *FAN1* overexpression in HD-109 cells reduces CAG repeat expansion (Goold et al. 2019). McAllister et al. show that nuclease-dead D960A *FAN1* mutation is associated with significantly faster expansion rates, which suggests that expansion is nuclease-dependent (McAllister et al. 2022). Further to this, Goold et al. show that *FAN1* can also inhibit CAG repeat expansion through its interaction with MLH1 (Goold et al. 2021). It does this by competing with MSH3 for MLH1 binding and, in doing so, restricts the formation of a functional mismatch repair assembly found to promote repeat expansion. Further to this, Porro et al. show in *in vitro* assays that the FAN1-MLH1 interaction promotes the repair of slipped-DNA structures formed by the CAG repeat, thought to be one of the mechanisms by which CAG repeats expand (Porro et al. 2021). These findings present new avenues for research and suggest that a better understanding of the DNA-protein and protein-protein interactions involved with repeat expansion and ways to modulate them will be critical for developing therapies.

5.8 Concluding remarks

Evidence pointing to somatic expansion being a key driver of HD pathology has emerged from increasingly numerous sources over the last few decades. Understanding the key determining factors in repeat expansion, including the role of DNA repair proteins, repeat flanking sequences and other onset modifying variants is likely to be critical to understanding how the disease progresses and is likely to generate novel therapeutic targets. Cell and animal models allow genotypes and phenotypes associated with expanded CAG repeats to be investigated in a controlled and timely manner. Being able to reliably sequence through long repeats will be essential to understand these associations and may be critical to the development of allele-specific gene-targeting therapies in a range REDs.

Appendices

Appendix 1: Flanking sequence windows and per-base Phred quality scores of 30 randomly selected reads with a "CAACAG" flanking sequence in PacBio sequence data of *FAN1*$^{+/+}$ and *FAN1*$^{-/-}$ 109NI neurons.

CAGs counted by RepeatDecoder restrictive profile. Flanking sequence, coloured blue, is the region between the 3' ends of the restrictive and permissive profiles. Per-base quality scores: Phred quality scores ranging from 0 to 93 (see explanation in section 1.6.2). Condition takes the format: cell line, d: harvest day, r: biological replicate, p: PCR replicate. (A) "", i.e., loss of the CAA. (B) "C". (C) "CAACAGCAACAG". (D) "CAACAACAG". (E) "CAAGCAG".

A

Read ID	CAGs	Condition	Sequence	Per-base quality scores
4194442	127	FAN1+/+ d71 r2 p1	**CAGCAG**CCGCCACCG	**[25, 22, 35, 34, 60, 57]**[][13, 34, 55, 27, 57, 50, 48, 71, 35]
4194604	136	FAN1+/+ d50 r1 p1	**CAGCAG**CCGCCACCG	**[66, 60, 57, 93, 93, 93]**[][93, 42, 93, 93, 28, 93, 86, 80, 93]
4194739	106	FAN1+/+ d37 r3 p2	**CAGCAG**CGCCGCCGC	**[93, 93, 93, 93, 93, 74]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4260112	122	FAN1−/− d16 r2 p1	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4260474	136	FAN1+/+ d71 r1 p2	**CAGCAG**CCGCCACCG	**[33, 23, 41, 19, 20, 24]**[][17, 8, 8, 30, 11, 9, 16, 16, 24]
4325597	137	FAN1−/− d71 r2 p1	**CAGCAG**CCGCCAACA	**[93, 93, 46, 93, 93, 93]**[][93, 93, 93, 3, 71, 60, 93, 93, 93]
4325641	56	FAN1−/− d37 r3 p1	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4325897	149	FAN1−/− d71 r2 p1	**CAGCAG**CGCCAGCCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 9, 93, 93, 93]
4325946	125	FAN1+/+ d37 r1 p1	**CAGCAG**CCCCGCCGC	**[54, 46, 43, 50, 43, 42]**[][52, 38, 60, 70, 40, 51, 71, 60, 43]
4326074	124	FAN1−/− d37 r3 p1	**CAGCAG**CCGCCACCG	**[78, 67, 12, 22, 91, 85]**[][55, 93, 93, 17, 93, 79, 59, 93, 93]
4391268	136	FAN1−/− d71 r1 p2	**CAGCAG**CGACAGCCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4391457	133	FAN1−/− d50 r2 p2	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4391806	130	FAN1+/+ d50 r3 p2	**CAGCAG**CCGCCCGCC	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4456756	73	FAN1+/+ d37 r3 p2	**CAGCAG**CCGCCACCG	**[57, 40, 44, 58, 32, 65]**[][69, 25, 65, 84, 35, 57, 72, 54, 66]
4456910	136	FAN1−/− d50 r3 p2	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4457141	140	FAN1+/+ d37 r3 p2	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4522344	114	FAN1+/+ d50 r1 p2	**CAGCAG**CCGCCGCCG	**[80, 93, 93, 93, 93, 93]**[][93, 93, 93, 85, 93, 93, 93, 93, 81]
4522697	131	FAN1−/− d50 r3 p1	**CAGCAG**CGACAGCCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4522753	127	FAN1+/+ d16 r1 p1	**CAGCAG**CCGCCAACC	**[93, 42, 93, 58, 28, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4588459	91	FAN1+/+ d37 r1 p1	**CAGCAG**CGACAGCCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4653312	132	FAN1−/− d37 r2 p1	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4653422	128	FAN1−/− d50 r1 p1	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 31]**[][38, 63, 12, 28, 93, 77, 41, 77, 53]
4653591	131	FAN1−/− d37 r2 p2	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4653642	150	FAN1+/+ d71 r1 p2	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4653838	131	FAN1+/+ d71 r2 p2	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4653846	84	FAN1+/+ d16 r2 p1	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4719083	126	FAN1+/+ d50 r1 p2	**CAGCAG**CCAACGCCG	**[60, 72, 56, 70, 91, 93]**[][88, 89, 62, 36, 71, 49, 38, 57, 73]
4719527	135	FAN1−/− d71 r2 p2	**CAGCAG**CCGCCACCG	**[93, 15, 62, 93, 45, 49]**[][93, 93, 93, 93, 93, 93, 93, 25, 76]
4784698	137	FAN1+/+ d37 r2 p1	**CAGCAG**CCGCCACCG	**[93, 93, 93, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]
4784851	145	FAN1−/− d16 r2 p1	**CAGCAG**CCGCCACCG	**[93, 64, 11, 93, 93, 93]**[][93, 93, 93, 93, 93, 93, 93, 93, 93]

B

Read ID	CAGs	Condition	Sequence	Per-base quality scores
4194650	127	FAN1−/− d37 r2 p1	**CAGCAG**CAAACAGCCG	**[93, 84, 93, 93, 93, 93][93]**[72, 61, 3, 35, 54, 81, 93, 93, 91]
4194777	132	FAN1+/+ d37 r2 p2	**CAGCAG**CAAACAGCCG	**[24, 62, 75, 50, 86, 93][79]**[31, 30, 93, 93, 93, 93, 93, 93, 93]
4194802	128	FAN1−/− d37 r1 p1	**CAGCAG**CAACTGCCGC	**[93, 93, 93, 93, 93, 93][93]**[93, 93, 93, 93, 93, 93, 93, 93, 93]
4195153	130	FAN1−/− d37 r1 p1	**CAGCAG**CAACACCCGC	**[93, 93, 93, 93, 93, 93][93]**[93, 93, 93, 93, 62, 93, 93, 93, 93]
4195248	155	FAN1+/+ d71 r2 p1	**CAGCAG**CAACGGCCGC	**[93, 93, 93, 93, 93, 93][93]**[93, 93, 93, 93, 93, 93, 93, 93, 93]
4260735	126	FAN1+/+ d37 r3 p2	**CAGCAG**CAAACAGCCG	**[34, 56, 67, 20, 60, 93][39]**[8, 24, 86, 93, 93, 93, 63, 93, 93]
4325890	134	FAN1−/− d71 r2 p1	**CAGCAG**CACAGCCGCC	**[93, 93, 93, 93, 93, 93][70]**[34, 85, 93, 93, 66, 93, 93, 33, 93]
4326325	123	FAN1−/− d37 r1 p1	**CAGCAG**CACAGCCGCC	**[93, 91, 93, 93, 51, 93][56]**[23, 42, 93, 93, 93, 93, 93, 15, 93]
4391709	123	FAN1+/+ d50 r2 p1	**CAGCAG**CAACTAGCCG	**[70, 78, 41, 28, 93, 82][55]**[60, 44, 29, 6, 26, 82, 60, 93, 76]
4457103	125	FAN1+/+ d50 r1 p1	**CAGCAG**CACAGCCGCC	**[93, 93, 93, 78, 53, 93][49]**[47, 93, 93, 93, 93, 93, 93, 32, 93]
4457117	135	FAN1−/− d50 r3 p1	**CAGCAG**CAAACAGCCG	**[93, 93, 93, 93, 93, 93][93]**[15, 63, 93, 93, 93, 93, 93, 93, 93]
4457280	130	FAN1+/+ d37 r3 p2	**CAGCAG**CAAACAGCCG	**[82, 50, 67, 93, 71, 73][93]**[11, 50, 93, 93, 77, 93, 74, 93, 93]
4457454	132	FAN1−/− d50 r3 p1	**CAGCAG**CAAACAGCCG	**[93, 93, 93, 93, 93, 93][93]**[93, 93, 93, 93, 93, 93, 93, 93, 93]
4522401	102	FAN1+/+ d16 r2 p1	**CAGCAG**CAACGGCCGC	**[83, 93, 10, 64, 60, 44][62]**[42, 70, 44, 69, 93, 88, 93, 66, 91]
4522678	135	FAN1−/− d50 r1 p2	**CAGCAG**CAAGCCGCCA	**[93, 93, 93, 93, 66, 93][93]**[74, 93, 93, 93, 93, 93, 93, 93, 93]
4522876	130	FAN1+/+ d37 r1 p1	**CAGCAG**CAACACCGCC	**[85, 25, 52, 93, 83, 93][93]**[47, 93, 79, 3, 36, 68, 93, 85, 93]
4587689	86	FAN1−/− d50 r3 p1	**CAGCAG**CACAGCCGCC	**[93, 93, 93, 93, 93, 93][93]**[63, 93, 93, 93, 93, 93, 93, 93, 93]
4588270	130	FAN1+/+ d71 r3 p2	**CAGCAG**CACAGCCGCC	**[54, 12, 17, 57, 83, 86][72]**[72, 90, 75, 76, 58, 49, 55, 52, 60]
4653155	130	FAN1+/+ d50 r1 p1	**CAGCAG**CAAGCCGCCA	**[93, 93, 93, 93, 93, 93][93]**[93, 93, 93, 93, 93, 93, 93, 93, 93]
4653326	130	FAN1−/− d50 r1 p2	**CAGCAG**CACAGCCACC	**[93, 93, 93, 93, 93, 93][93]**[93, 93, 93, 93, 93, 93, 93, 93, 93]
4653567	129	FAN1+/+ d16 r2 p2	**CAGCAG**CACAGCCGCC	**[64, 39, 36, 78, 74, 76][73]**[51, 65, 56, 52, 58, 51, 71, 69, 77]
4653775	137	FAN1−/− d71 r1 p2	**CAGCAG**CAAGCCGCCA	**[93, 80, 93, 59, 58, 91][32]**[17, 11, 58, 66, 93, 93, 64, 93, 93]
4718713	124	FAN1−/− d71 r2 p1	**CAGCAG**CACAGCCGCC	**[32, 93, 93, 76, 88, 28][93]**[3, 93, 93, 93, 93, 93, 93, 93, 93]
4718778	131	FAN1−/− d50 r3 p1	**CAGCAG**CACAGCCGCC	**[61, 93, 93, 82, 93, 93][93]**[26, 93, 93, 93, 93, 93, 93, 82, 93]
4719587	124	FAN1+/+ d71 r2 p2	**CAGCAG**CAAGCCGCCA	**[84, 61, 57, 45, 87, 85][93]**[93, 86, 92, 93, 78, 92, 93, 93, 93]
4784946	130	FAN1+/+ d37 r2 p1	**CAGCAG**CAAACAGCCG	**[35, 72, 54, 55, 73, 57][90]**[53, 36, 77, 60, 32, 82, 70, 63, 70]
4850092	124	FAN1+/+ d16 r1 p1	**CAGCAG**CACGCCGCCG	**[88, 4, 47, 53, 17, 79][93]**[71, 77, 90, 93, 93, 93, 93, 93, 93]
4850277	144	FAN1−/− d16 r3 p2	**CAGCAG**CAAACAGCCG	**[93, 93, 93, 93, 93, 93][93]**[80, 68, 93, 93, 93, 93, 93, 93, 93]
4850429	130	FAN1+/+ d50 r3 p1	**CAGCAG**CAACCAGCCG	**[23, 24, 45, 44, 59, 74][73]**[93, 93, 75, 86, 89, 85, 93, 87, 82]
4915594	70	FAN1+/+ d71 r2 p1	**CAGCAG**CAACAACGGC	**[93, 93, 93, 93, 93, 93][93]**[93, 93, 93, 93, 93, 93, 93, 93, 93]

