KU Leuven

Biomedical Sciences Group

Faculty of Medicine

Department of Human Genetics

DETERMINATION OF THE VARIABILITY AND ASSOCIATED EPIGENETIC SIGNATURE OF TANDEM REPEATS BY SINGLE MOLECULE SEQUENCING

Simon ARDUI

Jury:

Promoter:	Prof. dr. Joris R. Vermeesch
Co-promoter:	Prof. dr. Gert Matthijs
Chair:	Prof dr. Anton Roebroek
Secretary:	Prof. dr. Diether Lambrechts
Jury members:	Prof. dr. Adam Ameur
	Prof. dr. Diether Lambrechts
	Prof. dr. Karen Sermon
	Prof. dr. Kevin J. Verstrepen

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor in Biomedical Sciences

December 2018

Author: Simon Ardui

Cover design: Simon Ardui

Table of Content

Chapter 1: General Introduction

Partly based on:

Ardui, S.[1], Ameur, A.[2], Vermeesch, J. [1], Hestand, M.[3] (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 46, 2159–2168.

[1]Department of Human Genetics, KU Leuven, Leuven 3000, Belgium
[2]Department of Immunology, Genetics and Pathology, Uppsala University, Science for Life Laboratory, Uppsala 75108, Sweden,
[3]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia and 4Department of Clinical Genetics, VU University Medical Center, Amsterdam 1081 BT, The Netherlands

1.1. Repeats in the Human Genome

The human genome is made up of billions of DNA building blocks that encode all the necessary information to create a functional human being. A human body contains around 10^{13} cells which all carry their own copy of the human genome. Although all cells carry the exact same genetic information, they manage to develop in different cell types (muscle, brain, skin,…) with completely different functions since the human genome is a dynamic entity which can selectively display the information needed for each specific cell type.

Our knowledge of the human genome has greatly increased over the past decades and culminated in the human genome project in 2001 (Lander et al., 2001). This project revealed the genetic composition of the human genome which is surprising to what one would expect without a priori knowledge. For example, genes make up only 1.5% of the human genome while repetitive regions cover up to 50% (Jelinek et al., 1980; Lander et al., 2001). The repetitive regions can be categorized into 2 groups: interspersed and tandem repeats. Long terminal repeats (LTR), short interspersed nuclear elements (SINE) and long interspersed nuclear elements (LINE) are examples of the interspersed repeats and are spread across the complete human genome. This is in contrast with tandem repeats which are either large blocks of DNA (e.g. centromeric regions) or short units (e.g. short tandem repeats (STR)) repeated head-to-tail at the same locus (Gregory, 2005).

Currently repetitive regions are understudied. They have often been dismissed from genomic analyses due to their status as uninteresting junk DNA and due to technical limitations which make them difficult (sometimes even impossible) to interpret (Doolittle and Sapienza, 1980; Ohno, 1972). It is now becoming apparent that repeats have crucial roles in many biological processes. Therefore, this thesis focuses on STRs and contributes to the intensive and state of the art research which is currently performed on the repeatome.

1.2. Short Tandem Repeats

1.2.1. Introduction

STRs are also called microsatellites or simple sequence repeats (SSRs) and are made up of short units of 2-6 repeat units that are repeated head to tail (e.g. CGG → CGGCGGCGGCGG). On average, a STR is encountered every 2kb in the human genome making these elements extremely common (Lander et al., 2001). They are not only found in intergenic regions, but also in introns, 3' or 5' untranslated regions (UTR), promotors, enhancers and even in exons (Ellegren, 2004; Sawaya et al., 2013). Tandem repeats with a repeat unit larger than 6 bases also exist and are called minisatellites (Jeffreys et al., 1985).

STRs behave differently compared to other genetic variants, such as single nucleotide polymorphism (SNPs). Not only do STRs have a higher magnitude of instability, they also have a different mode of instability. Where the mutation profile of SNPs is binary (e.g. C →T), the number of STR building blocks can vary by adding or deleting one or more repeat units (e.g. $(CGG)_{10}$ → $(CGG)_{15}$). This reflects a digital mutation profile of STRs whereby many possibilities can be created. Instability of STRs is therefore also referred to as dynamic mutations (Nithianantharajah and Hannan, 2007). Duitama et al. (2014) showed that STR are 7.4X more polymorphic than SNPS (9% versus 1.25%). The degree of instability depends on different factors like the length of the motif, the total length of the STR, the presence of interruptions (e.g. CGGCGG**AGG**CGG), epigenetic modifications and the parent (paternal versus maternal) from which the STR is inherited (Brinkmann et al., 1998; Legendre et al., 2007). Instability can occur meiotically, mitotically and also postmitotically (Gonitel et al., 2008; Martorell et al., 1997; Rifé et al., 2004). Due to this instability, usually there is a range of tandem repeat lengths exhibited in the population for each tandem repeat. STRs are also (post)mitotically unstable and consequently somatic mosaicism is also common (Jiraanont et al., 2017; Martorell et al., 1997; Pretto et al., 2014a). However, due to technical and experimental limitations this type of mosaicism is usually overlooked. Therefore, its possible biological and clinical impact is unclear.

1.2.2. Understanding the Function of STRs

<u>Tandem Repeats and Disease</u>

STRs are useful molecular markers in human genetics studies due to their high degree of polymorphism. However, nowadays it is clear that they also play crucial roles in many biological processes. This is exemplified by the tremendous impact mutated STRs have on the phenotype of an individual by causing various cancers (such as Lynch syndrome and prostate cancer) and more than 40 repeat expansion disorders (López Castel et al., 2010; Pearson et al., 2005b; Tsujimoto et al., 2004). These repeat expansion disorders are primarily neuromuscular and neurodegenerative disorders caused by a significant expansion of a trinucleotide repeat which can be located in both coding and non-coding regions (Fondon et al., 2008; Usdin, 2008). Some representative examples of this group are shown in Table 1 (for a complete list see Castel et al., 2010).

Although each repeat expansion disorder is characterized by its own specifics, they all behave in a similar way. Instead of one normal allele, a whole range of normal STRs with a low number of repeat units occur in healthy individuals. Within the normal range STRs are very stable, both meiotically and mitotically. However, sometimes a small increase of a few units might occur. If multiples of these small expansions occur over different generations, at a certain moment the STR passes a critical size threshold and arrives in the premutation zone. In this zone STRs usually do not cause disease yet but become more unstable. Large expansions with tens to hundreds of units can now occur in a single generation. These expansions will become pathogenic above a specific length (Lee and McMurray, 2014). Interestingly, the minimal expansion to be pathogenic is smaller for coding STRs compared to non-coding STRs. This is because coding expansions lead to dysfunctional proteins, while there is less selection on non-coding STRs (Table 1).

The phenotype of the diseases differs significantly with varying repeat sizes within the disease range. Larger STR expansions will induce more and worse symptoms, a higher penetrance and an earlier manifestation of the disorder (Bagni and Oostra, 2013). Typically, the severity of repeat expansion disorders phenotype gradually worsens in further generations since the repeat size continues to grow within the pedigree. This is known as genetic anticipation or the Sherman Paradox (Fu et al., 1991; Pratte et al., 2015).

How Do STRs Cause Disease?

Mutated STRs can disturb biological processes on the DNA, RNA and protein level (Hamada et al., 1984). On the DNA level, genes can be silenced by long STRs located in the 5'UTR (e.g. FXS & FRAXE) or intronic regions (e.g. FRDA)(Figure 1). Here, the presence of these expanded STRs inhibits transcription or transcriptional elongation of the gene. In FXS, the expanded CGG repeat also triggers methylation of the 5'UTR which contributes to the complete silencing of *FMR1*.

Once expanded STRs are transcribed, they can boycott the regular biological processes on RNA level through a variety of mechanisms. Firstly, the presence of expanded CUG (DM1), CCUG (DM2) or CCG repeats (Fragile X-associated tremor/ataxia syndrome (FXTAS)) can add a toxic RNA gain-of-function to the RNA. The transcripts are retained in foci within the cell nucleus where they are able to sequester RNA binding proteins. In DM1 and FXTAS CUG triplet repeat RNA binding protein 1 and musclebind-like protein are sequestered by the expanded repeats (Orr and Zoghbi, 2007; Sofola et al., 2007). Since both proteins are involved in alternative splicing, their misregulation will cause the aberrant splicing of different genes. Secondly, some STRs might induce a more open chromatin conformation which will boost transcription. This is seen in FXTAS where a premutated CGG allele induces an overexpression of *FMR1* which will further magnify the toxic RNA gain-of-function (Tassone et al., 2007). Thirdly, also repeat associated non-ATG (RAN) translation occurs in different repeat expansion disorders such as HT, DM1 and FXTAS (Green et al., 2016). RNA molecules can normally only be translated when an AUG codon is present. However, if the RNA molecule contains an expanded repeat (e.g. CAG), these transcripts can be translated even in the absence of a start codon. RAN-translation can start from all 3 possible reading frames (e.g. CAG, AGC or GCA) and each of these translations can produce potentially a toxic homopolymeric protein contributing to neuronal toxicity (Green et al., 2016).

Since transcription often occurs bidirectional, one expanded repeat can generate 2 toxic RNA molecules which can produce together up to 7 toxic proteins (3 RAN translated from each transcript + 1 from the AUG start codon)(Pearson, 2011).

Table 1: Characteristics of tandem repeat expansion disorders.

			Range		
Disease	Repeat Unit	Gene Name	Normal	Premutation	Disease
Disorders caused by coding STRs					
DRPLA	CAG	*ATN1*	6-35	35-48	49-88
HD	CAG	*HTT*	6-29	29-37	38-180
OPMD	GCN	*PAPBN1*	10	/	11-17
SCA1	CAG	*ATXN1*	6-39	40	41-83
SCA2	CAG	*ATXN2*	31	31-32	32-200
SCA3	CAG	*ATXN3*	12-40	41-60	60-87
SCA6	CAG	*CACNA1A*	<18	19	20-33
SCA7	CAG	*ATXN7*	4-17	28-33	>36
SCA17	CAG	*TBP*	25-42	43-45	45-66
SMBA	CAG	*AR*	13-31	32-39	40
Disorders caused by non-coding STRs					
DM1	CTG	*DMPK*	5-37	37-50	>50
DM2	CCTG	*CNBP*	<30	31-74	75-11.000
EPM1	C_4GC_4GCG	*CSTB*	2-3	4-29	>29
FRAX-E	GCC	*AFF2*	4-39	40-200	>200
FRDA	GAA	*FXN*	5-30	31-66	67-1500
FXS	CGG	*FMR1*	6-54	55-200	>200
HDL2	CTG	*JPH3*	6-27	29-35	36-57
SCA8	CTG	*ATXN8OS*	15-34	35-88	89-250
SCA10	ATTCT	*ATXN10*	10-29	30-399	400-4500
SCA12	CAG	*PPP2R2B*	7-28	29-66	67-78
FTD/ALS	GGGGCC	*C9ORF72*	2-25	/	>25

DRPLA, dentatorubral-pallidoluysian atrophy; *ATN1*, atrophin 1; HD, huntington's disease: *HTT*; huntingtin; OPMD, oculopharyngeal muscular dystrophy; *PAPBN1*, poly(A) binding protein nuclear 1;SCA, spincerebellar ataxia; *ATXN*, ataxin; *CACNA1A*, calcium voltage-gated channel subunit alpha1 A; *TBP*, TATA-box binding protein, SMBA, Spinal and bulbar muscular atrophy; *AR*, androgen receptor; DM, myotonic dystrophy; DMPK, Dystrophic Myotonic Protein Kinase; CNBP, CCHC-type zinc finger nucleic acid binding protein; FRAX-E, fragile XE syndrome; *AFF2*, AF4/FMR2 family member 2; FRDA, friedreich's ataxia; *FXN*, frataxin; FXS, fragile X syndrome; *FMR1*, fragile-X mental retardation 1; HDL2, Huntington disease-like 2, *JPH3*, junctophilin 3; *ATXN8OS*, ataxin 8 opposite strand, *PPP2R2B*, protein phosphatase 2 regulatory subunit B beta, C9FTD/ALS, Frontotemporal Dementia and Amyotrophic Lateral Sclerosis. EPM1, progressive myoclonic epilepsy 1 (based on Lopez Castel et al., 2010).

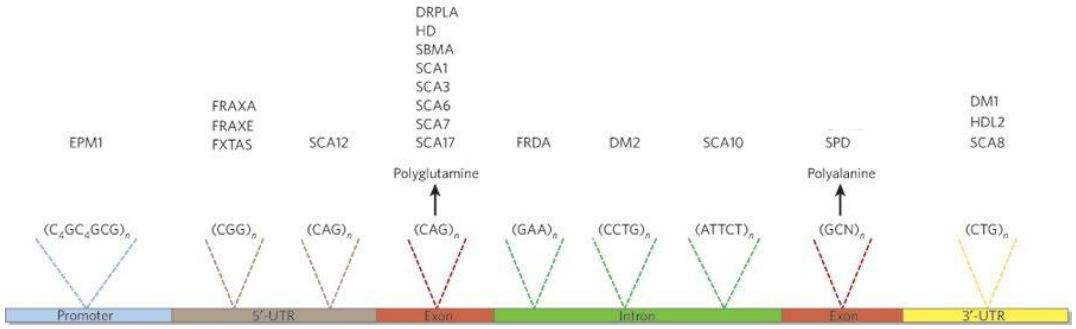


Figure 1: Disease causing STRs are found in promotors, 3' and 5' untranslated regions (UTR), promotors, introns and even in exons (adapted from Mirkin, 2007).

Some STRs are embedded within coding regions and hence are also translated into protein (Figure 1). Up to now CAG and GCN repeats are identified which are translated into polyglutamine (PolyQ) or polyalanine (PolyA) tracks respectively (Table 1). The presence of these long tracks can cause a reduction of the normal functioning of the protein and can add a toxic gain-of-function. This toxic gain-of-function is dependent on the hosting protein, but often involves the sequestration of glutamine-rich transcription factors like CREB-binding protein and specificity protein 1, thereby interfering with transcription (Orr and Zoghbi, 2007; Usdin, 2008).

Expanded STRs can modulate biological processes in different ways. Although for some repeat expansion disorder the pathogenesis is largely understood (e.g. FXS & DM1) this remains opaque for other diseases (e.g. SCA10, SCA12 & HDL2). Therefore, the described mechanisms will most probably be complemented with additional mechanisms in the future.

The function of tandem repeats beyond disease

Bearing in mind the detrimental effects of expanded STRs, it is surprising that evolution did not eliminate these harmful DNA elements. One hypothesis is that natural selection is simply not powerful enough to eliminate the highly polymorphic STRs from the genome. On the other hand, a functional role of STRs in bacteria and yeast has already been demonstrated (Stern et al., 1986; Tirosh et al., 2009; Verstrepen et al., 2005; Weiser et al., 1989), and hence it is possible that STRs also offer an added value for the functioning of humans.

Evidence is accumulating that STR variation shapes the phenotype of healthy individuals. As stated already above, hundreds of thousands of STRs are widespread across the human genome and occur in introns, UTRs, enhancers and in the coding regions of up to 20% of genes (Ellegren, 2004; Gemayel et al., 2010; Sawaya et al., 2013). Hence, they can influence a wide range of biological processes by affecting gene expression, RNA and protein function, which hints towards a functional impact of STRs (Gemayel et al., 2010; Vinces et al., 2009). Furthermore, STRs are especially enriched in genes important for neurological and developmental functions (Legendre et al., 2007; Nithianantharajah and Hannan, 2007; Riley and Krieger, 2009). In both drosophila and dogs it has been shown that a quantitative correlation exists between STR variation and head shape (Birge et al., 2010; Fondon and Garner, 2004). In humans, the normal variation of the *HTT* CAG has been shown to correlate with intelligence (Lee et al., 2018).

The high degree of STR polymorphisms not only allows fine-tuning of gene expression, but also provides the human genome with a rich source of variability upon which natural selection may act (Frenkel and Trifonov, 2012; King et al., 1997). Variation in STRs may arise randomly but will subsequently serve as a template whereupon natural selection can act. In this way, STRs can serve as tuning knobs in evolutionary processes (Kashi and King, 2006).

1.2.3. Mechanisms Involved in Tandem Repeat Instability

STR instability arises during meiosis, mitosis and even in non-dividing cells (De Temmerman et al., 2004; Gonitel et al., 2008; Pretto et al., 2014b). This indicates that instability is not caused by one simple mechanisms, but rather by several pathways and mechanisms (Pearson et al., 2005b). Here, we discuss how different mechanisms, including recombination, replication and the DNA repair machinery, are empowered to induce STR instability.

- Recombination

Recombination is known to repair double-stranded breaks and to introduce variation during meiosis. It is thus not surprising that recombination can also introduce STR variation by the unequal crossing over between 2 sister chromatids, homologous chromosomes or even within the same chromosome (Figure 2)(Paques et al., 1998; Richard and Pâques, 2000; Sia et al., 1997; Smith, 1976). Furthermore, recombination can also occur within the same STR leading to contractions (Gemayel et al., 2010).

Expansions of the GCN repeat in polyA disorders are often caused by recombination. Since the third letter of the codon is an N, this codon will always encode Alanine, regardless of the DNA letter at the third position. Hence, there is a lot of variation at this position since there is no selective pressure. By tracing variation in this letter, recombination events can be picked up (Albrecht and Mundlos, 2005). For other tandem repeat disorders, the role of recombination is not clear. If recombination occurs within a STR, this could induce variability without any change in the flanking regions. Hence, variability introduced by recombination might be missed (Mirkin, 2007). On the other hand it is also hypothesized that the role of recombination in the creation of STR variation is only limited, but rather becomes important for larger repeat sizes like minisatellites (Richard and Pâques, 2000).

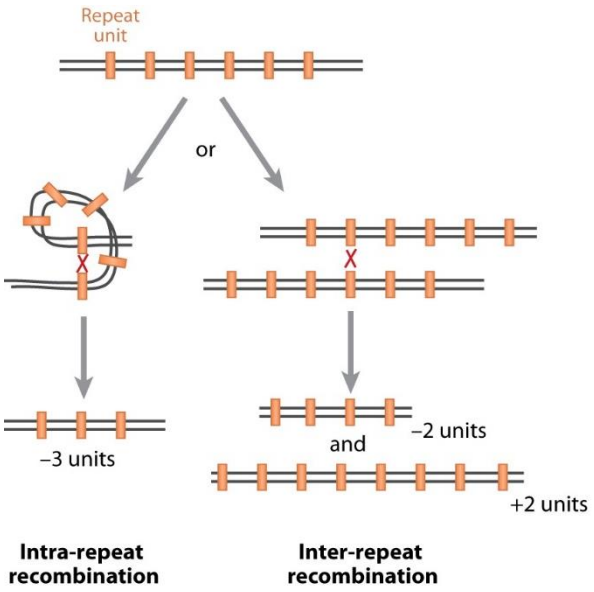


Figure 2: Unequal crossing over can introduce STR variation (adapted from Gemayel et al, 2010).

- Replication

Replication plays an important role in tandem repeat instability (Tachida and Iizuka, 1992). During replication duplex DNA is temporally unwound due to the progression of the replication fork. When this happens at a locus containing a STR, the ssDNA will have the tendency to form stable hairpins. After a second round of replication, this leads to contractions and expansions depending on the location of the hairpin which might be either the leading or the lagging strand (Figure 3a). One hypothesis explaining expansions suggests that the polymerase stalls within the STR (Figure 3b). This can be followed by double stranded breaks which will induce instability upon repair (Pearson et al., 2005a). Alternatively, the replication fork can also back-up forming a four-way junction which resembles a chicken foot (Figure 3b, lower pathway). The leading strand can now use the newly formed daughter strand as a template in order to synthesize enough DNA to pass the STR. Subsequently, the replication fork is flipped back whereafter replication can continue. This will lead to the expansion of the STR if a stable hairpin is formed during the fork reversal in the daughter of the leading strand. Additionally, hairpins can also form before the replication complex arrives. If this happens on the lagging strand, one or more Okazaki fragments might be skipped before synthesis of the lagging strand restarts. Since amplification of the leading strand might still proceed, an imbalance between the leading and lagging strand is created. If afterwards the gap on the lagging strand containing the STR is skipped during the repair of this gap, a contraction is formed (Figure 3b; upper pathway)(Mirkin, 2007; Voineagu et al., 2009).

The variability introduced by replication can explain some of the typical characteristics of STR instability. For example, the expansions and contractions formed during replication will be rather small since they arise from the formation of stable hairpin-loops. Interestingly, this explains why in both DM1 and FXS the normal and premutation repeats are more unstable when they are inherited paternally rather than maternally (Nolin et al., 2015; Pratte et al., 2015; Sullivan et al., 2002). Since sperm cells undergo significantly more divisions compared to oocytes (hundreds versus dozens), the chance of variability is also higher in sperm cells than in oocytes (Nolin et al., 1999). In addition, in FXS the *FMR1* CGG repeat is transmitted more stable when AGG units interrupting the CGG repeat are present. These AGG units destabilize the hairpins formed during replication and thus reduce the amount of instability (Nolin et al., 2015).

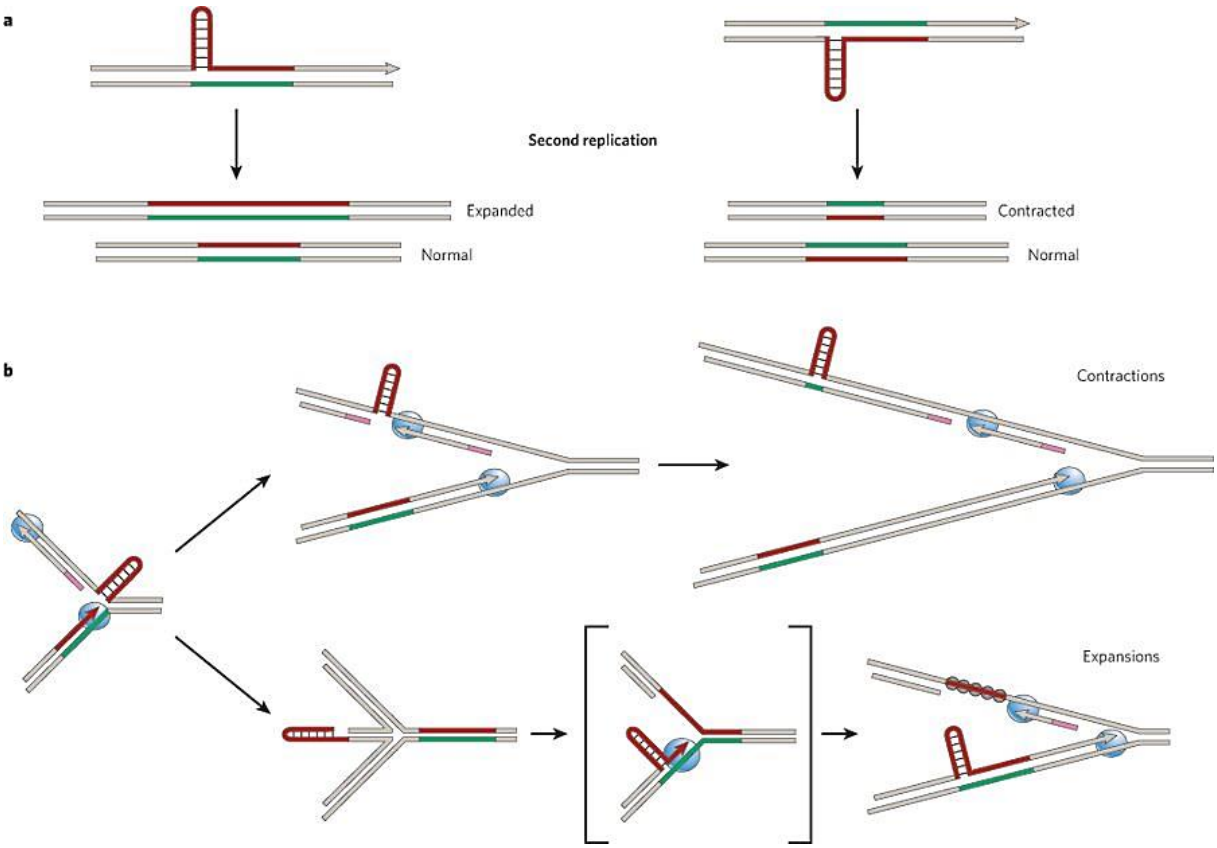


Figure 3: Mechanisms by which STR instability is induced by replication. (A) Stable hairpin structures are formed during replication which lead to contractions and expansions depending on the location of the hairpin. (B) Hypothetical mechanisms through which contractions and expansions might arise (adapted from Mirkin, 2007).

DNA Repair

Although the main task of the DNA repair machinery is to safeguard genomic integrity, it can also fuel STR instability. As previously mentioned, ssDNA containing STRs has the tendency to form stable hairpin structures. These secondary structures arise during replication, but also during other molecular processes like DNA repair. This is exemplified by the base excision repair (BER) pathway and the transcription-coupled repair (TCR) pathway (Lin and Wilson, 2007; Liu and Wilson, 2012).

BER is responsible for removal of damaged bases from DNA and starts with a 7,8-dihydro-8-oxoguanine DNA glycosylase (OGG1) which removes damaged guanine bases from a DNA strand. In addition, OGG1 will also create a single stranded cut in the double stranded DNA by nicking the phosphodiester backbone next to the damaged base. Afterwards, polymerase β can start from the 3' end of the nick and repair the gap. Since this polymerase has strand displacement activity, it will create a stretch of ssDNA containing the STR which will loop back and form a hairpin loop (Asagoshi et al., 2010). Normally this dissociated stretch is removed by flap endonuclease 1. However, this enzyme is unable to remove hairpins since the 5' end is hidden within the hairpin (Spiro et al., 1999).

Hence, the hairpin can be incorporated leading to an expansion. TCR repairs lesions within genes which are actively transcribed. The complex consists out of xeroderma pigmentosum complementation group F and excision cross complementing repair 1 which will make a cut at the 5' end of the lesion. Normally a second cut is made 20 nucleotides downstream by xeroderma pigmentosum complementation group G, but if the DNA contains a STR this can inhibit the excision by formation of hairpins (Staresincic et al., 2009).

The question remains why the DNA repair pathway is not able to remove these hairpins after their genesis. The mismatch repair (MMR) system, another component of the DNA repair system, is thought to be responsible for this failure (Schmidt and Pearson, 2016a). The MutS homologue 2 (MSH2) is part of the MMR system and functions as a genomic guardian with the aim of recognizing the hairpins. After recognition, it will sequester other components of MMR (MSH3, the MutL complex, proliferating cell nuclear antigen,…) to the hairpins (Usdin et al., 2015). Normally this should be followed by hairpin removal whereafter a polymerase would restore the original DNA composition (Lang et al., 2011). Strikingly, MSH2 has an unusual high affinity for STR-containing hairpins, probably due to the presence of mismatches present in the hairpins, and hence cannot detach from the hairpins (Owen et al., 2005). The MMR machinery will then introduce a gap at the opposite strand. Exonuclease 1 will subsequently create a single stranded gap which can be filled in again by polymerase δ. During this restoration of the gap, the hairpin loop can be incorporated causing STR variability (Lang et al., 2011; Schmidt and Pearson, 2016a). Hence, DNA repair not only fails to remove stable STR-hairpins, but it even stimulates the formation of more hairpins, contributes to the stabilization of these hairpins and facilitates their integration in the genome.

Interestingly, DNA repair can explain some of the largest expansions in tandem repeat disorders. For instance, in FXS and DM1 very large expansions with hundreds to thousands of units take place in meiotically arrested oocytes (De Temmerman et al., 2004; Nolin et al., 2003a). Hence, these expansions cannot be caused by replication or recombination. Therefore, the most likely explanation is that these expansions are caused by the DNA repair pathway.

Although an expansion caused by the BER pathway will only change the repeat by a few repeat units, base oxidation occurs up to 50,000 times per cell per day and hence several rounds of the BER pathway within a STR could culminate in a large expansion (Bjelland and Seeberg, 2003; Martin, 2008). Similarly, the progressive growth of STRs in non-dividing cells which occur in some expansion disorders like HD and DM can also be explained by the DNA repair pathway (Kennedy et al., 2003; Morales et al., 2012; Swami et al., 2009).

Since base oxidation occurs continuously and during the entire life of an organism, this provides plenty of opportunities for multiple rounds of oxidation followed by an inadequate repair leading to expansion (Morales et al., 2016).

1.2.4. From Epigenetics to Tandem Repeats and Back

Epigenetics encompass a large group of heritable DNA modifications that alter gene expression without directly changing the primary DNA sequence. This includes different nucleotide modifications, histone modifications and nucleosome positioning from which DNA methylation is the most studied member. This occurs mostly at CpG dinucleotides residing in CpG islands (He and Todd, 2011).

Epigenetic marks are important for packaging the human genome. They also regulate gene expression through which they influence a plethora of biological processes, such as X-inactivation, gene imprinting, differentiation, tumorigenesis and aging (Field et al., 2018; Seisenberger et al., 2013; Verma et al., 2014). It is therefore not surprising that epigenetics is also an important factor in mediating tandem repeat disorders.

Expanded STRs are able to induce epigenetic modifications that influence the clinical outcome of tandem repeat disorders. One of the most prominent examples is the FXS. Here, the presence of large expanded repeats (> 200 CGG units) triggers methylation and histone modifications via an RNA-directed mechanism (Colak et al., 2014). Methylation usually runs from the upstream promotor region over the *FMR1* CGG repeat up until intron 1 (Brasa et al., 2016). The long repeat tract in combination with the presence of the epigenetic modifications completely silence *FMR1* and no fragile X mental retardation protein (FMRP) will be produced. However, if the degree of methylation is not 100%, some *FMR1* mRNA and FMRP will still be present. Interestingly, this can contribute to a more positive phenotype since even the presence of low amounts of FMRP positively impact cognitive function in FXS (Pretto et al., 2014b). In addition to methylation and histone modifications, STRs can also influence gene expression by remodeling the chromatin structure (Kumari and Usdin, 2009; Wang and Griffith, 1995). For instance, in expanded *DMPK* CTG repeats the flanking sequences up- and downstream of the repeat become methylated. Since the repeat resides in the 3'UTR of the gene, this does not influence the transcription of *DMPK* (Nakamori and Thornton, 2010). However, *DMPK* is located within a very gene-dense region and it is located within an insulator element with 2 CTCF binding sites downstream of the repeat. Methylation of these sites inhibits the anchoring of CTCF protein that causes loss of the insulator element and a reduced expression of the downstream gene *SIX5* that probably contributes to the phenotype of DM1 (Barbé et al., 2017; Cho et al., 2005; Yanovsky-Dagan et al., 2015). Interestingly, in DM1 the correlation between repeat size and phenotype is less strong compared to other tandem repeat disorders, suggesting that other factors like epigenetic modifications can potentially have a strong impact on the phenotype of DM1 patients (Evans-Galea et al., 2013a).

Epigenetic modifications in turn also influence the genetics of STRs. It is known that epigenetics can help maintaining the integrity of the human genome, and similarly they can also reduce STR variability (Putiri and Robertson, 2011). This is exemplified by the stabilization of long *FMR1* CGG repeats after methylation in fragile-X patients with a full mutation. Even when fibroblasts from these patients were cultured for multiple passages, no instability could be observed (Wohrle et al., 1993). Probably long CGG repeats are only unstable during the first cell divisions, thereafter they become stabilized by methylation. This explains why often very similar mosaic patterns are observed between different fetal or post-mortem tissues of the same individual (Devys et al., 1992; Ferreira et al., 2013; Wohrle et al., 1993; Wöhrle et al., 1992). Also, in contrast to DM1, progressive instability with increasing repeat lengths during the lifetime of a human is not observed in FXS thanks to presence of methylation (Nakamori and Thornton, 2010).