C

Read ID	CAGs	Condition	Sequence	Per-base quality scores
4260044	135	FAN1−/− d71 r1 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4260268	129	FAN1+/+ d50 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4260495	131	FAN1+/+ d71 r1 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 81, 58, 67, 82, 62][62, 63, 92, 71, 93, 84, 83, 52, 92, 63, 89, 50][68, 93, 93, 48, 93, 65, 50, 85, 77]
4260778	119	FAN1+/+ d37 r3 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4260853	126	FAN1−/− d16 r3 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4326342	129	FAN1−/− d50 r2 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 91, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 84, 93][93, 93, 93, 93, 91, 85, 93, 93, 93]
4456843	130	FAN1+/+ d37 r3 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4457210	137	FAN1+/+ d16 r1 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4522117	129	FAN1+/+ d50 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 66, 93, 83]
4522654	129	FAN1+/+ d71 r3 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4522968	130	FAN1+/+ d50 r3 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4653294	134	FAN1−/− d71 r1 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 42, 93, 93, 93, 93, 3, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4653516	137	FAN1−/− d37 r3 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 80, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4849973	125	FAN1+/+ d37 r3 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4916026	130	FAN1+/+ d71 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4980870	130	FAN1−/− d50 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 60, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 49, 93, 93, 93]
5112004	133	FAN1−/− d50 r3 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5112363	113	FAN1−/− d37 r1 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5112662	129	FAN1−/− d37 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5242992	53	FAN1−/− d37 r3 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5243736	119	FAN1+/+ d71 r3 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[76, 62, 70, 82, 84, 72][93, 88, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5308838	121	FAN1+/+ d71 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[79, 93, 75, 70, 89, 93][93, 68, 44, 49, 68, 68, 79, 59, 66, 67, 54, 53][56, 62, 44, 29, 51, 42, 56, 60, 64]
5309406	125	FAN1+/+ d16 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5374638	132	FAN1+/+ d16 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5570887	132	FAN1−/− d16 r3 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 82][93, 93, 76, 91, 93, 93, 54, 93, 93, 89, 93, 93][93, 93, 93, 93, 93, 93, 90, 93, 93]
5636755	131	FAN1−/− d37 r2 p2	CAGCAGCAACAGCAACAGCCCCACCGC	[86, 37, 23, 72, 8, 53][93, 47, 86, 93, 93, 93, 31, 55, 93, 81, 67, 93][13, 93, 54, 63, 7, 42, 93, 93, 41]
5636901	141	FAN1+/+ d50 r1 p2	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5701954	129	FAN1+/+ d16 r3 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[92, 63, 73, 93, 76, 78][93, 71, 18, 54, 24, 93, 52, 30, 39, 74, 66, 53][64, 53, 60, 56, 64, 85, 56, 14, 54]
5702520	94	FAN1−/− d16 r1 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5702552	132	FAN1+/+ d71 r2 p1	CAGCAGCAACAGCAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]

D

Read ID	CAGs	Condition	Sequence	Per-base quality scores
4260662	130	FAN1−/− d37 r3 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4391622	130	FAN1+/+ d37 r3 p2	CAGCAGCAACAACAGCCGCCACCG	[33, 45, 57, 47, 59, 56][43, 54, 51, 42, 52, 53, 34, 41, 44][9, 9, 17, 43, 53, 53, 48, 43, 34]
4456565	132	FAN1−/− d37 r2 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4588183	135	FAN1+/+ d16 r3 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4718829	121	FAN1+/+ d37 r2 p1	CAGCAGCAACAACAGCCGCCACCG	[22, 52, 28, 48, 57, 50][28, 21, 32, 49, 64, 11, 34, 6, 28][54, 67, 56, 57, 65, 47, 48, 18, 49]
4719121	136	FAN1+/+ d71 r2 p1	CAGCAGCAACAACAGCCGCCACCG	[67, 93, 92, 57, 78, 58][72, 82, 73, 76, 92, 83, 75, 71, 82][87, 40, 75, 42, 25, 66, 64, 76, 90]
4719319	177	FAN1−/− d37 r2 p1	CAGCAGCAACAACAGCCGCCACCG	[42, 64, 82, 32, 66, 59][44, 69, 8, 67, 57, 6, 63, 25, 23][27, 65, 29, 38, 14, 55, 76, 93, 55]
4719514	130	FAN1−/− d50 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4784605	33	FAN1+/+ d37 r3 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4784754	120	FAN1+/+ d16 r1 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 83, 93, 93][92, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 43, 93, 93]
4915939	151	FAN1−/− d71 r3 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4981331	128	FAN1−/− d37 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[66, 93, 93, 64, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4981375	126	FAN1−/− d16 r3 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5309122	132	FAN1−/− d16 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5374933	134	FAN1+/+ d50 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[50, 93, 93, 65, 93, 93][67, 93, 93, 73, 93, 93, 89, 93, 93][68, 93, 93, 69, 93, 93, 93, 93, 93]
5439973	131	FAN1−/− d50 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5440185	123	FAN1+/+ d37 r1 p2	CAGCAGCAACAACAGCCGCCACCG	[57, 49, 71, 61, 40, 71][67, 55, 65, 51, 52, 61, 31, 8, 6][51, 71, 70, 73, 71, 70, 72, 71, 58]
5505393	125	FAN1+/+ d50 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5505644	131	FAN1−/− d16 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 58, 93][93, 93, 93, 93, 93, 93, 52, 93][93, 14, 93, 93, 44, 93, 93, 89, 93]
5505818	134	FAN1+/+ d37 r3 p2	CAGCAGCAACAACAGCCGCCACCG	[32, 61, 59, 42, 41, 82][81, 74, 93, 93, 80, 93, 93, 71, 93][93, 93, 93, 81, 93, 93, 64, 93, 93]
5505998	127	FAN1+/+ d50 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 89, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 91, 93, 93, 93, 93, 93, 85, 93]
5636324	128	FAN1+/+ d71 r3 p1	CAGCAGCAACAACAGCCGCCACCG	[91, 93, 93, 85, 93, 93][93, 17, 21, 69, 8, 93, 82, 43, 93][76, 93, 75, 61, 93, 65, 84, 93, 57]
5636826	127	FAN1−/− d16 r3 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5833686	103	FAN1+/+ d71 r1 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5899163	127	FAN1+/+ d71 r1 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5964509	131	FAN1−/− d50 r2 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5964601	134	FAN1−/− d16 r2 p2	CAGCAGCAACAACAGCCGCCACCG	[59, 93, 93, 93, 93, 93][93, 30, 57, 9, 90, 93, 93, 93, 93][93, 93, 93, 54, 93, 93, 69, 93, 93]
6029647	133	FAN1+/+ d71 r2 p1	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
6029952	127	FAN1+/+ d50 r1 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
6161003	127	FAN1−/− d50 r1 p2	CAGCAGCAACAACAGCCGCCACCG	[93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]