1.2.5. Therapeutics

No efficient therapy to cure repeat expansion disorders has been developed yet. However, the past decades several basic molecular processes underlying tandem repeat disorders were elucidated, enabling the identification of different targets upon which therapeutics could act. For example, in FXS FMRP is active at neuronal synapses where it functions as an antagonist of the metabotropic glutamate receptor (mGluR). Although promising results in mice were achieved with mGluR antagonist, this could not be replicated during clinical trials in humans (Zeidler et al., 2017). In DM1 the phenotype is mainly caused by another mechanism: a toxic RNA gain-of-function. Therefore, therapeutic approaches are focusing on reducing RNA toxicity, for example by stimulating the decay of the mRNA or by inhibiting its interaction with RNA binding proteins (Gao and Cooper, 2013; López-Morató et al., 2018). Although each disorder will require its own tailored therapy, approaches targeting a specific mechanism (e.g. RNA toxicity) could potentially work for many diseases.

The holy grail of therapy is to repair the disease-causing gene itself instead of its downstream effects. Repairing the mutation comes with several advantages: there is no need to understand the complete biology, the repair is irreversible and will cause less off-target effects. Recent advances in genome engineering, and especially the discovery of clustered regularly interspaced short palindromic repeats/CRISPR-associated systems (CRISPR/CAS9), make gene therapy more feasible (Shin and Lee, 2018). This approach was already validated in induced pluripotent stem cells (iPSCs) from a patient with FXS. Excision of the long CGG repeat reversed methylation and restored the expression of *FMR1* mRNA and protein (Park et al., 2015). Interestingly, even if systematically supplying an organism with CRISPR/CAS9 might remain challenging, applying gene correction on pluripotent stem cells followed by differentiation and transplantation might become possible in the near future (Maffioletti et al., 2015; Saverio Tedesco et al., 2012).

1.3. An Overview of Technologies to Analyze Short Tandem Repeats

The repetitive nature of STRs makes their analysis by current technologies challenging. Although there are multiple options to assess STRs, all platforms have their drawbacks and pitfalls. Here we discuss the most frequently used platforms for STR analysis.

1.3.1. Traditional Techniques

PCR

PCR is one of the most common tools in molecular biology and is used to exponentially amplify a specific region of the genome. Therefore, it can be used for the analysis of STRs whereby PCR is followed by fragment analysis to determine the repeat lengths. It is a wide-spread, easy and cheap method whereby regions up to ~15 kb can be amplified (Jia et al., 2014).

Unfortunately, PCR is very error prone when amplifying STRs due to their repetitive nature, sometimes in combination with a high GC-content (Table 2). This combination is so harsh that some regions are hitherto not amplifiable (Biasiotto et al., 2017; Braida et al., 2010). In addition, often a normal and an expanded repeat are present. Since PCR preferably amplifies the smaller, normal allele there is a high risk that the expanded allele will be missed (Chakraborty et al., 2016). This can partially be overcome by triplet-primed PCR (TP-PCR). This is an indirect method whereby the forward and reverse primer of a standard PCR are complemented with a third primer that will anneal right into the repeat. By adding the third primer, the PCR will produce a ladder of peaks that will be visible on an agarose gel or electropherogram as a smear, even if expanded STR alleles are too large for amplification with the forward and reverse primer. Furthermore, the presence of an interruption within the repeat will impede the annealing of the repeat primer that will be visible as a signal drop after fragment analysis. TP-PCR is therefore able to indicate the presence of repeat interruptions (Chen et al., 2010; Filipovic-Sadic et al., 2010; Seneca et al., 2012). In order to determine the length of a STR, PCR is followed by fragment analysis. Since this only reveals the size of the PCR product, it is not able to detect if a polymorphism occurs in the repeat or next to the repeat, causing misinterpretation of the results for some cases (Mononen et al., 2007; Radvansky et al., 2011).

PCR amplification followed by fragment analysis is currently the most frequent technology to analyze tandem repeats (Liljegren et al., 2016). Although PCR will introduce errors, this approach will allow the determination of the main allele length. This is sufficient for applications such as forensics or linkage mapping. However, since PCR is unable to faithfully replicate STRs, it will camouflage the underlying STR variability.

Southern blot

Southern blot is able to investigate one particular locus surrounded by a complex mixture of genomic DNA. This genomic DNA is digested with restriction enzymes. Subsequently, the DNA fragments are size separated by agarose gel electrophoresis.

After transfer to a nylon membrane, the fragment of interest can be visualized by a labeled probe complementary to the target locus. Southern blot allows to visualize large expanded alleles and, if methylation-sensitive restriction enzymes are used, also the presence of methylation can be identified (Table 2). These advantages make that Southern blot is also used for diagnostic purposes like for example the detection of long, expanded CGG repeats underlying FXS or DM1 (Figure 4)(Biancalana et al., 2015).

The use of Southern blot also comes with various problems. One of the largest drawbacks is the low resolution of its results. Additionally, expanded repeats are often very variable and are present as a diffuse smear rather than a single band. This is difficult to observe on a Southern blot and impedes the detection of minor alleles (Biancalana et al., 2015). Since Southern blot requires large amounts of DNA and is also time-consuming, labor intensive and not scalable, most scientists avoid the use of this technique (Ameur et al., 2018; Liu et al., 2017a).

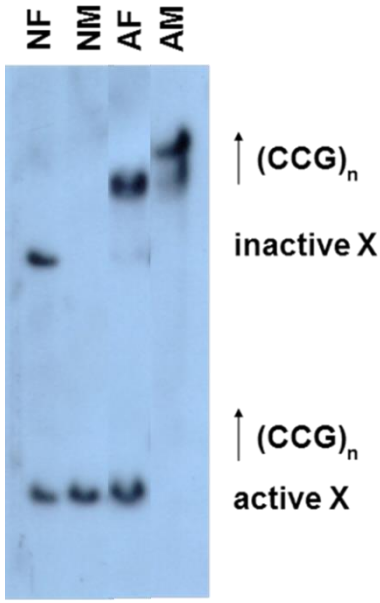


Figure 4: Southern blot of *FMR1*; A: affected; N: not affected; M: male; F: Female. Southern blot can differentiate methylated (inactive) alleles from non-methylated (active alleles). Affected individuals have bands corresponding to higher repeat numbers (picture courtesy of G.M. & V.R.).

1.3.2. Sequencing Techniques

Sequencing techniques are very useful for STR analysis since they reveal the composition of the DNA molecule on a nucleotide level. Here, both massively parallel sequencing (MPS) and long-read sequencing are discussed. The section on long-read sequencing is more elaborate since this is the main technology used in this thesis.

Massively Parallel Sequencing

The advent of MPS or next-generation sequencing (NGS) has unleashed a revolution in genetics. These sequencing technologies generate millions or billions of reads enabling scientists to investigate biological questions on a genome-wide scale and to screen for causative mutations in gene panels and full genomes in patients (Biesecker and Green, 2014).

There are different providers of MPS on the market, but their technologies all share some similar characteristics: they generate millions or billions of short reads (up to 600 bp) from clonally amplified DNA clusters in a high-throughput and cheap manner (Table 2). Most MPS data focusing on STRs was produced by 454 pyrosequencing since this platform can produce relative long reads with a length of 600 bp (Børsting and Morling, 2015; Duitama et al., 2014). The most commonly used MPS platforms are the sequencing by synthesis apparatuses from Illumina (Alkan et al., 2011). Although these platforms can generate billions of reads, the length is limited to 300 bp (Table 3). In spite of these short read lengths, Illumina sequencers are also being used for STR analysis (Bornman et al., 2012; Sharma et al., 2017; Zeng et al., 2015).

STR analysis by MPS involves some major drawbacks (Table 2). Firstly, the quality of the reads containing STRs is often inferior because the clonal amplification brings in the disadvantages associated with PCR (cf. supra) while the repetitive nature of the STR stimulates the loss of phase coherence of the sequencing polymerases (Ameur et al., 2018; Loomis et al., 2013). Even if a STR is sequenced successfully, the read length remains too short to span (long) STRs and will generate reads without or with only one flanking sequence. Therefore, it is impossible to decipher the repeat length and/or the location.

The poor performance of STR analysis by MPS has gained notoriety and different strategies have been developed to overcome this. For example, the advent of a PCR-free library preparation method and the development of several, novel algorithms have improved STR detection (Bahlo et al., 2018; Dolzhenko et al., 2017). Unfortunately, STR analysis by MPS will always be impacted by the clonal amplification and limited read lengths that belong to the inherent nature of the technologies. Hence, MPS is not suited to analyze (expanded) STRs.

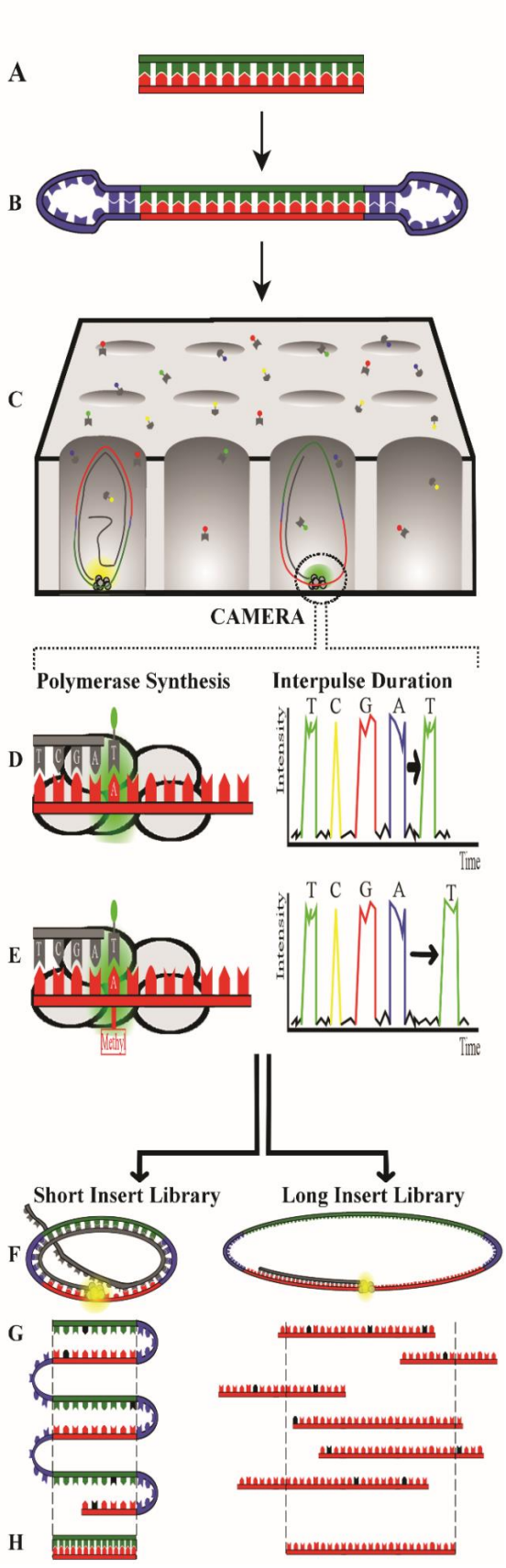
Long-read Sequencing

Long-read or third-generation sequencing are single molecule sequencing techniques and are fundamentally different from clonal based second-generation sequencing methods. Two technologies, Single Molecule Real-Time (SMRT) sequencing developed by Pacific Biosciences (Eid et al., 2009) and nanopore sequencing developed by Oxford Nanopore Technologies (ONT)(Clarke et al., 2009), are currently commercially available.

SMRT sequencing is a sequencing technology developed by Pacific Biosciences. Before sequencing, a library is prepared from double stranded DNA (Figure 5A) input material to which hairpin adapters are ligated (Figure 5B)(Travers et al., 2010). Sequencing of a library is performed on a SMRT Cell that contains 150,000 nanoscale observation chambers (Zero Mode Waveguides (ZMWs)) for the RSII system and up to a million on the newer Sequel platform. Ideally one molecule is loaded per ZMW to maximize throughput and read lengths (Figure 5C). In practice, about 1/3 to 1/2 of the ZMWs per SMRT cell are loaded. Hence a SMRT cell produces on average 55,000 reads for the RSII system and 365,000 reads for the Sequel system (Table 3). The actual sequencing reaction occurs within each ZMW (Rhoads and Au, 2015). The fluorescence emitted by the nucleotides is recorded by a camera in real-time. These signals are converted to long sequences termed continuous long reads (CLR), linear reads, or polymerase reads. Due to the circular structure of the library, a short insert will be covered multiple times by the continuous long read (CLR). Each pass of the original DNA molecule is termed a subread, which can be combined into one highly accurate consensus sequence termed a circular consensus sequence (CCS) or reads-of- insert (ROI) (Figure 5F–H, left panel). Though SMRT sequencing always uses a circular template, long insert libraries typically only have a single pass and hence generate a linear sequence with single pass error rates (Figure 5F-G, black nucleotides at right panel). Afterwards, overlapping single passes can be combined into one consensus sequence of high quality (H, right panel). Overall, CCS reads have the advantage of being very accurate while single passes stand out for their long read lengths (>20 kb). Due to the real-time detection of the nucleotide incorporation rate, the pace of the polymerase progressing through the DNA strand is registered during sequencing (Eid et al., 2009). Since this varies with modifications present on top of the DNA, recording the time between nucleotide incorporations allows detection of DNA modifications (Figure 5D-E).

Since a polymerase is tied to about twelve nucleotides, a DNA modification on one nucleotide can actually affect the incorporation rate of surrounding nucleotides. This results in a "fingerprint" some of which have been characterized, such as for 6-mA, 4-mC, and (Tet-converted) 5-mC. Unfortunately, 5-mC, one of the most widespread modifications in humans, has only a very subtle influence on the pace of the polymerase. Hence, a high coverage (100-250X) is necessary to pick up this modification type (Feng et al., 2013; Schadt et al., 2013).

Figure 5: Overview of SMRT sequencing technology. Double stranded input DNA before (A) and after library preparation (B). A library is then loaded and sequenced on a SMRT Cell (C). Note that not all ZMWs will contain a DNA molecule because the library is loaded by diffusion. Since the emitted fluorescence of the incorporated bases is emitted in real-time, also the time between nucleotide incorporation that is called the interpulse duration (IPD) (D, right panel). When a sequencing polymerase encounters a nucleotide on the DNA strand containing a modification, like for example a 6-methyl adenosine modification (E, left panel), then the IPD will be delayed (E, right panel) compared to non-methylated DNA (D, right panel). The long reads can span a short insert multiple times that can be combined in a highly accurate consensus molecule (F-H, right panel). Long insert libraries will only be passed one time and contain sequencing errors (black nucleotides) but can be combined in a consensus sequence if they overlap.



PacBio data differs from short read MPS sequencing technologies in several aspects. Reads are not a set read length, but a distribution of read lengths depending on how long each individual polymerase is active (Table 3). Since there is no need for amplification during the library preparation, nor during the sequencing process, biases such as GC-skewing are near absent (Table 2). In contrary to MPS platforms, raw PacBio reads also differ in error types (more indels than mismatches) and have a much higher abundance (~13-15%, Table 3). However, sequencing errors are spread randomly across the reads (Figure 6)(Carneiro et al., 2012; Chaisson et al., 2015). This randomness enables highly accurate consensuses (>99.9%) to be build up rapidly by sequencing multiple times the same molecule (CCS reads) or by combining different CLRs derived from the same locus (Figure 5G-H)(Hestand et al., 2016b; Koren et al., 2012).