E

Read ID	CAGs	Condition	Sequence	Per-base quality scores
4325748	95	FAN1+/+ d37 r3 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 42, 60, 14, 87, 93, 93][79, 93, 93, 49, 35, 53, 12, 74, 35]
4457387	128	FAN1+/+ d16 r2 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[55, 44, 93, 66, 93, 93]**[66, 32, 6, 15, 14, 30, 93][74, 93, 93, 72, 93, 93, 73, 93, 93]
4653989	131	FAN1+/+ d50 r1 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
4915434	130	FAN1+/+ d37 r3 p1	**CAGCAG**CAAGCAGCCACCGCCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5112475	130	FAN1−/− d71 r2 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[82, 66, 62, 52, 28, 68]**[54, 21, 26, 13, 14, 15, 58][45, 26, 52, 24, 93, 86, 28, 81, 93]
5112649	130	FAN1−/− d50 r1 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 57, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5243504	127	FAN1+/+ d16 r3 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 34, 93, 89, 93][93, 93, 77, 91, 93, 93, 93, 93, 93]
5374359	136	FAN1−/− d50 r2 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 37, 93, 19, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5439728	129	FAN1−/− d16 r1 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 76, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
5571441	125	FAN1+/+ d71 r2 p2	**CAGCAG**CAAGCAGCCACCGCCC	**[93, 67, 93, 85, 78, 75]**[34, 7, 39, 49, 56, 10, 71][13, 29, 88, 56, 85, 78, 7, 62, 46]
5767617	143	FAN1−/− d50 r3 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 90, 93, 93, 48]**[93, 82, 93, 93, 93, 93, 93][93, 84, 93, 93, 93, 93, 89, 93]
5833328	129	FAN1−/− d50 r2 p1	**CAGCAG**CAAGCAGCCGCCACAG	**[93, 58, 69, 93, 92, 92]**[39, 27, 42, 12, 64, 75, 75][58, 59, 37, 33, 42, 42, 5, 44, 57]
5898423	84	FAN1+/+ d50 r1 p2	**CAGCAG**CAAGCAGCCGCCCACA	**[39, 41, 44, 42, 35, 39]**[38, 15, 10, 40, 41, 19, 48][26, 26, 16, 12, 10, 14, 7, 37, 34]
6095386	129	FAN1−/− d16 r1 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 57, 93]**[93, 66, 93, 53, 46, 66, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
6685233	122	FAN1−/− d37 r3 p1	**CAGCAG**CAAGCAGCCGCCACAG	**[93, 93, 93, 93, 93, 93]**[93, 63, 93, 66, 93, 93, 93][93, 93, 23, 40, 40, 90, 42, 93, 93]
6816428	109	FAN1+/+ d16 r1 p2	**CAGCAG**CAAGCAGCCAACCGCC	**[93, 93, 93, 93, 71, 75]**[93, 77, 93, 93, 85, 93, 93][93, 93, 93, 93, 93, 93, 53, 93, 93]
6816742	132	FAN1−/− d50 r3 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
6881754	97	FAN1−/− d71 r3 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
6882064	116	FAN1−/− d71 r3 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[26, 18, 28, 57, 65, 56]**[40, 19, 30, 4, 40, 46, 67][9, 54, 43, 43, 57, 48, 26, 22, 37]
6882200	132	FAN1+/+ d50 r3 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 20]**[93, 53, 75, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
6947712	122	FAN1−/− d37 r1 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 84, 93]**[78, 5, 61, 4, 87, 93, 93][77, 41, 93, 59, 93, 93, 23, 53, 46]
7144011	78	FAN1−/− d37 r2 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 31, 87, 3, 93, 93, 93][93, 78, 62, 93, 80, 58, 89, 60, 54]
7340168	102	FAN1+/+ d50 r2 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[49, 48, 88, 93, 41, 91]**[93, 23, 69, 5, 93, 47, 71][90, 93, 83, 87, 78, 50, 64, 93, 93]
7405969	130	FAN1−/− d71 r2 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 78, 43, 29, 93, 56, 93][93, 93, 93, 85, 93, 93, 93, 93, 93]
7406334	123	FAN1−/− d16 r1 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[93, 93, 93, 93, 93, 93]**[93, 39, 4, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
7406416	119	FAN1−/− d37 r2 p1	**CAGCAG**CAAGCAGCCGCCACAG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 70, 93][93, 93, 93, 17, 36, 80, 93, 93, 93]
7602966	140	FAN1+/+ d37 r3 p2	**CAGCAG**CAAGCAGCCGCCACCG	**[42, 11, 30, 39, 34, 47]**[43, 27, 58, 8, 64, 32, 34][27, 41, 23, 37, 33, 9, 34, 55, 38]
7603066	128	FAN1−/− d71 r1 p1	**CAGCAG**CAAGCAGCGACAGCCG	**[93, 93, 93, 93, 93, 93]**[93, 93, 93, 93, 93, 93, 93][93, 93, 93, 93, 93, 93, 93, 93, 93]
7930669	126	FAN1+/+ d71 r1 p1	**CAGCAG**CAAGCAGCCGCCACCG	**[39, 90, 83, 87, 93, 92]**[70, 30, 28, 4, 76, 37, 93][66, 93, 46, 72, 93, 52, 93, 93, 85]
7995856	131	FAN1−/− d50 r1 p2	**CAGCAG**CAAGCAGCCGCCAACA	**[78, 93, 93, 84, 93, 88]**[93, 87, 93, 78, 93, 93, 93][93, 44, 20, 60, 93, 27, 31, 72, 93]

Appendix 2: EMBOSS Water alignment of Sanger sequencing data generated using primers TOM54 and TOM55 on first-round 3 kb amplicons to *in silico* PCR product generated using primer sequences TOM48 and TOM49 on hg38. Pairwise sequence alignment using EMBOSS Water (Smith-Waterman algorithm) with default alignment parameters (Matrix: DNAfull, GAP OPEN: 10, GAP EXTEND: 0.5).

Forward sequence

```
########################################
# Program: water
# Rundate: Tue 12 Jun 2018 09:59:28
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20180612-095924-0647-28939696-p2m.asequence
#    -bsequence emboss_water-I20180612-095924-0647-28939696-p2m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: htt
# 2: 70EI43_09724556_09724556
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1053
# Identity:    1047/1053 (99.4%)
# Similarity:  1047/1053 (99.4%)
# Gaps:           3/1053 ( 0.3%)
# Score: 5193.0
#
#
```

```
#=====================================

htt            2624 GAACTTTATTTATTTATTTATTTTTGAGACAGAGTCTCACTCTTGTCACC   2673
                    |.||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     1 GGACTTTATTTATTTATTTATTTTTGAGACAGAGTCTCACTCTTGTCACC     50

htt            2674 CAGGCTGGAGTGCAATGGCATGATCTTGGCTCACTGCAACCTCCACCTCC   2723
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245    51 CAGGCTGGAGTGCAATGGCATGATCTTGGCTCACTGCAACCTCCACCTCC    100

htt            2724 CAGGTTCAAGCAATTCTGCCTCAGCCTCCGGAATAGCTGGGACTACAGGC   2773
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   101 CAGGTTCAAGCAATTCTGCCTCAGCCTCCGGAATAGCTGGGACTACAGGC    150

htt            2774 ATGCACCACTACACCCGGCTAATTTTTGTATTTTTAGTAGAGACAGGGTT   2823
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   151 ATGCACCACTACACCCGGCTAATTTTTGTATTTTTAGTAGAGACAGGGTT    200

htt            2824 TCGCCATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCTGGTGATCTGC   2873
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   201 TCGCCATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCTGGTGATCTGC    250

htt            2874 CTGCCTTGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCGCACC   2923
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   251 CTGCCTTGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCCACCGCACC    300

htt            2924 TCGCTGGAACTTAATTTTTTTAGAGACAGTGTCGCTCTATCACCCAAGCT   2973
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   301 TCGCTGGAACTTAATTTTTTTAGAGACAGTGTCGCTCTATCACCCAAGCT    350

htt            2974 GGAGTGCAGTGGTGCAATCCTAGCTCACTTGCAGCCTCAAATTCCTGGGT   3023
                    |.||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   351 GTAGTGCAGTGGTGCAATCCTAGCTCACTTGCAGCCTCAAATTCCTGGGT    400

htt            3024 TCAGGTGATCCTCCCACATCAGCCTCCCAAGAACTGGGAACTAACAGCTG   3073
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   401 TCAGGTGATCCTCCCACATCAGCCTCCCAAGAACTGGGAACTAACAGCTG    450

htt            3074 TTTCTCTGCTGTCCTTCTCAAGAAAAGGGAGGCTACTGCTACCCCACTGG   3123
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245   451 TTTCTCTGCTGTCCTTCTCAAGAAAAGGGAGGCTACTGCTACCCCACTGG    500

htt            3124 GGACAATGCTGGGTTTCCCTTTAGGACAGGCTCTGAGACAAGGCGGAGGT   3173
```

```
                     ||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     501 GGACAATGCTGGGTTTCCCTTTAGGACAGGCTCTGAGACAAGGCGGAGGT    550

htt              3174 GCTGTTTGTGGCCACAGAGCAGGGGACTCTGGGTTGCAGGTGTGGCCTGG   3223
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     551 GCTGTTTGTGGCCACAGAGCAGGGGACTCTGGGTTGCAGGTGTGGCCTGG    600

htt              3224 CTAAAGTAGGCTTTACTGGGCTCCTCTCTGCCTGCATCACCCCCCGGCTG   3273
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     601 CTAAAGTAGGCTTTACTGGGCTCCTCTCTGCCTGCATCACCCCCCGGCTG    650

htt              3274 GGCGGTTGTCTCTGAGGCCAACCTTACTCCCTGCTGGGCAGGCTGGACAG   3323
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     651 GGCGGTTGTCTCTGAGGCCAACCTTACTCCCTGCTGGGCAGGCTGGACAG    700

htt              3324 CTGCCCTCTCCGTTTGCCCCTCTACCACCCAAAAGGCAGGAGGCTCTGGA   3373
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     701 CTGCCCTCTCCGTTTGCCCCTCTACCACCCAAAAGGCAGGAGGCTCTGGA    750

htt              3374 GACCAGGACCCTGCCCGCCACGGCCTGTGTCCCAGGCGTGAGGGGGTGCC   3423
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     751 GACCAGGACCCTGCCCGCCACGGCCTGTGTCCCAGGCGTGAGGGGGTGCC    800

htt              3424 CCACAGACCTCTGCTGAGCTGCTGCTGAATGACGCCCCTTGGGGGTCCTG   3473
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     801 CCACAGACCTCTGCTGAGCTGCTGCTGAATGACGCCCCTTGGGGGTCCTG    850

htt              3474 CCGGAAGGTCAGAGCAGGGGTGCACTCCCATAAAGAAACGCCCCCAGGTC   3523
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     851 CCGGAAGGTCAGAGCAGGGGTGCACTCCCATAAAGAAACGCCCCCAGGTC    900

htt              3524 GGGACTCATTCCTGTGGGCGGCATCTTGTGGCCATAGCTGCTTCTCGCTG   3573
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     901 GGGACTCATTCCTGTGGGCGGCATCTTGTGGCCATAGCTGCTTCTCGCTG    950

htt              3574 CACTAATCACAGTGCCTCTGTGGGCAGCAGGCGCTGACCACCCAGGCCTG   3623
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
70EI43_097245     951 CACTAATCACAGTGCCTCTGTGGGCAGCAGGCGCTGACCACCCAGGCCTG   1000

htt              3624 CCCC-AGACCCTCTCCTCCCTTCC-GGGGCGCTGCGCTGGG-ACCGATGG   3670
                     |||| |.|||||||||||||||||| |||||||||||||||| ||||||||
70EI43_097245    1001 CCCCAAAACCCTCTCCTCCCTTCCGGGGGCGCTGCGCTGGGAACCGATGG   1050
```

```
htt             3671 GGG   3673
                     |||
70EI43_097245   1051 GGG   1053


#---------------------------------------
#---------------------------------------
```