SMRT sequencing has gained notoriety in the STR field thanks to the promising potential of the long reads and the high accuracy. Loomis et al.(2013) showed they could sequence through a long full FMR1 mutation allele of 750 units that equals 2 kb of 100% GC and repetitive content. Another example of tackling a tandem repeat by SMRT sequencing is the ATTCT repeat embedded in intron 9 of *ATXN10* gene causing SCA10. Long read sequencing allowed to reconstruct for the first time the full length of an expanded ATTCT repeat. This revealed both known and novel interruptions that influence the phenotype of SCA10 patients. Hence, knowing the exact repeat structure allows for better genotype-phenotype correlations (McFarland et al., 2015).

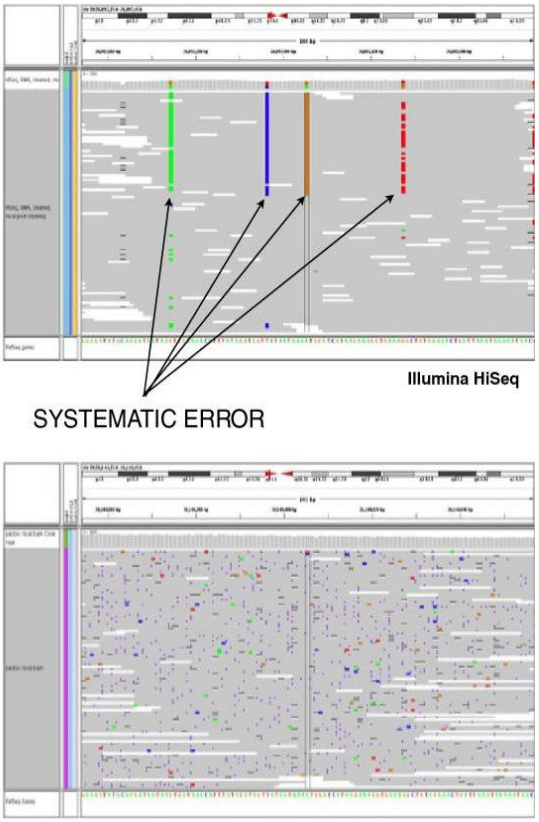


Figure 6: IGV snapshot of data from Illumina NGS sequencing (upper panel) and SMRT sequencing on a PacBio RSII instrument (lower panel). Illumina data hints towards the presence of heterozygosity due to context specific errors. SMRTs sequencing produces more errors, but they are spread completely randomly. SMRT sequencing shows that there is no event in the displayed region. (Figure adapted from Carneiro et al., 2012).

Despite the clear advantages of SMRT sequencing, some limitations remain. For example, large input amounts of typically 5 micrograms are necessary and only a limited throughput is achieved (Travers et al., 2010).

Table 2: Overview of the advantages and disadvantages of different technologies

	Advantages	Disadvantages
PCR	• Cheap • Analysis of regions up to 10 kb • High-throughput • Indication of interruptions	• Error prone, especially in GC and repetitive regions • Erasure of epigenetic patterns • Risk of misinterpretation
Southern Blot	• Information on repeat size and DNA methylation • Sizing of long, expanded repeats	• Low resolution: no detection of minor alleles • Time-consuming • Results are difficult to interpret • Low throughput
Massively Parallel Sequencing	• High throughput • Cheap • Single-nucleotide resolution	• Short reads • Inferior quality
SMRT Sequencing	• Long reads • Detection of base modification in DNA • High accuracy • Single-nucleotide resolution • No GC-bias	• High amount of input DNA necessary • Low-throughput • Expensive
Nanopore Sequencing	• Long Reads • Detection of base modification in DNA and RNA • Single-nucleotide resolution • Compact Devices • Cheap • Fast library preparation	• High error rate • Presence of systematic errors

Nanopore sequencing is also a third generation single-molecule sequencing platform. Here, a nanopore is embedded within an electrically-resistant polymer membrane that separates 2 ionic solutions. An electric current is applied over this membrane and monitored at the nanopores by a sensor chip. The electric current will change when a DNA molecule is dredged through a nanopore. Actually, since each base (A, G, C, T) generates a different distortion of the electric current, monitoring these distortions reveals the code of the sequenced DNA strand. Different nanopores are organized in flow cells to scale-up the throughput. For example, the Promethion platform consists of 48 flow cells that each contain 3000 nanopore channels (Table 3).

The main advantages of nanopore sequencing are the absence of PCR, the production of very long reads and the identification of base modifications due to the real-time monitoring of the current (Rand et al., 2017; van Dijk et al., 2018).

The read lengths generated in nanopore sequencing are only limited by the length of the provided DNA templates that exceptionally results in reads exceeding 1 Mb (Rand et al., 2017). This is in contrast with SMRT sequencing from Pacific Biosciences where the lifetime of the sequencing polymerase is the main factor determining the read length, limiting the maximum read lengths (Table 3). Furthermore, nanopore sequencing has some additional advantages. The nanopore sequencing suit consists of portable instruments with the smallest one not larger than a USB stick (MinION), which makes this apparatus very interesting for sequencing in the field (van Dijk et al., 2018). The sequencers are also cheap, require a low investment cost and only need a short library prep (Table 3).

It should be noted that nanopore sequencing is not yet fully matured. For example, the run parameters promised by ONT are in general impossible to replicate by field sites and production problems regularly occur (van Dijk et al., 2018). Furthermore, nanopore sequencing is not suited for applications that require a high accuracy. In contrast to PacBio sequencing, error rates (15%) cannot be reduced by consensus reads. However, nowadays it is possible to sequence both strands of the DNA template (called $1D^2$) which reduces the error rate to 3% (Table 3). Unfortunately it is not very efficient, it negatively impacts the throughput and some systematic errors remain present in the data (Mitsuhashi et al., 2017). In spite of the high error rate, some publications do report on the analysis of STRs by nanopore sequencing. Although significant variation was present in their data, they could determine approximate STR read length (Ishiura et al., 2018; Liu et al., 2017b; Roeck et al., 2018; Tang et al., 2017). However, base calling accuracy within repeat region was higher for SMRT sequencing than nanopore sequencing (Ebbert et al., 2018).

Table 3: Comparison of different sequencing platforms

	MPS	**Long-Read Sequencing**			
	Illumina	**PacBio**		**Nanopore**	
	Hiseq 4000	**RSII**	**Sequel**	**MinION**	**PromethIon**
Output (Tb)	1.5	0.001	0.01	0.02	6
Mean Read lengths	2*150 bp	>20 kb	>20 kb	variable	variable
Max. read lengths	300 bp	>80 kb	>300 kb	>1 Mb	>1 Mb
Error rate (%)	~ 0,1	0.001 (CCS) 13-15 (CLR)	0.001 (CCS) 13-15 (CLR)	3 ($1D^2$) 15 (1D)	3 ($1D^2$) 15 (1D)
Error type	Substitutions	Indels	Indels	Deletions	Deletions
Run Time	<1-3.5 days	0.5- 6 hours	0.5- 10 hours	2 days	2 days
Instrument cost (€)	800,000	600,000	300,000	1000	120,000
Library cost (€)	100	250	250	500	500
Cost/run (€)	2500	350	750	600	1150
Hands-on time	5h	4h	4h	10 min	10 min

Numbers from company websites https://nanoporetech.com and www.illumina.com queried on 15-09-2018 and the publications: Akogwu et al. (2016), Ameur et al. (2018) and van Dijk et al. (2018).

1.4. CRISPR/CAS9

In 2012 scientists showed that CRISPR/CAS9 allows precise and efficient genome editing both *in vitro* and in living eukaryotic cells (Jinek et al., 2012). From then on, CRISPR/CAS9 became very rapidly widespread in the scientific community where it revolutionized genome editing and (medical) research.

The CRISPR/CAS9 system is the backbone of the adaptive immune system in many bacteria and archaea (Karvelis et al., 2013a). First, during an immunization phase, a memory is built from the invading, viral, phage or plasmid DNA. This is done by integrating snippets of the invading DNA into the CRISPR locus as spacers (Figure 7, upper panel)(Mali et al., 2013). In a second phase, the stored information is transcribed into RNA (crRNA) together with other components of the CRISPR/CAS9 system (inter alia tracrRNA and CAS9). Afterwards, a complex is formed between crRNA, tracrRNA and CAS9 and processed by RNaseIII. Once the bacteria or archaea is now invaded with new viral DNA, the matured CRISPR/CAS9 complex will start screening the DNA for protospacer-adjacent motif (PAM) sequences in the viral DNA. At every PAM site CRISPR/CAS9 will check if the region is complementary with the crRNA. Finally, the viral DNA will be degraded by a double stranded cut if a successful match occurs between the crRNA and the viral DNA (Figure 7, lower panel)(Horvath and Barrangou, 2010). A PAM site is for example NGG in *Streptococcus pyogenes* and has 2 important functions. Firstly, it serves as a key to differentiate between the DNA from the host and the virus. Additionally, it also allows a fast screen of the viral DNA since complementary between the crRNA and the virus is only checked at these PAM sites (Sternberg et al., 2014).

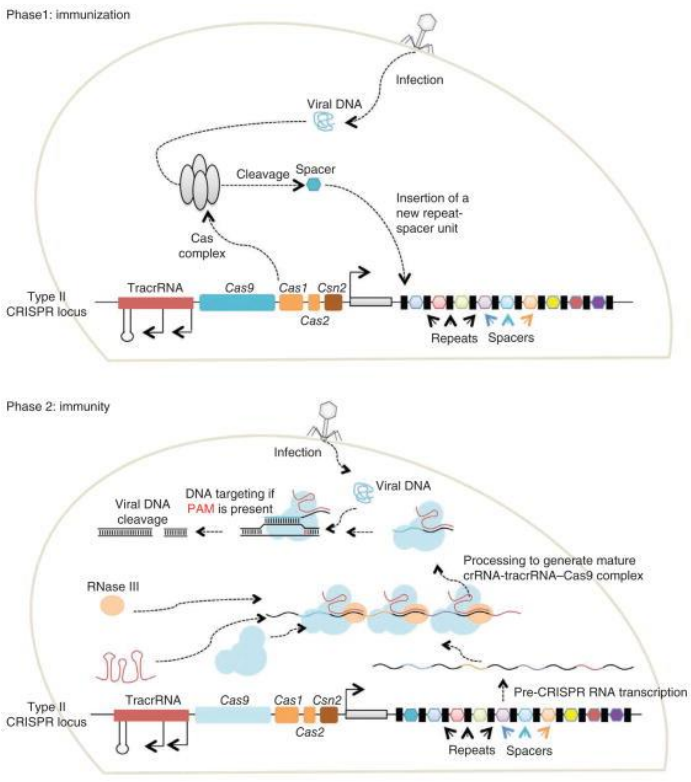


Figure 7: The functioning of CRISPR/CAS9 as an immune defence mechanism in bacteria and archaea. Firstly, invading viral DNA is processed and small DNA fragments are integrated in the host DNA separated by repeats. Afterwards, the integrated viral DNA fragments will be transcribed, form a complex with tracrRNA and Cas9 and processed by RNAIII. If viral DNA invades the host again during a second invasion, the matured crRNA-tracrRNA-CAS9 complex will screen the invading DNA at every PAM site for complementary with the crRNA. If there is a successful match, the viral DNA will be cleaved by CAS9 (figure adapted from Mali et al. (2013)).

After understanding the mechanisms of CRISPR/CAS9, scientists realized its potential as a genome engineering tool. In bacteria the system can be used as it is, but for its use in human small modifications needed to be done. For example, the codon sequence of Cas9 is optimized for humans and an appropriate nuclear localization signal was added to the protein (Hsu et al., 2013). Another improvement was the fusion of the tracrRNA and crRNA into only one chimeric short-guide RNA (sgRNA)(Figure8). This comes with the advantage that only two components are necessary for genome editing: the Cas9 protein and a sgRNA. This makes the CRISPR/CAS9 system not only powerful and precise, but also very user friendly (Doench et al., 2014). In addition it outperforms other genetic modification tools like zinc-finger nucleases (ZFN) and transcription-activator like effector nucleases (TALEN) that are complex, costly and time-consuming (Gaj et al., 2013) and RNAi that only results in temporary inhibition of gene expression (Elbashir et al., 2002).

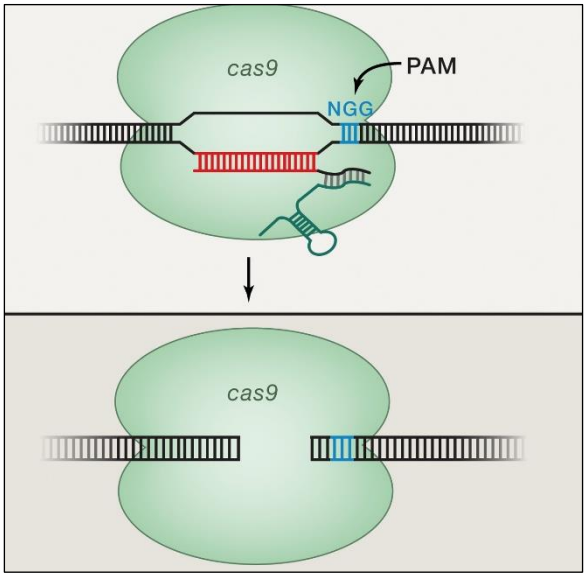


Figure 8: Schematic representation of CRISPR/CAS9 cutting. Upon the presence of a NGG PAM site, complementary between the sgRNA and the DNA region is checked (upper panel). A successful match between the sgRNA and the DNA indicates that the CRISPR/CAS9 complex identified the target region after which a double stranded is made (lower panel)(figure adapted from Lander (2016)).

The possibilities of CRISPR/CAS9 in molecular and cell biology are numerous. The system is mostly used i*n vivo* where CRISPR/CAS9 is used to knock-out a specific gene in eukaryotic cells from human, mouse, zebrafish and yeast, amongst others (Lander, 2016) . This can be done by making a double stranded cut whereafter repair by non-homologous end joining may subsequently introduce small indels and induce gene silencing (Hsu et al., 2013). Besides, also more specific mutations can be created when in addition to CRISPR/CAS9 also a template DNA for repair is provided. This will stimulate the homology directed repair pathway after creation of the double stranded break and allow to introduce for instance a specific point mutation or deletion (Mali et al., 2013a; Xu et al., 2014).

Interestingly, CRISPR/CAS9 can also be used *in vitro*, although this has received much less attention (Liu et al., 2015). Different tools and websites have been developed to help scientists with designing efficient guides with minimal off-target effects (e.g. Benchling or https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design). After designing sgRNA's, they can be produced locally but lately they can also be ordered from suppliers of nucleic acids like IDT. The latter is especially interesting because it ensures quality of the used sgRNA's, avoids contamination and is relatively cheap.

The CAS9 protein can also be produced on site or can easily be purchased from different providers (e.g. New England Biolabs (NEB), IDT or ThermoFischer). By combining a sgRNA, CAS9 and DNA, a double stranded cut can be made at almost any genomic position in a single reaction tube in ≈15 minutes. This system is for example more flexible than restriction enzymes which do not cut uniquely and are limited to a fixed recognition sequence (Wang et al., 2015). This opens up novel possibilities in the design of molecular assays which will be explored in this thesis (Karvelis et al., 2013b). For example, by using not one but two sgRNA's closely located to each other, a target locus can be completely excised from the genome. In addition to CAS9, also a CAS9 nickase and a dead CAS9 (dCAS9) have been engineered and further fuel the use of CRISPR/CAS9 (Mali et al., 2013a). CAS9 nickases makes a single stranded break and can be used to increase the specificity by designing 2 sgRNA targeting the opposite strands at a particular locus (Sternberg et al., 2014). dCAS9 does not cut the DNA at all, but still has the possibility to match the sgRNA with the target locus. Interestingly, dCAS9 can be coupled to enzymes inducing base modifications or adding fluorescent groups at a specific locus (Eid et al., 2018; Mali et al., 2013b; WareJoncas et al., 2018).

Chapter 2: Research Objectives

The general aim of this thesis is to make advantage of the power of Single Molecule Real-Time sequencing developed by Pacific Biosciences to study STRs. The goal is to develop novel methodologies that make maximal use of the assets of long-read sequencing.

Specifically:

1. We aim to develop an amplification-free enrichment method targeting the *FMR1* CGG repeat causing FXS (Chapter 3). This would allow the determination of the true biological genetic and epigenetic variation of the repeat in an economically efficient manner.

2. We aim to develop a method directly detecting AGG interruptions residing within the *FMR1* CGG repeat (Chapter 4). This method can positively impact *FMR1* research and diagnostics since these AGG's influence the stability of the *FMR1* repeat.

3. We aim to develop methods that facilitate the study of the *DMPK* CTG repeat underlying DM1 (Chapter 5). More specifically, we intent to establish a method to investigate the genetic variability of large *DMPK* CTG repeats and to determine the efficiency of CRISPR/CAS9 excision of the repeat.

Chapter 3: Development of an amplification-free enrichment method targeting the FMR1 CGG repeat

3.1. Abstract

Single Molecule Real-Time sequencing is a powerful technology to assess the genetic and epigenetic patterns of short tandem repeats (STRs). However, whole genome single molecule sequencing is too costly for single locus analyses. Hence, it is necessary to develop a targeted amplification-free enrichment method to reduce costs and increase throughput, while at the same time avoiding biases created from amplification. Therefore, we have explored a CRISPR-CAS9 based approach to excise the *FMR1* CGG repeat in combination with restriction enzymes. We present a proof-of-concept for profiling human *FMR1* CGG repeats for size determination and methylation status.

3.2. Introduction

Short tandem repeats (STRs) have been the focus of intensive research because they are associated with various cancers and more than 40 hereditary disorders like fragile X syndrome (FXS), myotonic dystrophy (DM) and Huntington's disease (HD)(López Castel et al., 2010). Although each disease has its own specificities, they all share similar mechanisms and characteristics. In general, healthy individuals carry a moderate number of repeat units that can expand over several generations to large chains of head-to-tail repeated units causing different diseases (McMurray, 2010). In this study we have focused on (FXS) which is the most common heritable form of intellectual disability (Usdin et al., 2014). The CGG repeat in the 5' untranslated region (UTR) of the fragile X mental retardation 1 gene (*FMR1)* can expand to hundreds of repeat units, which depends mainly on the size of the repeat and on AGG units interspersing within the CGG repeat. In general, larger repeats with fewer AGG units have a higher risk that the CGG repeat will expand in future generations (Nolin et al., 2015). Besides genetic factors, epigenetic processes also play an important role in STR disorders (Evans-Galea et al., 2013b). This is most profound in FXS, where methylation of long CGG repeats contributes to the complete silencing of the gene (Penagarikano et al., 2007). Likewise epigenetic processes also influence the pathophysiology of other TR disorders (Evans-Galea et al., 2013b).

Despite the clinical importance of STRs, a good technology to assess their genetic and epigenetic profile is lacking so far. Where short STRs (< 150 bases) can still be characterized by standard molecular assays, like PCR followed by Sanger, and short read massively parallel sequencing (MPS), this becomes extremely cumbersome for the longer, disease causing STRs. One approach to overcome these limitations is triplet-primed PCR (TP-PCR). This can be used to flag patients with longer STRs and indicate the presence of interruptions, but it fails to determine the exact number of repeat units and is prone to false positives and negatives (Figure 9A)(Braida et al., 2010; Chen et al., 2010). Southern blot can detect large alleles and methylation patterns, but requires large DNA inputs, has a limited resolution and does not detect minor alleles (Figure 9B)(Loomis et al., 2013; Singh et al., 2014). The advent of Massively Parallel Sequencing (MPS) methods was also not able to solve any of the above problems. Additionally, due to the short read lengths, MPS is unable to span long STRs (Figure 9C – section 2)(Quilez et al., 2016). It is also a disadvantage that MPS is based on observing clusters of multiple molecules. When such a cluster contains a STR, the sequencing polymerases moves with a different speed across the different DNA molecules of the same group, which results in dephasing and inferior sequencing results for STRs (Duitama et al., 2014; Loomis et al., 2013). Even when MPS succeeds in sequencing STRs, alignment is often difficult and wrong repeat numbers might be called (Hannan, 2018a).

Another major drawback is that PCR is involved before and/or during sequencing. PCR amplification induces stutter and impedes the correct determination of STR variability (Figure 9C - section 3). Furthermore, some STRs, especially GC rich repeats, are completely recalcitrant to amplification (Van Blitterswijk et al., 2012). Moreover, none of the above methods is able to generate a direct read-out of STR epigenetics.

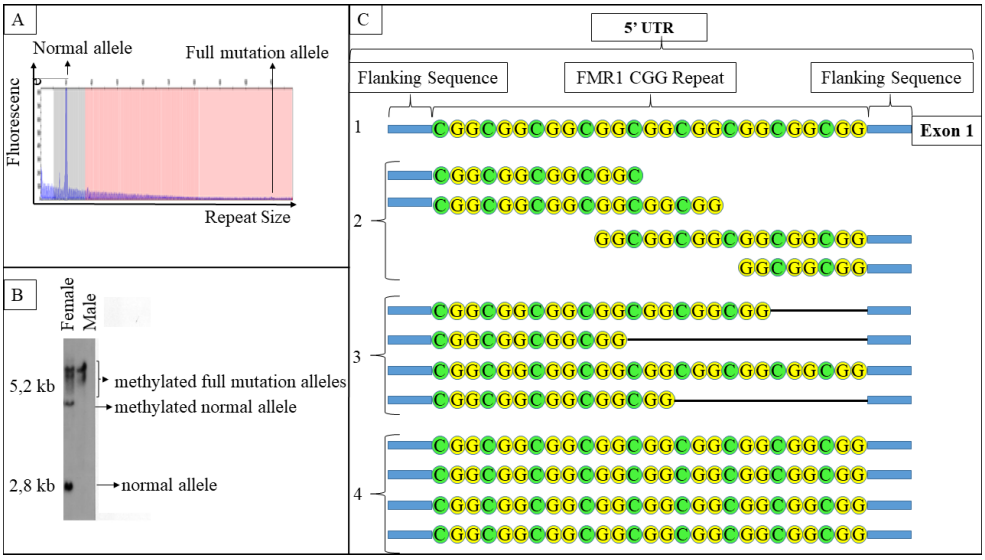


Figure 9: Result of a TP-PCR (A) and a Southern blot (B) of the *FMR1* CGG region. C1 shows the reference *FMR1* region. Two drawbacks of MPS are the short reads unable to span the entire repeat (C2) and the involvement of PCR before and/or during sequencing which induces stutter (C3). This is overcome by SMRT sequencing where the entire repeat can be spanned by the long reads (C4).

Technical shortcomings hamper the ability to accurately assess the degree of genetic and epigenetic mosaicism in STR disorders, which is important to completely understand the penetrance, complexity and phenotypical variation of the disorders (Quilez et al., 2016). For example, the degree of STR variability between and within individuals and their functional consequences is largely unknown (Duitama et al., 2014). *FMR1* pre- and full mutation alleles carry a large set of CGG repeats predisposing these alleles to instability (Pretto et al. 2014a, 2014b). Therefore, almost all full mutation carriers have repeats with different sizes within the full mutation zone instead of just a single allele size. Even more extreme size mosaicism is found in up to 41% of patients with FXS that carry both pre- and full mutations. Interestingly these patients have a better cognitive functioning because some fragile X mental retardation protein (FMRP) will be produced from the premutation allele. On the downside, the presence of premutation alleles also entails the risk of developing additional clinical features such as fragile X-associated tremor/ataxia syndrome (FXTAS; MIM#300623) and fragile X-associated primary ovarian insufficiency (FXPOI; MIM#311360)(Santa María et al., 2013). These disorders are typically associated with the premutation allele. In addition to size mosaicism, the extent of methylation can also vary in fragile X-patients. For instance, a patient can have a completely methylated full mutation and in addition carry unmethylated alleles ranging in size from the premutation all the way up to the full mutation zone. Unmethylated pre- and full mutation alleles can be transcribed, albeit with a decreased efficiency, and hence also influence the fragile X phenotype (Pretto et al., 2014b).

Single-Molecule Real-Time (SMRT) sequencing developed by Pacific Biosciences holds the potential to reveal DNA modifications on the *FMR1* CGG repeat. Firstly, this technology generates long and accurate reads that span the expanded, disease causing STRs completely. Secondly, it can detect DNA modifications simultaneously with the genetic code (Loomis et al., 2013; Nakano et al., 2017a).

Some modifications like N4-methylcytosine (m4C) and N6-methyladenosine (6mA) have a big influence on the kinetics of the polymerase and thus are easy to detect in the sequencing data. By contrast, the influence of other modifications like 5-methylcytosine (5mC) is subtler. Unfortunately, SMRT sequencing only has a limited throughput as it generates around 50,000 reads or 0.75 billion bases per SMRT Cell on the RSII instrument (Rhoads and Au, 2015). This makes it economically unfeasible to sequence the entire human genome only to look at the *FMR1* locus. Typical amplification methods of the repeat would introduce variation in size and remove methylation signals. In order to evade these drawbacks, a targeted, amplification-free method enriching the *FMR1* CGG repeat was developed.

3.3. Materials and Methods

3.3.1. DNA Samples

The BAC RP11-37p24 was used to validate the developed enrichment methods. This BAC contains a *FMR1* CGG repeat with 11 units, but it is smaller (143.5 kb) compared to the human genome ($3*10^9$ kb). Hence, it was easier to assess enrichment of the BAC molecule. BAC molecules were extracted from E. *Coli* with the NucleoBond® Xtra BAC (Macherey-Nagel, Duren, Germany).

DNA from a male FXS patient with 600 *FMR1* CGG repeats (for the sequencing of long STRs) and DNA from a healthy male with 19 *FMR1* CGG repeats (for the validation of 5mC detection) were isolated from peripheral white blood cells according to standard procedures. Male individuals were selected since they carry only a single X chromosome and thus also only one *FMR1* CGG repeat. The targeted, amplification-free enrichment method was also applied to DNA from a healthy female with 20 and 23 *FMR1* CGG repeats according to PCR sizing. DNA from this female was extracted from cultured, human dermal fibroblasts with the Wizard Genomic DNA Purification Kit (Promega, Madison, WI) that yields large DNA molecules.

This study was approved by the local ethical committee and informed consent was obtained from all patients.

3.3.2. Sequencing long FMR1 CGG repeats

Before developing an enrichment method, the potential of SMRT sequencing to span long STRs was explored. The full mutation allele of a male fragile X patient with 600 CGG repeats was amplified with the AmplideX *FMR1* PCR kit (Asuragen, Austin, TX) and with previously published primers (Filipovic-Sadic et al., 2010). Afterwards, the PCR product was purified with Ampure Beads (Beckman Coulter, Brea, CA) and subjected to PacBio sequencing.

3.3.3. Design of sgRNAs

In order to enrich the *FMR1* CGG repeat without amplification, first of all a cut was made both up- and downstream of the repeat with the CRISPR-CAS9 system. This system uses an RNA molecule (sgRNA) that contained a specific target to guide the CRISPR-CAS9 protein to a specific position where the DNA strand will be cut. Four guides that cut upstream and 12 guides that cut downstream were designed with a webtool from the Zhang lab (http://crispr.mit.edu/)(Hsu et al., 2013). An overview of the ordered sgRNA's is shown in Table 4. All guides were ordered directly as RNA molecules at IDT (Coralville, IA).

Table 4: Overview of designed sgRNA's.

		sgRNA		qPCR primers	
Name		**Sequence**	**Distance to CGG**	**Forward**	**Reverse**
Upstream	1	GGTTTCACTTCCGGTGGA	91	CCCAGGCCACTTGAAGAGAG	CGCCGCCCGCTCAGA
	2	ACAGCGTTGATCACGTGACG	176	ACTTGAAGAGAGGGGCGGG	CTCCACCGGAAGTGAAACCG
	3	CGCGCGTCTGTCTTTCGACC	294	ACCTCTGCAGAAATGGGCGT	TCTCTTCAAGTGGCCTGGGAG
	4	GGGTCGAAAGACAGACGCGC	307	ACCTCTGCAGAAATGGGCGT	TCTCTTCAAGTGGCCTGGGAG
Downstream	1	GCGGGCTCCCGGCGCTAGCA	52	AGCCCACCTCTCGGGG	CTGCCGCCAAGTACCTTGTA
	2	CGACACCAAGAAGAAAAGGG	181	GCTCCAATGGCGCTTTCTAC	GGAGAGGGGCTTCCAACAGG
	3	CACGCTCGGCGGGATGTTGT	254	TGGAAAAATCACATGTTGGAGA	AAGGGCAATCAGACTGTAGGAA
	4	CCGGCTCTAGTACCTGCCGC	410	TGTCTTTTGGTCAGAGTGAAGC	GCAAAGGAGAGAAATGAAGGA
	5	TCGGCGCTAAGTGACGGCGA	460	TTTGTGGTGTTGCAGTGGAC	CCAACAAACGCACTACTGCTA
	6	TGTCGTGTGGGTAGTTGTGG	529	CCCAGGCCACTTGAAGAGAG	GCTCCTCCACAACTACCCAC
	7	GAGTAGTAAGAAGCGGTAGT	684	CCAATGGCGCTTTCTACAAG	CCTCGCTGGTCTCTCATTTC
	8	CTTTATATAGGCATTCAT	1112	CTCCGTTTCGGTTTCACTTC	CCTTGTAGAAAGCGCCATTG
	9	TTGGCAATAGAAGGTGCGTG	1330	GATGGCTTATTCCCCCTTTC	CAGTGGAGCTCTCCGAAGTC
	10	TCTCAAATGGTCTGCACTGA	2186	CAGGGCTGAAGAGAAGATGG	AGTCCTTCCCTCCCAACAAC
	11	TCATTTGGTTAGTAGTATTA	3081	GCGAGGAGAGGGTTCTCTTT	CTCCACAACTACCCACACGA
	12	TGTAAATAAACTTGCACTCG	3281	GGAGAGCTCCACTGTTCTGG	GGCCATGTTAGGGTCTTCCT

3.3.4. Cas9 Digestion

Validation of sgRNA's

In order to select the most potent up- and downstream sgRNA, the efficiency of all sgRNA's was determined. Therefore, Cas9 nuclease (NEB) was used to digest 2 µg of the RP11-37p24 BAC according to the manufacturer's instructions. After purification, the efficiency of the Cas9 digestion was determined by qPCR, where the primer pair spanned the cleavage site (Table 4). SYBR Green Master Mix (Thermofisher, Waltham, MA) was used for the 20 ul qPCR reactions (10 ul SYBR Green Master mix, 2 ul DNA input (50 ng/ul), 2 µl forward primer (5 µM), 2 µl reverse primer (5 µM) and 4 ul H_2O) which was incubated at 95° for 5 min followed by 45 cycles at 95°C for 10 sec, 60°C for 20 sec and 72°C for 30 sec were performed. Afterwards, a melting curve was constructed by incubation the sample at 95°C for 5", 65°C for 1 min followed by slow heating to 97°C. Finally, a relative quantification was performed using the $\Delta\Delta C_T$ method. The efficiency could subsequently be determined by comparing the number of cleaved molecules in the Cas9 sample with an untreated control sample.

Digestion of genomic DNA

For each enrichment experiment up to 2 µg BAC DNA and/or 20 µg human DNA was subjected to Cas9 digestion with the most efficient up- and downstream sgRNA together in the same reaction. The Cas9 treatment was performed according to the manufacturer's instructions whereby reagents were scaled if necessary. This reaction has a total volume of 30 µl (10 µl DNA, 10 µL nuclease-free water, 3 µl of 300 nm upstream sgRNA, 3 µl of 300 nm downstream sgRNA and 1 ul of CAS9 protein) and was incubated for 16 hours at 37°C. The efficiency of Cas9 digestion of the human/BAC DNA was controlled by qPCR with primers across the up- and downstream digestion sites.