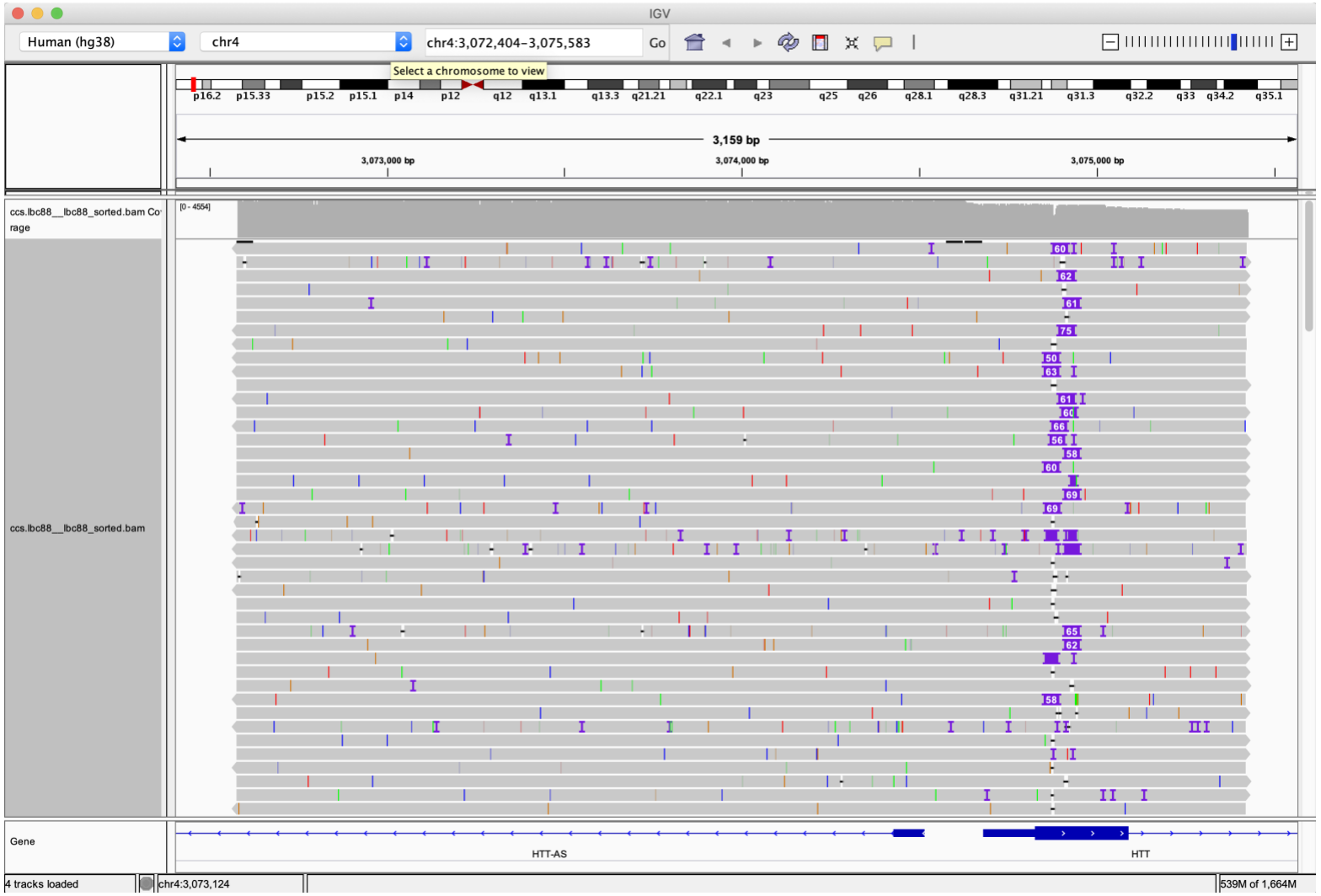
Reverse sequence

```
########################################
# Program: water
# Rundate: Tue 12 Jun 2018 10:21:10
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20180612-102108-0900-14301780-p2m.asequence
#    -bsequence emboss_water-I20180612-102108-0900-14301780-p2m.bsequence
#    -datafile EDNAFULL
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -snucleotide1
#    -snucleotide2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: hg38_dna_htt_locus
# 2: 70EI44_09724945_09724945
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 506
# Identity:     505/506 (99.8%)
# Similarity:   505/506 (99.8%)
# Gaps:           0/506 ( 0.0%)
# Score: 2521.0
```

```
#
#
#=======================================

hg38_dna_htt_   4879 AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG   4928
                     ||.||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249      3 AGTAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG     52

hg38_dna_htt_   4929 CAGCAGCAACAGCCGCCACCGCCGCCGCCGCCGCCGCCTCCTCAGCT   4978
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249     53 CAGCAGCAACAGCCGCCACCGCCGCCGCCGCCGCCGCCTCCTCAGCT    102

hg38_dna_htt_   4979 TCCTCAGCCGCCGCCGCAGGCACAGCCGCTGCTGCCTCAGCCGCAGCCGC   5028
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    103 TCCTCAGCCGCCGCCGCAGGCACAGCCGCTGCTGCCTCAGCCGCAGCCGC    152

hg38_dna_htt_   5029 CCCCGCCGCCGCCCCCGCCGCCACCCGGCCCGGCTGTGGCTGAGGAGCCG   5078
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    153 CCCCGCCGCCGCCCCCGCCGCCACCCGGCCCGGCTGTGGCTGAGGAGCCG    202

hg38_dna_htt_   5079 CTGCACCGACCGTGAGTTTGGGCCCGCTGCAGCTCCCTGTCCCGGCGGGT   5128
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    203 CTGCACCGACCGTGAGTTTGGGCCCGCTGCAGCTCCCTGTCCCGGCGGGT    252

hg38_dna_htt_   5129 CCCAGGCTACGGCGGGGATGGCGGTAACCCTGCAGCCTGCGGGCCGGCGA   5178
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    253 CCCAGGCTACGGCGGGGATGGCGGTAACCCTGCAGCCTGCGGGCCGGCGA    302

hg38_dna_htt_   5179 CACGAACCCCCGGCCCCGCAGAGACAGAGTGACCCAGCAACCCAGAGCCC   5228
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    303 CACGAACCCCCGGCCCCGCAGAGACAGAGTGACCCAGCAACCCAGAGCCC    352

hg38_dna_htt_   5229 ATGAGGGACACCCGCCCCCTCCTGGGGCGAGGCCTTCCCCCACTTCAGCC   5278
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    353 ATGAGGGACACCCGCCCCCTCCTGGGGCGAGGCCTTCCCCCACTTCAGCC    402

hg38_dna_htt_   5279 CCGCTCCCTCACTTGGGTCTTCCCTTGTCCTCTCGCGAGGGGAGGCAGAG   5328
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    403 CCGCTCCCTCACTTGGGTCTTCCCTTGTCCTCTCGCGAGGGGAGGCAGAG    452

hg38_dna_htt_   5329 CCTTGTTGGGGCCTGTCCTGAATTCACCGAGGGGAGTCACGGCCTCAGCC   5378
                     ||||||||||||||||||||||||||||||||||||||||||||||||
70EI44_097249    453 CCTTGTTGGGGCCTGTCCTGAATTCACCGAGGGGAGTCACGGCCTCAGCC    502
```

```
hg38_dna_htt_   5379 CTCTCG   5384
                     ||||||
70EI44_097249    503 CTCTCG    508


#---------------------------------------
#---------------------------------------
```

Appendix 3: Minimap2 alignment of PacBio sequencing data to hg38 viewed in IGV. Default parameters used (see here for details: https://lh3.github.io/minimap2/minimap2.html). Thick purple lines are insertions. Multi-coloured lines are mismatches. Gaps are deletions.



References

Adegbuyiro, A. et al. 2017. Proteins Containing Expanded Polyglutamine Tracts and Neurodegenerative Disease. *Biochemistry* 56(9), pp. 1199–1217. doi: 10.1021/acs.biochem.6b00936.

Alexander, G.E. 1994. Basal ganglia-thalamocortical circuits: Their role in control of movements. *Journal of Clinical Neurophysiology* . doi: 10.1097/00004691-199407000-00004.

Andrew, S.E. et al. 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nature Genetics* 4(4), pp. 398–403. doi: 10.1038/ng0893-398.

Ardui, S. et al. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research* 46(5), pp. 2159–2168. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29401301 [Accessed: 23 September 2020].

Arnulf, I. et al. 2008. Rapid Eye Movement Sleep Disturbances in Huntington Disease. *Archives of Neurology* 65(4), p. 482. doi: 10.1001/archneur.65.4.482.

Arrasate, M. et al. 2004. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature* 431(7010), pp. 805–10. doi: 10.1038/nature02998.

Aylward, E.H. et al. 1998. Frontal lobe volume in patients with Huntington's disease. *Neurology* 50(1), pp. 252–8. doi: 10.1212/wnl.50.1.252.

Aziz, N.A. et al. 2008. Weight loss in Huntington disease increases with higher CAG repeat number. *Neurology* 71(19), pp. 1506–13. doi: 10.1212/01.wnl.0000334276.09729.0e.

Aziz, N.A. et al. 2011. Parent-of-origin differences of mutant HTT CAG repeat instability in Huntington's disease. *European Journal of Medical Genetics* 54(4), pp. e413–e418. doi: 10.1016/j.ejmg.2011.04.002.

Basu, S. et al. 2010. Purification of specific cell population by fluorescence activated cell sorting (FACS). *Journal of visualized experiments : JoVE* (41). Available at: http://www.ncbi.nlm.nih.gov/pubmed/20644514 [Accessed: 14 February 2022].

References

Bates, G.P. et al. 2014. *Huntington's Disease*. 4th ed. Oxford University Press.