3.3.5. Library Preparation

Library preparation started with a 0.6X Ampure Bead purification of the input material, whereafter the eluted product was diluted to 135 ng/µl. Next, the genomic DNA was repaired by supplying the DNA with 1X ThermoPol reaction buffer, 1X NAD^+, 10 nmol/µg ATP, 1nmol/µg dNTP's and 0.5 µl/µg PreCR repair mix (NEB) and incubation (37°C, 20 min). This was followed by adding 0.5 µl/µg Nebnext End Repair Enzyme mix (NEB) and incubation (25°C, 5 min) in order to repair the ends of the DNA molecules, which is important to ensure an efficient ligation of the PacBio adapters in the next step. After a 0.6X Ampure Bead Purification, the PacBio adapters were ligated by adding 10 µl/µg of PacBio's blunt adapter, 1X template prep buffer (Pacific Biosciences, Menlo Park, CA), 0.5 nmol/µg ATP, 0.25 µl/µg Quick T4 DNA ligase (NEB) and 0.5 µl/µg H_2O to the mixture, followed by incubation (25°C, 16h). Afterwards, the reaction was stopped by heating the mixture 10 minutes at 65°C. Subsequently, an exonuclease treatment was performed that removes all molecules to which no or only one PacBio adapter annealed (non-circularized molecules). This was done by adding 0.125 µl/µg ExoIII (Westburg, Leusden, The Netherlands), 0.25 µl/µg Exo VII (NEB), 0.25 µl/µg Styl-HF (NEB), 0.25 µl/µg MscI (NEB) and 0.25 µl/µg Nsi-HF (NEB) and 125 µl/µg PUC18 plasmid followed by incubation (37°C, 60 min). The plasmid was not sequenced because it does not contain PacBio adapters. It functions as carrier DNA to avoid the exonuclease treatment being too harsh for the *FMR1* CGG fragment. This is especially important when only little genomic DNA is left at the end of the exonuclease treatment. Finally, 2 successive 0.5X ampure bead purifications were performed.

3.3.6. Complexity reduction

Size selection

Manual or automated size selection can be used to isolate a fragment with a specific size. Automated size selection was done with the BluePippin Size selection System (Sage system, Beverly, MA) whereby the DNA mixture was loaded on a 0.75% gel cassette followed by recovery of the target fragment. For the manual size selection, the DNA mixture was electrophoretically resolved on a 1% agarose gel. Afterwards, a clean scalpel was used to excise the target fragment that was subsequently purified with a gel extraction kit (Qiagen, Hilden, Germany). Then, libraries were prepared from the isolated fragments.

Restriction Digestion

Restriction digestion can be used to make cuts at specific locations in the genome. A standard library preparation was done, except that 3 restriction enzymes (StyI-HF, MscI and NsiI-HF) were supplied to the sample simultaneously with the exonuclease treatment. These enzymes cut the human genome frequently but did not cut within the target fragment excised with CRISPR-CAS9.

3.3.7. SMRT Sequencing

All libraries were prepared and sequenced in duplicates. Each library was loaded on one SMRT cell of a PacBio RSII instrument. Sequencing was done with the DNA/polymerase binding Kit P6 v2, DNA Sequencing Reagent Kit 4.0 v2 and the one-cell-per-well Magbead sequencing protocol for a 360-min movie.

3.3.8. Repeat Size Analysis

Reads were generated with a minimum of 1 full pass and a minimum predicted accuracy of 90% with the RS ReadsOfInsert.1 protocol from PacBio's analysis suite SMRTportal (v2.3.0). These settings assured that the maximum number of reads were recovered from the experiments.

To determine the repeat variability of the *FMR1* CGG allele, reads need to map to the locus and span the repeat completely. These reads (called "on-target" reads) were retrieved by aligning all generated reads using BWA-SW v0.7.10 (Li and Durbin, 2009) against the human reference genome hg19, downloaded from UCSC (Karolchik et al., 2004), followed by conversion of SAM to BAM by Samtools v1.3.1 (Li et al., 2009). Finally, BEDtools v2.20.1 was used to convert the BAM file to BED format and select the on-target reads (Quinlan and Hall, 2010).

The distribution of *FMR1* CGG repeat sizes was finally extracted by a custom python script that recognized the flanking sequences of the *FMR1* CGG repeat and subsequently identified the repeat size of each individual on-target molecule.

3.3.9. Kinetic analysis

Validation of 5mC detection

The influence of 5mC methylation in the *FMR1* CGG repeat was determined by sequencing both methylated and unmethylated control DNA and by determining the difference in the kinetic pattern of both DNA molecules afterwards. Therefore, 2 regions in the DNA from a healthy male were amplified by PCR: *CASK* (containing 5 CG dinucleotides) with forward primer GAGGCCTATGTTGCCTACCA and reverse primer GAGAGGTGGAGGAGTGG TGA and *FMR1* (containing a CGG repeat with 19 units) with previously published primers (Filipovic-Sadic et al., 2010). The PCR products were methylated in vitro by the CpG Methyltransferase M. SssI (New England Biolabs, Ipswich, MA) according to the instructions provided by the supplier. Methylation efficiency was controlled by digestion of the *in vitro* methylated PCR products with the methylation sensitive restriction enzymes AatII and AciI (NEB) for *CASK* and *FMR1* respectively. Finally, a pool of unmethylated *FMR1* and *CASK* PCR products and a pool of methylated *FMR1* and *CASK* PCR products was made whereafter each pool was sequenced on a separate SMRT cell on the RSII PacBio instrument.

Analysis

The kinetic data was analyzed with the RS_Modification_and_Motif_Analysis.1 tool that is included in PacBio's SMRTPortal. If no unamplified control was present, an *in silico* control was used to detect modifications. After analysis both the kinetic profile of the polymerase that is called the Interpulse Duration (IPD) and the detected modifications could be retrieved.

3.4. Results

Long-read single molecule sequencing allows to span long STRs and detect DNA modifications simultaneously (Supplementary Figure 18 and 19, Clark et al., 2013; Loomis et al., 2013). To make full use of these two properties, we developed an amplification-free enrichment method targeting the *FMR1* CGG repeat based on CRISPR-CAS9.

3.4.1. Development of an amplification-free enrichment method targeting the *FMR1* CGG region

In order to get one human *FMR1* CGG molecule from one SMRT cell, an enrichment of at least 16X should be achieved. However, before enriching human DNA, the development of the methodology was done on BAC DNA containing the *FMR1* CGG repeat that is smaller (143 kb) compared to the human genome (3.2 Gb). Firstly, the goal was to excise the *FMR1* CGG repeat from the genome. Therefore, the input DNA (Figure 11A) was treated with CRISPR-CAS9 that was guided by two specific RNA molecules towards a specific up- and downstream position (Figure 11B) where a double stranded cut was made by CRISPR-CAS9 (Figure 11C). Therefore, 4 guides that cut upstream and 12 guides that cut downstream of the repeat were designed and validated. qPCR analysis of all guides showed that upstream guide 2 and downstream guide 12 were the most powerful guides with an efficiency of 90 and 92 percent respectively. This means that by combining these guides a target molecule of 3525 bases will be excised in 83% of the DNA input (Figure 10).

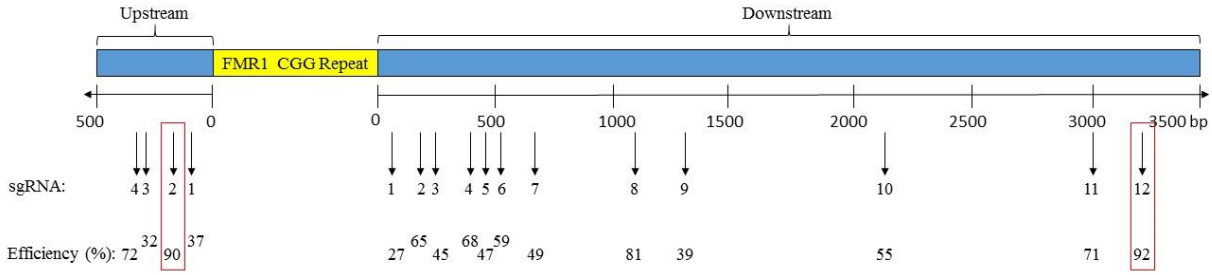


Figure 10: Location and efficiency of the designed guide RNA's. Upstream guide 2 and downstream guide 12 were the most efficient. Therefore, these guides were selected to excise a fragment of 3525 bases centered around the *FMR1* CGG repeat.

Following excision by CRISPR-CAS9 a lot of off-target DNA molecules remained present in the library mixture. Therefore, a quest for an adequate method to enrich the target molecule was started. Interestingly, SMRT sequencing preferentially sequences the smallest molecules in a sequencing library (Loomis et al., 2013). Hence, the excised fragment would be sequenced preferentially over the large genomic DNA. So, firstly the complete mixture of DNA molecules was subjected to a standard PacBio library prep followed by sequencing (Figure 11D & J). The sequencing data showed that the PacBio RSII instrument indeed has a tendency to preferentially sequence the smaller target fragment over the larger DNA molecules (Figure 12). This strategy yielded 32,832 on-target reads, which represents an enrichment factor of 25 (Table 5).

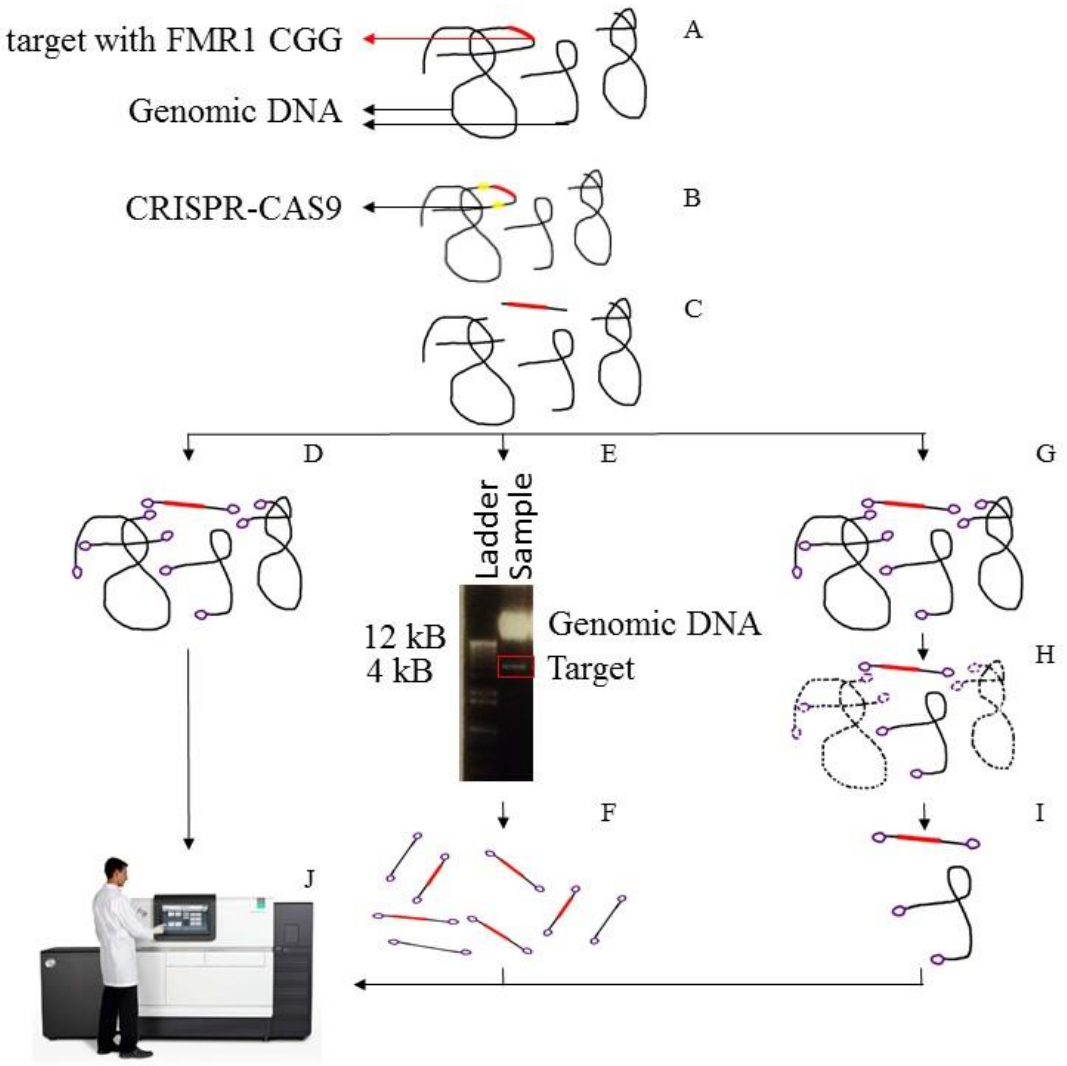


Figure 11: Schematic overview of the different strategies explored to enrich the *FMR1* CGG repeat without amplification. First, DNA was purified (A) and subjected to CRISPR-CAS9 treatment that excised the *FMR1* CGG repeat from the genomic DNA (C). Subsequently PacBio adapters were ligated onto the mixture of DNA fragments (D) followed by sequencing (J). Secondly, the excised fragment was isolated by size selection (E) before library preparation (F) and sequencing (J). Thirdly, PacBio adapters were ligated onto the entire mixture of DNA fragments (G) followed by a simultaneous restriction- and exonuclease digestion (H). The final DNA mixture contained molecules with the *FMR1* CGG repeat and molecules that resisted the restriction digestion and exonuclease treatment (I). This mixture was sequenced completely on a PacBio RSII (J).

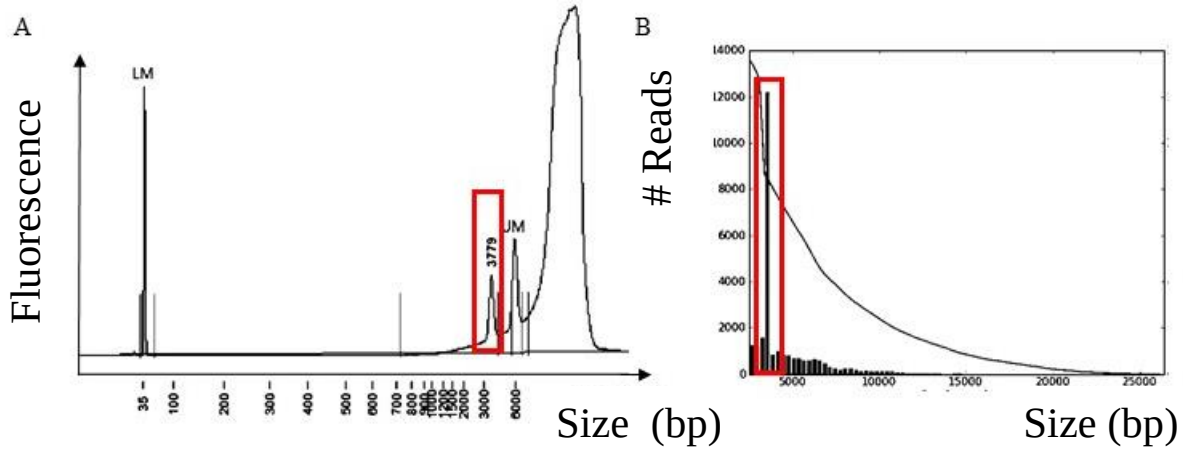


Figure 12: Size distribution of the BAC DNA after CRISPR-CAS9 treatment and library preparation (A) and after SMRT sequencing (B). The target is highlighted with a red box. After the library preparation the amount of target is less compared to the number of large molecules. Strikingly, the distribution shifts completely after sequencing because the PacBio RSII instrument preferentially sequences smaller molecules.