Bates, G.P. et al. 2015. Huntington disease. *Nature Reviews Disease Primers* 1(1), p. 15005. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27188817 [Accessed: 14 August 2018].

Bäuerlein, F.J.B. et al. 2017. In Situ Architecture and Cellular Interactions of PolyQ Inclusions. *Cell* 171(1), pp. 179-187.e10. doi: 10.1016/j.cell.2017.08.009.

Bettencourt, C. et al. 2016. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Annals of Neurology* 79(6), pp. 983–990. Available at: http://doi.wiley.com/10.1002/ana.24656 [Accessed: 15 August 2018].

Bruyn, G.W. 1968. Huntington's chorea: historical, clinical and laboratory synopsis. *Handbook of Clinical Neurology* 6, pp. 298–378.

Budworth, H. et al. 2015. Suppression of Somatic Expansion Delays the Onset of Pathophysiology in a Mouse Model of Huntington's Disease. *PLOS Genetics* 11(8), pp. 1–22. doi: 10.1371/journal.pgen.1005267.

Cahan, P. and Daley, G.Q. 2013. Origins and implications of pluripotent stem cell variability and heterogeneity. *Nature Reviews Molecular Cell Biology* . doi: 10.1038/nrm3584.

Camnasio, S. et al. 2012. The first reported generation of several induced pluripotent stem cell lines from homozygous and heterozygous Huntington's disease patients demonstrates mutation related enhanced lysosomal activity. *Neurobiology of Disease* . doi: 10.1016/j.nbd.2011.12.042.

Le Cann, K. et al. 2021. The difficulty to model Huntington's disease in vitro using striatal medium spiny neurons differentiated from human induced pluripotent stem cells. *Scientific reports* 11(1), p. 6934. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33767215 [Accessed: 14 February 2022].

Cannavo, E. et al. 2007. Characterization of the interactome of the human MutL homologues MLH1, PMS1, and PMS2. *The Journal of Biological Chemistry* 282(5), pp. 2976–86. doi: 10.1074/jbc.M609989200.

Cardoso, F. 2017. Nonmotor Symptoms in Huntington Disease. *International Review of Neurobiology* 134, pp. 1397–1408. doi: 10.1016/bs.irn.2017.05.004.

References

Charles, P. et al. 2007. Are interrupted SCA2 CAG repeat expansions responsible for parkinsonism? *Neurology* 69(21), pp. 1970–5. doi: 10.1212/01.wnl.0000269323.21969.db.

Chaudhury, I. et al. 2014. FANCD2-controlled chromatin access of the Fanconi-associated nuclease FAN1 is crucial for the recovery of stalled replication forks. *Molecular and Cellular Biology* 34(21), pp. 3939–54. doi: 10.1128/MCB.00457-14.

Chintalaphani, S.R. et al. 2021. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathologica Communications* 9(1), p. 98. Available at: https://actaneurocomms.biomedcentral.com/articles/10.1186/s40478-021-01201-x [Accessed: 21 June 2021].

Choudhry, S. et al. 2001. CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. 10(21), pp. 2437–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11689490 [Accessed: 23 January 2022].

Ciosi, M. et al. 2018. Library preparation and MiSeq sequencing for the genotyping-by-sequencing of the Huntington disease HTT exon one trinucleotide repeat and the quantification of somatic mosaicism. *Protocol Exchange* . Available at: http://www.nature.com/protocolexchange/protocols/6621 [Accessed: 19 July 2019].

Ciosi, M. et al. 2019. A genetic association study of glutamine-encoding DNA sequence structures, somatic CAG expansion, and DNA repair gene variants, with Huntington disease clinical outcomes. *EBioMedicine* 48, pp. 568–580. Available at: https://pubmed.ncbi.nlm.nih.gov/31607598/ [Accessed: 12 May 2020].

Ciosi, M. et al. 2021. Approaches to Sequence the HTT CAG Repeat Expansion and Quantify Repeat Length Variation. *Journal of Huntington's disease* 10(1), pp. 53–74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33579864 [Accessed: 2 January 2022].

Covarrubias-Pazaran, G. et al. 2016. Fragman: an R package for fragment analysis. *BMC Genetics* 17(1), p. 62. Available at: http://bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0365-6 [Accessed: 22 November 2021].

Craufurd, D. et al. 2015. Diagnostic genetic testing for Huntington's disease. *Practical neurology* 15(1), pp. 80–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25169240 [Accessed: 19 January 2022].

Daunay, A. et al. 2019. Low temperature isothermal amplification of microsatellites drastically reduces stutter artifact formation and improves microsatellite instability detection in cancer. *Nucleic acids research* 47(21), p. e141. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31584085 [Accessed: 3 January 2022].

Davenport, C. 1912. Heredity in Relation to Eugenics. *The Annals of the American Academy of Political and Social Science* . doi: https://doi.org/10.1177/000271621204200148.

Davies, S.W. et al. 1997. Formation of Neuronal Intranuclear Inclusions Underlies the Neurological Dysfunction in Mice Transgenic for the HD Mutation. *Cell* 90(3), pp. 537–548. doi: 10.1016/S0092-8674(00)80513-9.

Díaz-Gay, M. and Alexandrov, L.B. 2021. Unraveling the genomic landscape of colorectal cancer through mutational signatures. *Advances in cancer research* 151, pp. 385–424. Available at: http://www.ncbi.nlm.nih.gov/pubmed/34148618 [Accessed: 13 February 2022].

DiFiglia, M. et al. 1997. Aggregation of Huntingtin in Neuronal Intranuclear Inclusions and Dystrophic Neurites in Brain. *Science* 277(5334), pp. 1990–1993. Available at: https://www.science.org/doi/10.1126/science.277.5334.1990 [Accessed: 10 January 2022].

Dimos, J.T. et al. 2008. Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. *Science* . doi: 10.1126/science.1158799.

Donaldson, J. et al. 2021. What is the Pathogenic CAG Expansion Length in Huntington's Disease? *Journal of Huntington's disease* 10(1), pp. 175–202. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33579866 [Accessed: 26 January 2022].

Donaldson, J.J. 2019. *Huntingtin CAG repeat expansions in induced pluripotent stem cell models of Huntington's Disease*. Available at: https://orca.cardiff.ac.uk/129860/1/2020Donaldson J PhD.pdf [Accessed: 21 December 2021].

Dorsman, J.C. et al. 2002. Interruption of perfect CAG repeats by CAA triplets improves the stability of glutamine-encoding repeat sequences. *BioTechniques* 33(5), pp. 976–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12449369 [Accessed: 11 February 2022].

Douglas, I. et al. 2013. Juvenile Huntington's disease: a population-based study using the General Practice Research Database. *BMJ Open* 3(4), p. e002085. doi: 10.1136/bmjopen-2012-002085.

van Duijn, E. et al. 2007. Psychopathology in verified Huntington's disease gene carriers. *The Journal of Neuropsychiatry and Clinical Neurosciences* 19(4), pp. 441–8. doi: 10.1176/jnp.2007.19.4.441.

van Duijn, E. et al. 2008. Cross-sectional study on prevalences of psychiatric disorders in mutation carriers of Huntington's disease compared with mutation-negative first-degree relatives. *The Journal of clinical psychiatry* 69(11), pp. 1804–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19026253 [Accessed: 8 January 2022].

Duyao, M. et al. 1995. Inactivation of the mouse Huntington's disease gene homolog Hdh. *Science* 269(5222), pp. 407–410. doi: 10.1126/science.7618107.

Ebbert, M.T.W. et al. 2018. Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Molecular neurodegeneration* 13(1), p. 46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30126445 [Accessed: 25 January 2022].

Fang, L. and Wang, K. 2018. Identification of copy number variants from SNP arrays using PennCNV. In: *Methods in Molecular Biology*. doi: 10.1007/978-1-4939-8666-8_1.

Findlay Black, H. et al. 2020. Frequency of the loss of CAA interruption in the HTT CAG tract and implications for Huntington disease in the reduced penetrance range. *Genetics in Medicine* 22(12), pp. 2108–2113. Available at: http://www.ncbi.nlm.nih.gov/pubmed/32741964 [Accessed: 23 January 2022].

Fisher, E.M.C. and Bannerman, D.M. 2019. Mouse models of neurodegeneration: Know your question, know your mouse. *Science Translational Medicine* 11(493). Available at: https://www.science.org/doi/10.1126/scitranslmed.aaq1818 [Accessed: 1 February 2022].

Foox, J. et al. 2021. Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study. *Nature Biotechnology* 39(9), pp. 1129–1140. Available at: https://www.nature.com/articles/s41587-021-01049-5 [Accessed: 30 November 2021].

Fungtammasan, A. et al. 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research* 25(5), p. 736. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4417121/#!po=16.6667 [Accessed: 19 May 2020].

Gacy, A.M. et al. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* 81(4), pp. 533–40. doi: 10.1016/0092-8674(95)90074-8.

Gao, R. et al. 2008. Instability of expanded CAG/CAA repeats in spinocerebellar ataxia type 17. 16(2), pp. 215–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18043721 [Accessed: 23 January 2022].

GeM-HD Consortium, G.M. of H.D. 2019. Huntington's disease onset is determined by length of uninterrupted CAG, not encoded polyglutamine, and is modified by DNA maintenance mechanisms. *Cell* . doi: 10.1101/529768.

Giesselmann, P. et al. 2019. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nature Biotechnology* 37(12), pp. 1478–1481. Available at: http://www.nature.com/articles/s41587-019-0293-x [Accessed: 19 June 2020].

Gilpatrick, T. et al. 2020. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature biotechnology* 38(4), pp. 433–438. Available at: http://www.ncbi.nlm.nih.gov/pubmed/32042167 [Accessed: 2 January 2022].

Goffredo, D. et al. 2002. Calcium-dependent Cleavage of Endogenous Wild-type Huntingtin in Primary Cortical Neurons. *Journal of Biological Chemistry* 277(42), pp. 39594–39598. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0021925819723230 [Accessed: 10 January 2022].

Gold, B. 2017. Somatic mutations in cancer: Stochastic versus predictable. *Mutation research. Genetic toxicology and environmental mutagenesis* 814, pp. 37–46.

Available at: http://www.ncbi.nlm.nih.gov/pubmed/28137366 [Accessed: 13 February 2022].

Goold, R. et al. 2019. FAN1 modifies Huntington's disease progression by stabilizing the expanded *HTT* CAG repeat. *Human Molecular Genetics* 28(4), pp. 650–661. Available at: https://academic.oup.com/hmg/article/28/4/650/5144444 [Accessed: 11 June 2020].