Although this strategy worked extremely well for small BAC molecules, it was expected that the efficiency would drop when this would be applied to the human genome since it is significantly larger than a BAC molecule. Thus, the complexity of the DNA mixture needed to be reduced in order to increase the efficiency of the enrichment. A first strategy consisted of selecting the target fragment by size selection. This implied separating the DNA fragments by size on a gel whereafter the band containing the desired fragment (3525 bases) can be separated from smaller and larger off-target DNA molecules (Figure 11E). A manual excision of the target was chosen over an automatic BluePippin excision because the mean yield was larger for the manual method (0.9 ± 0.1 ng/µg input) compared to the BluePippin (0.4 ± 0.1 ng/µg input). Afterwards, again a library was prepared from the excised fragment (Figure 11F) followed by sequencing (Figure 11J). The percentage of on-target reads increased from 69% (without complexity reduction) to 81.4% after complexity reduction by size selection (Table 5). The percentage of on-target reads did not reach 100% because some molecules with a similar size as the target fragment were excised and sequenced as well (Figure 11F).

As a third approach, the complexity of the sample was reduced by a restriction digestion. Therefore, PacBio adapters were ligated onto the mixture of DNA molecules after CRISPR/CAS9 digestion (Figure 11G). Subsequently, the sample was treated with 3 restriction enzymes (StyI-HF, MscI and NsiI-HF) that cut frequently in the human genome but did not have any recognition sequence within the target molecule. Simultaneous with the restriction digestion, an exonuclease treatment was performed to remove all digested fragments and fragments that were not circularized because they do not contain 2 PacBio adapters (Figure 11H and Figure 13). Finally, also this library (Figure 11I) was subjected to sequencing (Figure 11J).

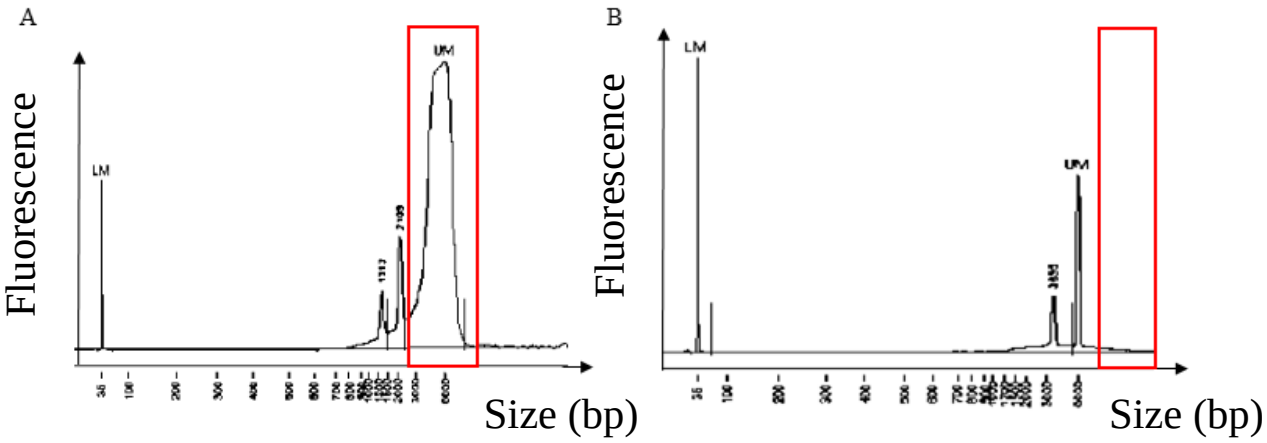


Figure 13: Difference of the size distribution of the BAC DNA after treatment with CRISPR-CAS9 followed by PacBio adapter ligation without (A) and with (B) restriction digestion and exonuclease treatment. In B the majority of the large off-target molecules were removed.

The percentage of on-target molecules of the restriction digestion method was higher (85%) compared to the size selection method (81.4%) but did not reach 100%. The sequencing data revealed that off-target regions with an uncut recognition sequence or without recognition sequence were still maintained in the final library (Figure 11I; Table 5). The distribution of the reads mapping to the BAC is shown in supplementary Figure 20. This indicates that the majority of reads perfectly fit with the position of the up- and downstream sgRNA's. The total number of mapped reads was lower when the complexity of the genome was reduced by gel excision and restriction digestion since the amount of input material was kept constant for all experiments. Restriction digestion outperformed the gel excision method because it yielded the highest number of on-target reads, the highest percentage of on-target reads and the highest enrichment factor.

Table 5: Overview of the performance of SMRT sequencing of the *FMR1* CGG repeat with and without enrichment.

	Reads			
Enrichment Method	**# Mapped**	**# On-target**	**% on-target**	**Enrichment**
No enrichment (theoretical)	50000	1800	3.6	0
CRISPR-CAS9	47583	32832	69.0	25X
CRISPR-CAS9 + Gel excision	161	131	81.4	29X
CRISPR-CAS9 + restriction digestion	713	606	85.0	30X

The on-target reads contained the *FMR1* CGG repeat and could thus be used to determine the repeat variability in the BAC DNA. The repeat distribution of the BAC molecule was calculated for all 3 strategies and is shown for the restriction digestion method (Figure 14A). The BAC molecules contained mainly 11 CGG units, but some variation between 9 and 14 CGG units was observed. The detected variability mirrors the underlying instability of the BAC grown in E. *Coli* since the analysis was not hampered by any confounding factors like PCR.

This approach also allowed identification of DNA modifications in the target region like m4C and 6mA (Figure 14B). Therefore, the data generated by the restriction digestion method was analyzed by the RS_Modification_and_Motif_Analysis.1 tool with an in silico reference. In total 209 positions carried a m4C modification and 868 carried a m6A. In Figure 14B, an example of a m4C and 6mA modification is shown. Strikingly, SMRT sequencing allowed to distinguish modifications on the positive strand from the negative strand which is exemplified in Figure 14B, where the m4C methylation only occurs at the forward strand while both the forward and the reverse strand carry a 6mA group. This allowed to detect strand-specific methylation patterns (hemimetyhlation). The modification analysis did not reveal any 5mC modifications, nor any modification inside the CGG repeat (data not shown).

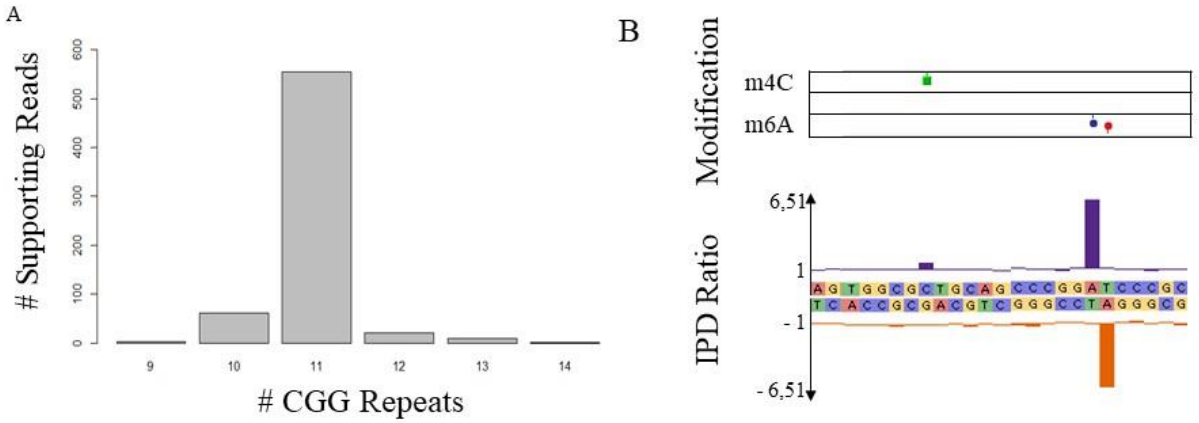


Figure 14: Repeat variability of the *FMR1* CGG repeat of the BAC grown in E. *Coli* (A). An example of a m4C modification on the positive strand and a 6mA modification on both the positive and the negative strand is shown in B.

3.4.2. Enrichment of Human DNA

The enrichment of human molecules was explored with the CRISPR-CAS9 method combined with restriction digestion since this was the best performing method. To monitor the enrichment, human genomic DNA was spiked with *FMR1* containing BAC molecules. Five different mixtures were prepared, sequenced and repeated multiple times: 1/1000 (2X), 1/100 (5X), 1/50 (1X), 1/1 (4X) and only human DNA (1/0; 2X). After long-read sequencing, the number of mapped reads, the number of on-target reads (reads that map to and span the *FMR1* CGG repeat), the percentage of on-target reads and the number of human molecules was determined. Since *FMR1* lengths differ between the BAC and human, the number of human molecules mapped were determined in each experiment.

In order to calculate the enrichment factor achieved in each experiment, the artificial genome size, the theoretical number of 4kb fragments and the theoretical percentage of on-target reads without enrichment needed to be calculated.

For example, for a 1/1000 mixture (Table 6-1):

$$\textbf{Artificial genome size} = \frac{1*\text{haploid human genome}+999*\text{BAC Molecule}}{1000} = \frac{3.2*10^9+999*1.4*10^5}{1000} = 3.35*10^6$$

 # 4kb fragments (the number of 4kb fragments in 1 artificial genome) =

$$\frac{Genome\ Size}{Library\ Size} = \frac{3.35*10^6}{4000} = 838$$

The **theoretical percentage of on-target reads** without enrichment =

$$\%FMR1 = \frac{\#\ CGG\ molecules}{\#\ molecules\ per\ CGG\ molecule} = \frac{1}{838} = 0.1193\%.$$

Enrichment factor: $\frac{\%\ On-target\ Reads\ after\ enrichment\ and\ sequencing}{Theoretical\ \%\ on-target\ reads\ without\ enrichment} = \frac{9.904}{0.1193} = 83$

A summary of the results of the different enrichment experiments is shown in Table 6. Interestingly, in 9 out of 14 experiments human molecules could be detected. In sample 7, 5 human molecules were retrieved from only a single SMRT cell. Another interesting sample is number 12, which yielded 4 human molecules and an enrichment factor of 102. The sample with the highest number of molecules did not correspond with the highest enrichment factor since the latter does not differentiate between BAC and human-derived *FMR1* molecules. The sequencing statistics showed significant variability between different experiments. This indicates that the experimental conditions and the loading of the SMRT cells could still have to be further improved.

Examples of the CGG repeat distribution of n° 1, 7, 12 and 13 are shown in Figure 15. The reads centered around 11 CGG repeats were derived from the BAC molecule whilst the reads containing 20 or 23 CGG units originated from the human DNA. No variability of the *FMR1* CGG repeat was detected in the DNA of the sequenced female. In Table 7, the repeat structure of the detected repeats is shown. Interestingly, except for the repeat size, also AGG units interrupting the CGG repeat can be detected with the developed method.

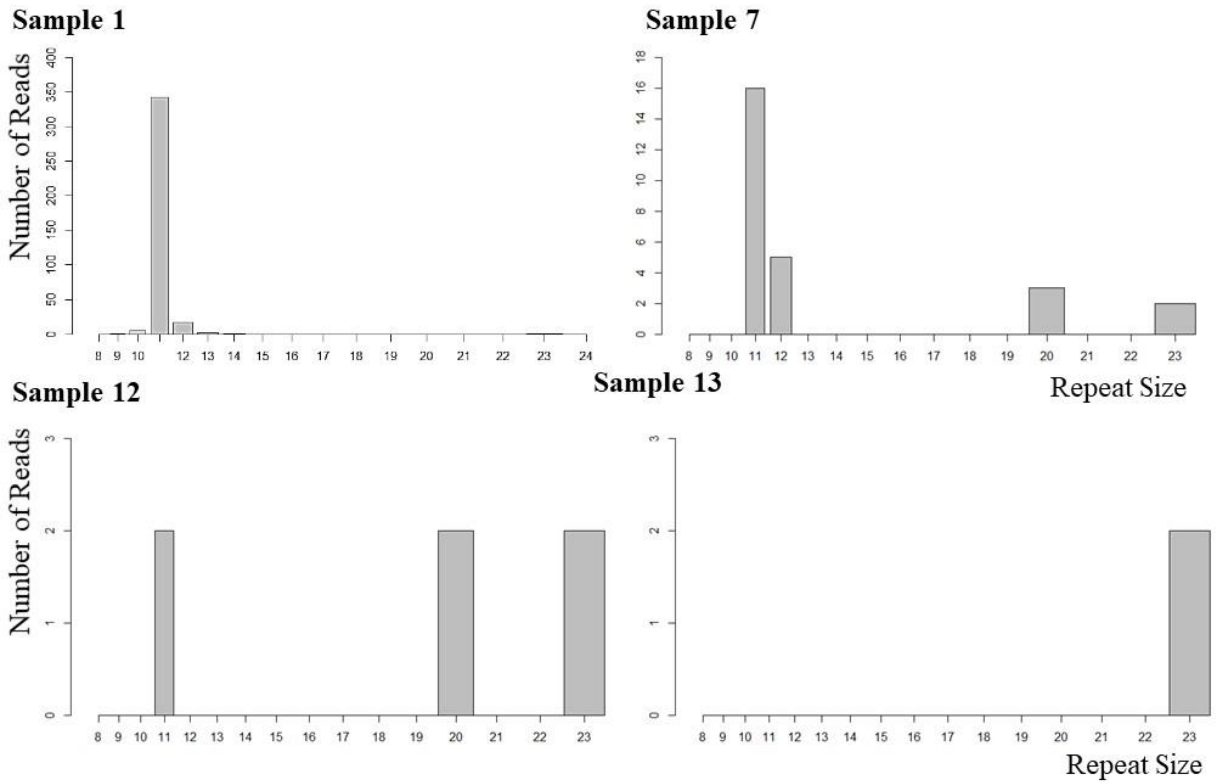


Figure 15: Repeat distrib-ution of sample 1 (Human/BAC ratio: 1/1000), 7 (1/100), 12 (1/1) and 13 (1/0). Repeat sizes centered around 11 CGG units are derived from the BAC molecule while the reads with 20 or 23 CGG units originate from the human DNA.

Table 6: Performance of enrichment on human DNA by CRISPR-CAS9 and complexity reduction by restriction digestion.

	Theoretical values				**Sequencing**				
Number	Human/BAC ratio	Artificial Genome Size	# 4 kb fragments	% *FMR1*	# Mapped Reads	# on-target reads	% on target reads	**Human Molecules**	Enrichment factor
1	1/1000	3.35E+06	838	0.1193%	3766	373	9.904%	**1**	83
2	1/1000				492	0	0.000%	**0**	0
3	1/100	3.22E+07	8059	0.0124%	27940	5	0.018%	**1**	1
4	1/100				5658	2	0.035%	**0**	3
5	1/100				3783	3	0.079%	**1**	6
6	1/100				2515	13	0.517%	**0**	42
7	1/100				23125	26	0.112%	**5**	9
8	1/50	6.31E+07	15766	0.0063%	28863	10	0.035%	**1**	5
9	1/1	1.60E+09	401179	0.0002%	32088	1	0.003%	**1**	13
10	1/1				10233	0	0.000%	**0**	0
11	1/1				33738	1	0.003%	**1**	12
12	1/1				23484	6	0.026%	**4**	102
13	1/0	3.21E+09	802322	0.0001%	35800	2	0.006%	**2**	45
14	1/0				340	0	0.000%	**0**	0

Human/BAC ratio: the ratio of the number of human genomes compared to the number of BAC molecules; Artificial Genome Size: the artificial genome size of the human/BAC mixture; # 4 kb fragments: the number of 4 kb fragments in 1 artificial genome; % *FMR1*: the theoretical percentage of reads that would map to the *FMR1* CGG repeat if a mixture would be sequenced without enrichment; Enrichment factor: indicates the amount of enrichment that is achieved for each experiment. # Mapped Reads: the number of reads that map to the human genome; # On-target reads: the number of reads that map to and span the *FMR1* CGG repeat; % on-target reads: the percentage of mapped reads that map and span the *FMR1* CGG repeat; Human molecules: the number of *FMR1* CGG molecules originating from the human genome.