Goold, R. et al. 2021. FAN1 controls mismatch repair complex assembly via MLH1 retention to stabilize CAG repeat expansion in Huntington's disease. *Cell reports* 36(9), p. 109649. doi: 10.1016/j.celrep.2021.109649.

Graybiel, A.M. 1998. The basal ganglia and chunking of action repertoires. In: *Neurobiology of Learning and Memory*. doi: 10.1006/nlme.1998.3843.

Gusella, J.F. et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306(5940), pp. 234–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6316146 [Accessed: 7 January 2022].

Gutekunst, C.A. et al. 1999. Nuclear and neuropil aggregates in Huntington's disease: relationship to neuropathology. *The Journal of Neuroscience* 19(7), pp. 2522–34.

Hafford-Tear, N.J. et al. 2019. CRISPR/Cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. *Genetics in Medicine* 21(9), pp. 2092–2102. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1098360021049996 [Accessed: 2 February 2022].

Höijer, I. et al. 2018. Detailed analysis of *HTT* repeat elements in human blood using targeted amplification-free long-read sequencing. *Human Mutation* 39(9), pp. 1262–1272. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29932473 [Accessed: 2 January 2022].

Hommelsheim, C.M. et al. 2015. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. *Scientific Reports* 4(1), p. 5052. Available at: http://www.nature.com/articles/srep05052 [Accessed: 20 May 2020].

Hsu, R.-J. et al. 2011. Long tract of untranslated CAG repeats is deleterious in transgenic mice. 6(1), p. e16417. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/21283659 [Accessed: 14 February 2022].

Hu, S. et al. 2016. Effects of cellular origin on differentiation of human induced pluripotent stem cell–derived endothelial cells. *JCI Insight* . doi: 10.1172/jci.insight.85558.

Huin, V. et al. 2021. Motor neuron pathology in CANVAS due to *RFC1* expansions. *Brain* . Available at: https://academic.oup.com/brain/advance-article/doi/10.1093/brain/awab449/6470371 [Accessed: 19 January 2022].

Huntington, G. 1872. On Chorea. *The Medical and surgical Reporter* 26(15), pp. 317–321. Available at: http://psychiatryonline.org/doi/abs/10.1176/jnp.15.1.109 [Accessed: 6 January 2022].

Jain, M. et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* 36(4), pp. 338–345. Available at: http://www.nature.com/doifinder/10.1038/nbt.4060 [Accessed: 14 August 2018].

Jama, M. et al. 2013. Triplet Repeat Primed PCR Simplifies Testing for Huntington Disease. *The Journal of Molecular Diagnostics* 15(2), pp. 255–262. Available at: https://www.sciencedirect.com/science/article/pii/S152515781200308X?via%3Dihub#bib31 [Accessed: 1 February 2022].

Johnson, E.B. et al. 2021. Dynamics of Cortical Degeneration Over a Decade in Huntington's Disease. *Biological psychiatry* 89(8), pp. 807–816. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33500176 [Accessed: 10 January 2022].

Kadyrov, F.A. et al. 2006. Endonucleolytic Function of MutLα in Human Mismatch Repair. *Cell* 126(2), pp. 297–308. Available at: https://www.sciencedirect.com/science/article/pii/S0092867406008129 [Accessed: 7 April 2022].

Kassubek, J. et al. 2004. Topography of cerebral atrophy in early Huntington's disease: A voxel based morphometric MRI study. *Journal of Neurology, Neurosurgery and Psychiatry*

Kaye, J. et al. 2021. Huntington's disease mouse models: unraveling the pathology caused by CAG repeat expansion. *Faculty reviews* 10, p. 77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/34746930 [Accessed: 24 January 2022].

Kennedy, L. et al. 2003. Dramatic tissue-specific mutation length increases are an

early molecular event in Huntington disease pathogenesis. *Human Molecular Genetics* 12(24), pp. 3359–3367. Available at: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddg352 [Accessed: 20 May 2020].

Kent, W.J. et al. 2002. The Human Genome Browser at UCSC. *Genome Research* 12(6), pp. 996–1006. Available at: https://genome.cshlp.org/content/12/6/996.abstract [Accessed: 14 May 2020].

Khristich, A.N. and Mirkin, S.M. 2020. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. *The Journal of biological chemistry* 295(13), pp. 4134–4170. Available at: http://www.ncbi.nlm.nih.gov/pubmed/32060097 [Accessed: 7 January 2022].

Kim, M. et al. 1999. Mutant huntingtin expression in clonal striatal cells: dissociation of inclusion formation and neuronal survival by caspase inhibition. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 19(3), pp. 964–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9920660 [Accessed: 10 January 2022].

Kita, H. and Kitai, S.T. 1988. Glutamate decarboxylase immunoreactive neurons in rat neostriatum: their morphological types and populations. *Brain Research* 447(2), pp. 346–52. doi: 10.1016/0006-8993(88)91138-9.

Kuemmerle, S. et al. 1999. Huntington aggregates may not predict neuronal death in Huntington's disease. *Annals of Neurology* 46(6), pp. 842–9.

Kumar, K.R. et al. 2019. Next-Generation Sequencing and Emerging Technologies. *Semin Thromb Hemost* 45, pp. 661–673. Available at: https://doi.org/ [Accessed: 25 September 2020].

de la Monte, S.M. et al. 1988. Morphometric demonstration of atrophic changes in the cerebral cortex, white matter, and neostriatum in Huntington's disease. *Journal of Neuropathology and Experimental Neurology* 47(5), pp. 516–25. doi: 10.1097/00005072-198809000-00003.

Lang, W.H. et al. 2011. Conformational trapping of mismatch recognition complex MSH2/MSH3 on repair-resistant DNA loops. *Proceedings of the National Academy of Sciences of the United States of America* . doi: 10.1073/pnas.1105461108.

Langbehn, D.R. et al. 2019. Association of CAG Repeats With Long-term Progression

in Huntington Disease. 76(11), p. [ahead of print]. Available at: https://pubmed.ncbi.nlm.nih.gov/31403680/ [Accessed: 10 January 2022].

Langley, C. et al. 2021. Fronto-striatal circuits for cognitive flexibility in far from onset Huntington's disease: evidence from the Young Adult Study. *Journal of neurology, neurosurgery, and psychiatry* 92(2), pp. 143–149. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33130575 [Accessed: 19 January 2022].

Lawlor, K.T. et al. 2011. Double-stranded RNA is pathogenic in Drosophila models of expanded repeat neurodegenerative diseases. *Human Molecular Genetics* 20(19), pp. 3757–3768. Available at: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddr292 [Accessed: 14 February 2022].

Lazar, A.S. et al. 2015. Sleep deficits but no metabolic deficits in premanifest H untington's disease. *Annals of Neurology* 78(4), pp. 630–648. Available at: https://onlinelibrary.wiley.com/doi/10.1002/ana.24495 [Accessed: 8 January 2022].

Lee, J.-M. et al. 2010. A novel approach to investigate tissue-specific trinucleotide repeat instability. *BMC systems biology* 4, p. 29. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20302627 [Accessed: 8 April 2022].

Lee, J.-M. et al. 2011. Quantification of Age-Dependent Somatic CAG Repeat Instability in Hdh CAG Knock-In Mice Reveals Different Expansion Dynamics in Striatum and Liver. Nollen, E. A. A. ed. *PLoS ONE* 6(8), p. e23647. Available at: https://dx.plos.org/10.1371/journal.pone.0023647 [Accessed: 20 May 2020].

Lee, J.-M. et al. 2015. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell* 162(3), pp. 516–526. Available at: https://www.sciencedirect.com/science/article/pii/S0092867415008405 [Accessed: 14 August 2018].

Lee, J.-M. et al. 2019. CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell* 178(4), pp. 887-900.e14. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0092867419307391 [Accessed: 12 November 2019].

Leeflang, E. et al. 1995. Single sperm analysis of the trinucleotide repeats in the huntington's disease gene: Quantification of the mutation frequency spectrum. *Human Molecular Genetics* . doi: 10.1093/hmg/4.9.1519.

Lemiere, J. et al. 2004. Cognitive changes in patients with Huntington?s disease (HD) and asymptomatic carriers of the HD mutation. *Journal of Neurology* 251(8), pp. 935–942. Available at: http://link.springer.com/10.1007/s00415-004-0461-9 [Accessed: 8 January 2022].

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Birol, I. ed. *Bioinformatics* 34(18), pp. 3094–3100. Available at: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778 [Accessed: 6 February 2022].

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14), pp. 1754–60. doi: 10.1093/bioinformatics/btp324.

Liquori, C.L. et al. 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. 293(5531), pp. 864–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11486088 [Accessed: 1 February 2022].

Liu, K.-Y. et al. 2015. Disruption of the nuclear membrane by perinuclear inclusions of mutant huntingtin causes cell-cycle re-entry and striatal cell death in mouse and cell models of Huntington's disease. *Human Molecular Genetics* 24(6), pp. 1602–16. doi: 10.1093/hmg/ddu574.

Liu, Q. et al. 2020. Genome-wide detection of short tandem repeat expansions by long-read sequencing. *BMC bioinformatics* 21(Suppl 21), p. 542. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33371889 [Accessed: 22 June 2021].

Liu, T. et al. 2010. FAN1 acts with FANCI-FANCD2 to promote DNA interstrand cross-link repair. 329(5992), pp. 693–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20671156 [Accessed: 20 January 2022].

Losekoot, M. et al. 2013. EMQN/CMGS best practice guidelines for the molecular genetic testing of Huntington disease. *European Journal of Human Genetics* 21(5), pp. 480–486. Available at: http://www.nature.com/articles/ejhg2012200 [Accessed: 2 January 2022].

Loupe, J.M. et al. 2020. Promotion of somatic CAG repeat expansion by Fan1 knock-out in Huntington's disease knock-in mice is blocked by Mlh1 knock-out. *Human molecular genetics* 29(18), pp. 3044–3053. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/32876667 [Accessed: 9 January 2022].

Macaulay, I.C. et al. 2015. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods* 12(6), pp. 519–522. Available at: http://www.nature.com/articles/nmeth.3370 [Accessed: 14 August 2018].

Macaulay, I.C. et al. 2016. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nature Protocols* 11(11), pp. 2081–2103. doi: 10.1038/nprot.2016.138.

MacDonald, M.E. et al. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72(6), pp. 971–983. Available at: https://www.sciencedirect.com/science/article/pii/009286749390585E [Accessed: 14 August 2018].

Madeira, F. et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research* 47(W1), pp. W636–W641. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30976793 [Accessed: 6 February 2022].

Maiuri, T. et al. 2017. Huntingtin is a scaffolding protein in the ATM oxidative DNA damage response complex. *Human Molecular Genetics* 26(2), pp. 395–406. doi: 10.1093/hmg/ddw395.

Mamedov, T.G. et al. 2008. A fundamental study of the PCR amplification of GC-rich DNA templates. *Computational Biology and Chemistry* 32(6), pp. 452–457. Available at: https://www.sciencedirect.com/science/article/pii/S1476927108000881#bib13 [Accessed: 20 May 2020].

Mangiarini, L. et al. 1997. Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nature Genetics* . doi: 10.1038/ng0297-197.

Mantere, T. et al. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in genetics* 10, p. 426. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31134132 [Accessed: 24 January 2022].

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Massey, T. et al. 2018. Methods for Assessing DNA Repair and Repeat Expansion in

Huntington’s Disease. *Methods in molecular biology (Clifton, N.J.)* 1780, pp. 483–495. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29856032 [Accessed: 15 February 2022].

Massey, T.H. and Jones, L. 2018. The central role of DNA damage and repair in CAG repeat diseases. *Disease Models & Mechanisms* 11(1), p. dmm031930. doi: 10.1242/dmm.031930.

Mattis, V.B. et al. 2014. HD iPSC-derived neural progenitors accumulate in culture and are susceptible to BDNF withdrawal due to glutamate toxicity. *Human Molecular Genetics* . doi: 10.1093/hmg/ddv080.

McAllister, B. et al. 2022. Exome sequencing of individuals with Huntington’s disease implicates FAN1 nuclease activity in slowing CAG expansion and disease onset. *Nature Neuroscience* 25(4), pp. 446–457. Available at: https://www.nature.com/articles/s41593-022-01033-5 [Accessed: 19 April 2022].

Mcallister, W.B. 2019. *Identification and characterisation of genetic variation that modifies age at onset in Huntington’s disease*. Available at: https://orca.cardiff.ac.uk/130027/1/2020McAllisterWB PhD.pdf [Accessed: 9 December 2021].

Migliore, S. et al. 2019. Genetic Counseling in Huntington’s Disease: Potential New Challenges on Horizon? *Frontiers in Neurology* 10, p. 453. doi: 10.3389/fneur.2019.00453.

Miller, D.E. et al. 2020. Targeted long-read sequencing resolves complex structural variants and identifies missing disease-causing variants. *bioRxiv* , p. 2020.11.03.365395. Available at: https://www.biorxiv.org/content/10.1101/2020.11.03.365395v1 [Accessed: 2 February 2022].

Milne, I. et al. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* 14(2), pp. 193–202. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs012 [Accessed: 15 February 2022].

Mirkin, S.M. and Frank-Kamenetskii, M.D. 1994. H-DNA and related structures. *Annual Review of Biophysics and Biomolecular Structure* . doi:

10.1146/annurev.bb.23.060194.002545.

Mitsuhashi, S. et al. 2019. Tandem-genotypes : robust detection of tandem repeat expansions from long DNA reads. *Genome Biology 2019 20:1* 20(1), p. 58. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1667-6 [Accessed: 18 July 2019].

Mochel, F. et al. 2007. Early Energy Deficit in Huntington Disease: Identification of a Plasma Biomarker Traceable during Disease Progression. Mueller, U. ed. *PLOS ONE* 2(7), p. e647. doi: 10.1371/journal.pone.0000647.

Moss, D.J.H. et al. 2017. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurology* 16(9), pp. 701–711. doi: 10.1016/S1474-4422(17)30161-8.

Mousavi, N. et al. 2018. Profiling the genome-wide landscape of tandem repeat expansions. *bioRxiv* . doi: 10.1101/361162.

Murray, V. et al. 1993. The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. *Nucleic acids research* 21(10), pp. 2395–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8506134 [Accessed: 12 April 2022].

Myers, R.H. et al. 1991. Factors Associated With Slow Progression in Huntington's Disease. *Archives of Neurology* 48(8), pp. 800–804. Available at: http://archneur.jamanetwork.com/article.aspx?articleid=591026 [Accessed: 8 January 2022].

Nasir, J. et al. 1995. Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. *Cell* 81(5), pp. 811–823. doi: 10.1016/0092-8674(95)90542-1.

Nattestad, M. et al. 2018. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome research* 28(8), pp. 1126–1135. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29954844 [Accessed: 13 February 2022].

NEB [no date]. Polymerase Fidelity: What is it, and what does it mean for your PCR? | NEB. Available at: https://www.neb.com/tools-and-resources/feature-articles/polymerase-fidelity-what-is-it-and-what-does-it-mean-for-your-pcr

[Accessed: 26 August 2020].

Nesic, K. et al. 2018. Targeting DNA repair: the genome as a potential biomarker. *The Journal of pathology* 244(5), pp. 586–597. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29282716 [Accessed: 13 February 2022].

Oh, S. et al. 2016. *Minimization of Chimera Formation and Substitution Errors in Full-Length 16S PCR Amplification*. Available at: www.pacb.com/support/documentation [Accessed: 8 December 2021].

Ooi, J. et al. 2019. Unbiased Profiling of Isogenic Huntington Disease hPSC-Derived CNS and Peripheral Cells Reveals Strong Cell-Type Specificity of CAG Length Effects. *Cell Reports* 26(9), pp. 2494-2508.e7. doi: 10.1016/j.celrep.2019.02.008.

Orr, H.T. and Zoghbi, H.Y. 2007. Trinucleotide Repeat Disorders. *Annual Review of Neuroscience* 30(1), pp. 575–621. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17417937 [Accessed: 14 August 2018].

Owen, B.A.L. et al. 2005. (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nature Structural & Molecular Biology* 12(8), pp. 663–70. doi: 10.1038/nsmb965.

Pacific Biosciences, P. 2018. *Procedure &Checklist - Amplicon Template Preparation and Sequencing*. Available at: https://github.com/AntWarland/doctoral_thesis/blob/main/PacBio/Procedure-Checklist-Amplicons.pdf [Accessed: 8 April 2022].

Pacific Biosciences, P. 2019. *Procedure & Checklist - Preparing SMRTbell® Libraries using PacBio® Barcoded Universal Primers for Multiplexing Amplicons*. Available at: https://github.com/AntWarland/doctoral_thesis/blob/main/PacBio/Procedure-Checklist-Preparing-SMRTbell-Libraries.pdf [Accessed: 8 April 2022].

Panegyres, P.K. et al. 2006. A study of potential interactive genetic factors in Huntington's disease. *European Neurology* . doi: 10.1159/000093867.

Paulsen, J.S. and Conybeare, R.A. 2005. Cognitive changes in Huntington's disease. *Advances in neurology* 96, pp. 209–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16385769 [Accessed: 8 January 2022].

Payne, A. et al. 2021. Readfish enables targeted nanopore sequencing of gigabase-

sized genomes. *Nature Biotechnology* 39(4), pp. 442–450. Available at: http://www.nature.com/articles/s41587-020-00746-x [Accessed: 14 February 2022].

Pearson, C.E. et al. 1998. Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry* 37(8), pp. 2701–8. doi: 10.1021/bi972546c.

Pollard, L.M. et al. 2004. Replication-mediated instability of the GAA triplet repeat mutation in Friedreich ataxia. *Nucleic acids research* 32(19), pp. 5962–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15534367 [Accessed: 1 February 2022].

Pollard, M.O. et al. 2018. Long reads: their purpose and place. *Human Molecular Genetics* 27(R2), pp. R234–R241. Available at: https://academic.oup.com/hmg/article/27/R2/R234/4996216 [Accessed: 14 August 2018].

Porro, A. et al. 2021. FAN1-MLH1 interaction affects repair of DNA interstrand cross-links and slipped-CAG/CTG repeats. *Science advances* 7(31), p. eabf7906. doi: 10.1126/sciadv.abf7906.

Pringsheim, T. et al. 2012. The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Movement Disorders* 27(9), pp. 1083–1091. Available at: https://onlinelibrary.wiley.com/doi/10.1002/mds.25075 [Accessed: 9 January 2022].

Quarrell, O. et al. 2012. The Prevalence of Juvenile Huntington's Disease: A Review of the Literature and Meta-Analysis. *PLoS currents* 4, p. e4f8606b742ef3. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22953238 [Accessed: 7 January 2022].

Ranen, N.G. et al. 1995. Anticipation and instability of IT-15 (CAG)n repeats in parent-offspring pairs with Huntington disease. *American Journal of Human Genetics* 57(3), pp. 593–602.

Robinson, J.T. et al. 2011. Integrative genomics viewer. *Nature Biotechnology* 29(1), pp. 24–26. Available at: http://www.nature.com/articles/nbt.1754 [Accessed: 7 December 2021].

De Roeck, A. et al. 2019. NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biology* 20(1), p. 239. Available at:

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1856-3 [Accessed: 1 February 2022].

Rolfsmeier, M.L. and Lahue, R.S. 2000. Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. 20(1), pp. 173–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10594019 [Accessed: 11 February 2022].

Rowe, R.G. and Daley, G.Q. 2019. Induced pluripotent stem cells in disease modelling and drug discovery. *Nature Reviews Genetics* 20(7), pp. 377–388. Available at: http://www.nature.com/articles/s41576-019-0100-z [Accessed: 21 December 2021].

Rubinsztein, D.C. 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(1), pp. 16–22.

Ruocco, H.H. et al. 2008. Longitudinal analysis of regional grey matter loss in Huntington disease: effects of the length of the expanded CAG repeat. *Journal of Neurology, Neurosurgery, and Psychiatry* 79(2), pp. 130–5. doi: 10.1136/jnnp.2007.116244.

Russ, J. et al. 2015. Hypermethylation of repeat expanded C9orf72 is a clinical and molecular disease modifier. *Acta neuropathologica* 129(1), pp. 39–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25388784 [Accessed: 13 February 2022].

Sanger, F. et al. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74(12). Available at: https://pubmed.ncbi.nlm.nih.gov/271968/ [Accessed: 10 August 2020].

Sathasivam, K. et al. 1999. Formation of polyglutamine inclusions in non-CNS tissue. *Human Molecular Genetics* . doi: 10.1093/hmg/8.5.813.

Saudou, F. and Humbert, S. 2016. The Biology of Huntingtin. *Neuron* 89(5), pp. 910–926. doi: 10.1016/j.neuron.2016.02.003.

Scahill, R.I. et al. 2020. Biological and clinical characteristics of gene carriers far from predicted onset in the Huntington's disease Young Adult Study (HD-YAS): a cross-sectional analysis. *The Lancet. Neurology* 19(6), pp. 502–512. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/32470422 [Accessed: 19 January 2022].

Schirmer, M. et al. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 17(1), p. 125. Available at: http://www.biomedcentral.com/1471-2105/17/125 [Accessed: 24 September 2020].

Semaka, A. and Hayden, M.R. 2014. Evidence-based genetic counselling implications for Huntington disease intermediate allele predictive test results. *Clinical Genetics* 85(4), pp. 303–311. doi: 10.1111/cge.12324.

Shelbourne, P.F. et al. 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.

Shen, W. et al. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. Zou, Q. ed. *PLOS ONE* 11(10), p. e0163962. Available at: https://dx.plos.org/10.1371/journal.pone.0163962 [Accessed: 11 April 2022].

Shendure, J. and Ji, H. 2008. Next-generation DNA sequencing. *Nature Biotechnology* 26(10), pp. 1135–1145. Available at: http://www.nature.com/articles/nbt1486 [Accessed: 26 August 2020].

Skodda, S. et al. 2014. Impaired motor speech performance in Huntington's disease. *Journal of neural transmission (Vienna, Austria : 1996)* 121(4), pp. 399–407. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24221215 [Accessed: 8 January 2022].

Smith-Geater, C. et al. 2020. Aberrant Development Corrected in Adult-Onset Huntington's Disease iPSC-Derived Neuronal Cultures via WNT Signaling Modulation. *Stem cell reports* 14(3), pp. 406–419. doi: 10.1016/j.stemcr.2020.01.015.

Soldner, F. et al. 2009. Parkinson's Disease Patient-Derived Induced Pluripotent Stem Cells Free of Viral Reprogramming Factors. *Cell* . doi: 10.1016/j.cell.2009.02.013.

Spada, A.R. La et al. 1991. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* 352(6330), pp. 77–79. Available at: http://www.nature.com/articles/352077a0 [Accessed: 7 January 2022].

Squitieri, F. et al. 2003. Homozygosity for CAG mutation in Huntington disease is associated with a more severe clinical course. *Brain* 126(Pt 4), pp. 946–55.

Stevens, J.R. et al. 2013. Trinucleotide repeat expansions catalyzed by human cell-free extracts. *Cell research* 23(4), pp. 565–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23337586 [Accessed: 9 January 2022].

Su, Y. et al. 2021. Deciphering Neurodegenerative Diseases Using Long-Read Sequencing. *Neurology* 97(9), pp. 423–433. Available at: http://www.ncbi.nlm.nih.gov/pubmed/34389649 [Accessed: 1 February 2022].

Suelves, N. et al. 2017. A selective inhibitor of histone deacetylase 3 prevents cognitive deficits and suppresses striatal CAG repeat expansions in Huntington's disease mice. *Scientific Reports* 7(1), p. 6082. Available at: http://www.nature.com/articles/s41598-017-05125-2 [Accessed: 7 April 2022].

Svrzikapa, N. et al. 2020. Investigational Assay for Haplotype Phasing of the Huntingtin Gene. *Molecular therapy. Methods & clinical development* 19, pp. 162–173. Available at: http://www.ncbi.nlm.nih.gov/pubmed/33209959 [Accessed: 25 January 2022].

Swami, M. et al. 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* . doi: 10.1093/hmg/ddp242.

Tabrizi, S.J. et al. 2009. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *The Lancet Neurology* 8(9), pp. 791–801. doi: 10.1016/S1474-4422(09)70170-X.

Tabrizi, S.J. et al. 2013. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *The Lancet Neurology* 12(7), pp. 637–49. doi: 10.1016/S1474-4422(13)70088-7.

Takahashi, K. et al. 2007. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* . doi: 10.1016/j.cell.2007.11.019.

Takahashi, K. and Yamanaka, S. 2006. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* 126(4), pp. 663–676. Available at: https://www.sciencedirect.com/science/article/pii/S0092867406009767?via%3Dihub [Accessed: 21 December 2021].

Tange, O. 2018. GNU Parallel 2018. Available at: https://zenodo.org/record/1146014#.YlQ0Qy8w3UY [Accessed: 11 April 2022].

Telenius, H. et al. 1994. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nature Genetics* . doi: 10.1038/ng0494-409.

Telezhkin, V. et al. 2016. Forced cell cycle exit and modulation of GABAA, CREB, and GSK3β signaling promote functional maturation of induced pluripotent stem cell-derived neurons. *American journal of physiology. Cell physiology* 310(7), pp. C520-41. doi: 10.1152/ajpcell.00166.2015.

The HD iPSC Consortium et al. 2012. Induced Pluripotent Stem Cells from Patients with Huntington's Disease Show CAG-Repeat-Expansion-Associated Phenotypes. *Cell Stem Cell* 11(2), pp. 264–278. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22748968 [Accessed: 21 December 2021].

The HD iPSC Consortium 2017. Developmental alterations in Huntington's disease neural cells and pharmacological rescue in cells and mice. *Nature Neuroscience* . doi: 10.1038/nn.4532.

Tsai, Y.-C. et al. 2017. Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions. *bioRxiv* , p. 203919. Available at: https://www.biorxiv.org/content/10.1101/203919v1 [Accessed: 31 January 2022].

Turner, D.J. 2011. Next-generation DNA Sequencing Technologies. In: *Encyclopedia of Analytical Chemistry*. Chichester, UK: John Wiley & Sons, Ltd. Available at: http://doi.wiley.com/10.1002/9780470027318.a9209.

Veitch, N.J. et al. 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. Available at: https://www.sciencedirect.com/science/article/pii/S1568786407000201#fig5 [Accessed: 8 February 2022].

Vnencak-Jones, C.L. 2003. Fluorescence PCR and GeneScan® Analysis for the Detection of CAG Repeat Expansions Associated with Huntington's Disease. In: *Neurogenetics*. New Jersey: Humana Press, pp. 101–108. Available at: http://link.springer.com/10.1385/1-59259-

330-5:101 [Accessed: 2 January 2022].

Vonsattel, J.P. and DiFiglia, M. 1998. Huntington disease. *Journal of Neuropathology and Experimental Neurology* 57(5), pp. 369–84.

Vonsattel, J.P.G. et al. 2008. Neuropathology of Huntington's disease. *Handbook of clinical neurology* 89, pp. 599–618. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18631782 [Accessed: 14 February 2022].

Waldvogel, H.J. et al. 2015. The Neuropathology of Huntington's Disease. *Current Topics in Behavioral Neurosciences* 22, pp. 33–80. doi: 10.1007/7854_2014_354.

Warner, J.P. et al. 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* . doi: 10.1006/mcpr.1993.1034.

Warner, J.P. et al. 1996. A general method for the detection of large CAG repeat expansions by fluorescent PCR. *Journal of medical genetics* 33(12), pp. 1022–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9004136 [Accessed: 24 January 2022].

Weirather, J.L. et al. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, p. 100. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28868132 [Accessed: 30 November 2021].

Wenger, A.M. et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* 37(10), pp. 1155–1162. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31406327 [Accessed: 7 December 2021].

Wenger, A.M. et al. 2020. Copy-Number Variant Detection with PacBio Long Reads Long Highly Accurate HiFi Reads. Available at: https://github.com/PacificBiosciences/pbsv [Accessed: 7 December 2021].

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* 101(10), pp. 3498–3503. Available at: https://pnas.org/doi/full/10.1073/pnas.0308679101.

White, J.K. et al. 1997. Huntingtin is required for neurogenesis and is not impaired by

the Huntington's disease CAG expansion. *Nature genetics* 17(4), pp. 404–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9398841 [Accessed: 10 January 2022].

Wick, R.R. et al. 2018. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS computational biology* 14(11), p. e1006583. Available at: http://www.ncbi.nlm.nih.gov/pubmed/30458005 [Accessed: 24 September 2020].

Wieben, E.D. et al. 2019. Amplification-free long-read sequencing of TCF4 expanded trinucleotide repeats in Fuchs Endothelial Corneal Dystrophy. *PloS one* 14(7), p. e0219446. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31276570 [Accessed: 2 January 2022].

Woerner, A.C. et al. 2016. Cytoplasmic protein aggregates interfere with nucleocytoplasmic transport of protein and RNA. *Science* 351(6269), pp. 173–6. doi: 10.1126/science.aad2033.

Wright, G.E.B. et al. 2019. Length of Uninterrupted CAG, Independent of Polyglutamine Size, Results in Increased Somatic Instability, Hastening Onset of Huntington Disease. *The American Journal of Human Genetics* 104(6), pp. 1116–1126. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0002929719301533 [Accessed: 19 May 2020].

Wright, G.E.B. et al. 2020. Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. *The Lancet Neurology* 19(11), pp. 930–939. Available at: https://www.sciencedirect.com/science/article/pii/S1474442220303434?via%3Dihub#bib6 [Accessed: 26 January 2022].

Xi, Z. et al. 2013. Hypermethylation of the CpG island near the G4C2 repeat in ALS with a C9orf72 expansion. *American journal of human genetics* 92(6), pp. 981–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23731538 [Accessed: 13 February 2022].

Xu, P. et al. 2020. Dynamics of strand slippage in DNA hairpins formed by CAG repeats: roles of sequence parity and trinucleotide interrupts. *Nucleic acids research* 48(5), pp. 2232–2245. Available at: http://www.ncbi.nlm.nih.gov/pubmed/31974547 [Accessed: 23 January 2022].

Xu, X. et al. 2017. Reversal of Phenotypic Abnormalities by CRISPR/Cas9-Mediated Gene Correction in Huntington Disease Patient-Derived Induced Pluripotent Stem Cells. *Stem Cell Reports* 8(3), pp. 619–633. doi: 10.1016/j.stemcr.2017.01.022.

Yoon, S.-R. et al. 2003. Huntington disease expansion mutations in humans can occur before meiosis is completed. *Proceedings of the National Academy of Sciences of the United States of America* 100(15), pp. 8834–8. doi: 10.1073/pnas.1331390100.

Zeitlin, S. et al. 1995. Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. *Nature Genetics* 11(2), pp. 155–163. doi: 10.1038/ng1095-155.

Zhang, N. et al. 2010. Characterization of human Huntington's disease cell model from induced pluripotent stem cells. *PLoS Currents* . doi: 10.1371/currents.RRN1193.

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