Table 7: Characteristics of the *FMR1* reads enriched in Sample 12.

# CGG units	**Repeat structure**	**# Reads**
11	(CGG)11	2
20	(CGG)10AGG(CGG)9	2
23	(CGG)13AGG(CGG)9	2

For sample 13 also an epigenetic analysis of the *FMR1* region was performed since this sample had 2 human *FMR1* molecules but no BAC molecules (Figure 16). The forward and the reverse strand had a 12X coverage because the 2 molecules covering the region were each passed 6 passes by the DNA polymerase. Despite the low coverage, the *FMR1* CGG region was still screened for the presence of epigenetic modifications, but this did not reveal the presence of DNA methylations.

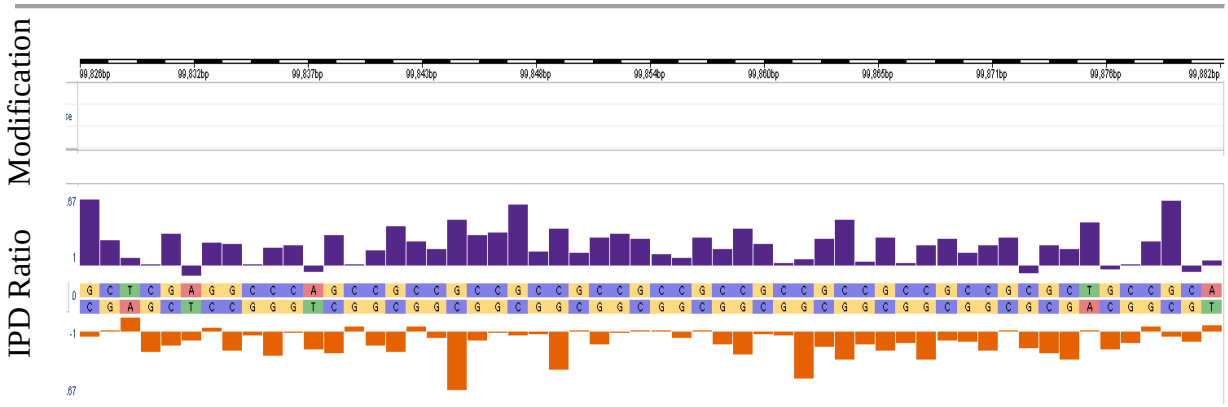


Figure 16: Epigenetic analysis of the *FMR1* CGG region of sample 13. The interpulse duration (IPD) ratio for the positive strand (purple) and the complementary negative strand (orange) are shown (lower panel). Pacbio's kinetic analysis tool could not detect any patterns which could indicate the presence of a DNA modification (upper panel).

Only a small percentage (0 - 9.9%) of the reads map on-target. Thus, it was also interesting to investigate if the remaining reads were mapping towards a few hotspots or rather spread randomly across the genome. The genome coverage of the sequenced samples reveals that the majority of off-target reads map towards the centromeres of the different chromosomes (Figure 17). This is not surprising since the restriction enzymes used to reduce the complexity of the sequencing library did not have any recognition sites inside these heavily repeated regions.

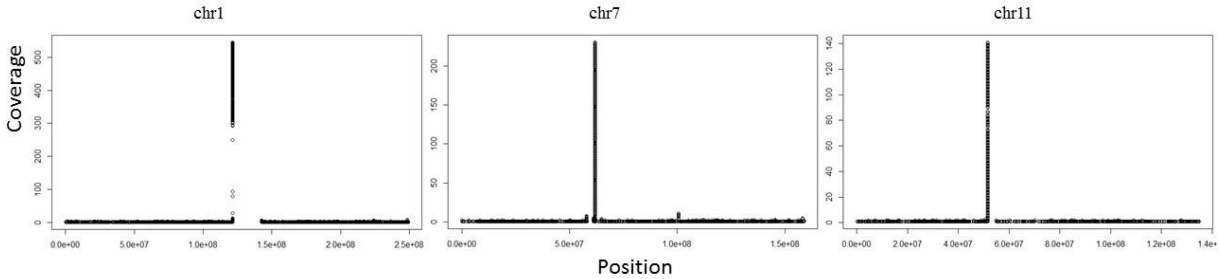


Figure 17: Coverage plot of chromosome 1, 7 and 11 for sample 12. Most off-target reads are derived from the centromere regions.

3.5. Discussion

The technical inability to accurately assess the degree of (epi)genetic STR variation hampers the knowledge on the influence of this variation on the penetrance, complexity and phenotypical variation of STR disorders. SMRT sequencing is a superb technology to analyze both the genetics and the epigenetics of STRs, but unfortunately has only a limited throughput. Thus, it is necessary to apply an enrichment strategy if one wants to study a particular locus. In order to evade the disadvantages inherent to enrichment strategies using amplification, here, a novel, targeted amplification-free methodology to enrich the *FMR1* CGG repeat was developed. This methodology is based on the molecular scissor CRISPR-CAS9 which cuts both up- and downstream of the CGG repeat whereby the repeat is excised. Afterwards different strategies were explored to enrich the target fragment from the genomic DNA.

CRISPR-CAS9 treatment followed by long-read sequencing of the BAC RP11 37p-24 containing the *FMR1* CGG repeat yielded almost 33,000 on-target reads. This very high number of reads was achieved by taking advantage of the tendency of the PacBio instrument to preferentially sequence the target fragment because this is smaller compared to the remaining, larger molecules. This allows to generate a very accurate picture of the *FMR1* CGG repeat variability of the BAC molecule and to identify DNA modifications. This strategy can be used to screen the (epi)genetics of STRs in BAC molecules, small genomes like viral populations, yeasts and mitochondria. However, a complexity reduction of the off-target genomic DNA molecules was necessary when the *FMR1* CGG repeat was enriched from a human genome. Hence, a restriction digestion was performed since this generated a superior result compared to a size selection of the target fragment. Although the methodology still showed some variability in human samples, enrichment factors over 100X were achieved. Interestingly, up to 5 reads with a *FMR1* CGG repeat derived from human DNA could be retrieved from a single SMRT cell. Without enrichment, 80 SMRT cells would have to be run to obtain the same coverage. The excised molecules have a size of 4kb. Hence, an accurate consensus could be constructed since the molecule can be passed multiple times by the sequencing polymerase generating multiple subreads of the same region. Although there are only few on-target reads generated, these reads faithfully reflect the repeat size of the original DNA molecule. Yet, unfortunately the coverage is still too low to determine the true underlying biological variability and the associated layer of DNA modifications of the *FMR1* CGG repeat in female DNA.

Further improvements are necessary to enhance the achieved enrichment factors. One option could be to add an additional restriction enzyme or a sgRNA targeting the centromere regions since coverage analysis shows that a large proportion of reads are derived from the centromere regions. Another improvement could be the use of a CRISPR-CAS9 nickase, which only cuts one strand, could be used instead of a CRISPR-CAS9 nuclease making a double stranded cut. By positioning the nickase on the positive strand 5 bases off from the negative strand, an overhang of 5 bases could be created. Afterwards these overhangs could be targeted by PacBio hairpin adapter carrying a commentary overhang. Since the PacBio adapters could only ligate to the target fragment, theoretically a very high on-target rate could be achieved. In addition, more on-target reads can be achieved by transferring the method from the PacBio RSII to the Sequel instrument that has around 7X more throughput.

An efficient amplification-free enrichment methodology will forge ahead knowledge on FXS. Generating a complete (epi)genetic picture of the *FMR1* CGG repeat will allow to improve genotype/phenotype correlations.

In the future it would be interesting to implement this methodology as a faster and more direct diagnostic tool surpassing current molecular assays (PCR, TP-PCR and Southern blot) used in most clinical laboratories today. Together, this will contribute to a better clinical management of *FMR1* and its associated disorders.

By designing new sgRNA's (and possibly new RE's), this method can not only be expanded to other STR disorders like Huntington's disease or myotonic dystrophy, but also to virtually any genomic region. Finally, this technology could also shed light on more fundamental STR properties like the inherent instability of these DNA elements within or between tissues.

Different amplification-free sequencing methods are currently being developed. For example, ExpansionHunter is an algorithm that detects long repeats from PCR-free whole-genome sequence data in Illumina data (Dolzhenko et al., 2017). Unfortunately, the applicability of the algorithm is limited due to the short reads and the bridge amplification during sequencing. Two other methodologies have been developed to be combined with long-read sequencing so far (Pham et al., 2016; Tsai et al., 2017). One method is developed by Pacific Biosciences and enriches the *FMR1* CGG repeat, but also the C9ORF72 G4C2, HTT CAG repeat and the Sca10 ATTCT repeat (Höijer et al., 2018; Schüle et al., 2017; Tsai et al., 2017). Although both methods produce a high number of on-target reads, they are difficult to transfer between laboratories (personal communication).

In conclusion, we developed a targeted amplification-free enrichment method for the *FMR1* CGG locus. The method is followed by SMRT sequencing and aims to determine the STR variability together with the DNA methylation pattern on a single-nucleotide level in BAC molecules. Even though the method also works on human DNA, more coverage is needed before a complete (epi)genetic overview of the *FMR1* CGG repeat can also be generated for this large genome. We showed the potential of combining targeted, amplification-free enrichment methods with SMRT sequencing which we believe will move the STR field forward in the near future.

3.6. Supplementary Data

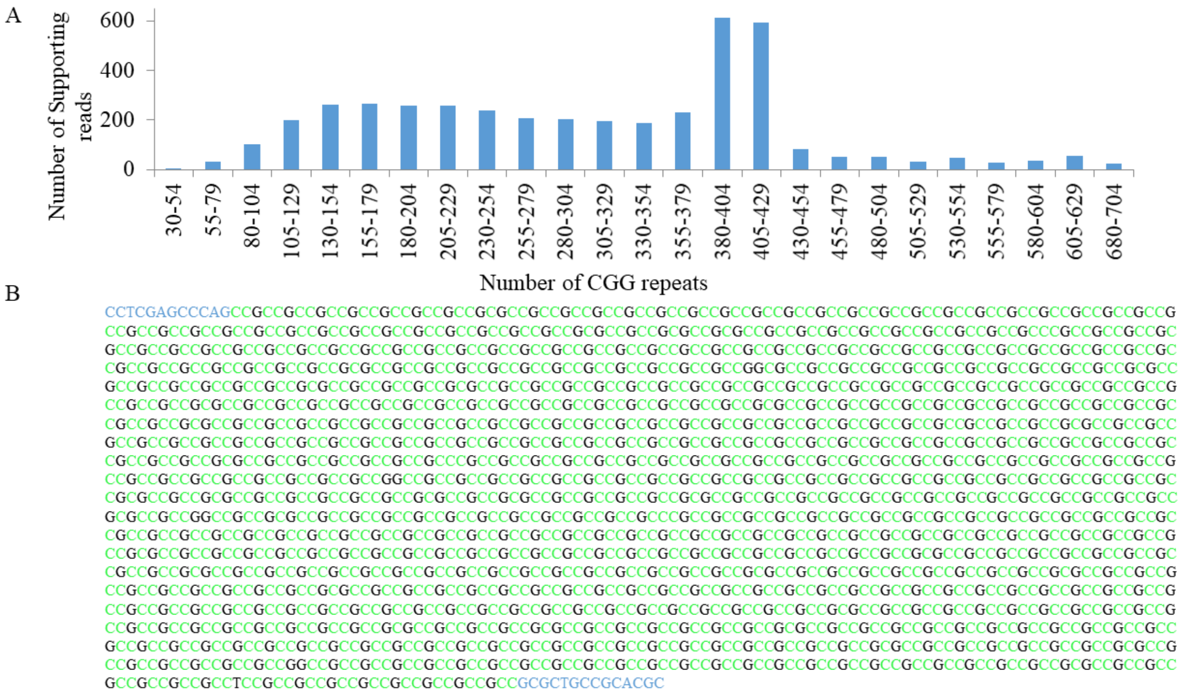


Figure 18: The distribution of the CGG repeat sizes after long-read sequencing of a 600 *FMR1* CGG full mutation is shown (panel A). The distribution is size corrected according to Loomis et al. (2013) because the library was loaded via distribution loading. An example of a read containing a long STR of 680 CGG units and flanking sequence (blue) is shown in panel B.

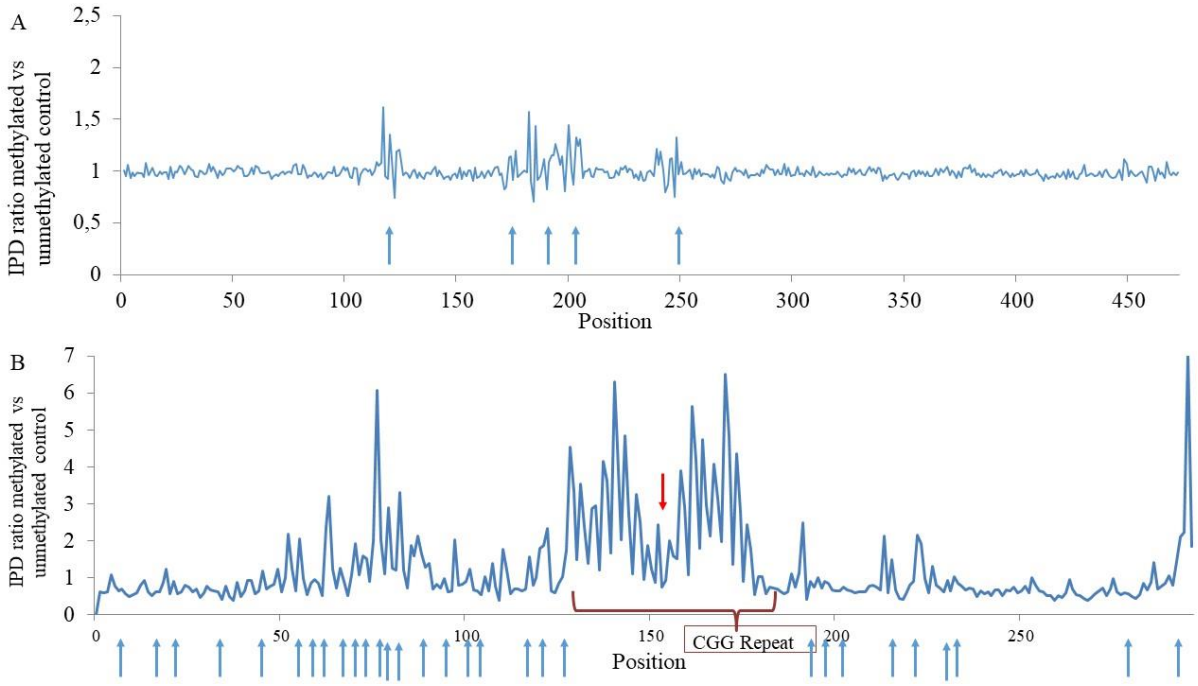


Figure 19: SMRT sequencing not only registers which base is incorporated, but also records the time between the incorporation of different nucleotides, which is called the interpulse duration (IPD). If the template molecule contains a DNA modification like 5mC, the IPD will be delayed compared to non-methylated DNA. Here, an amplicon containing 5 CG dinucleotides (panel A) and the *FMR1* CGG repeat (panel B) were sequenced, both unmethylated and in vitro methylated with M. SssI. Here, the evolution of the ratio of the methylated versus the unmethylated IPD is shown and CG dinucleotides are indicated with an arrow. Note that that the IPD not only changes at the modified nucleotide, but also at the surrounding sequence (panel A). Large perturbation of the IPD are observed at the CGG repeat position (panel B). Interestingly, the IPD dip in the middle of the repeat (red arrow) colocalizes with an AGG unit interrupting the repeat.

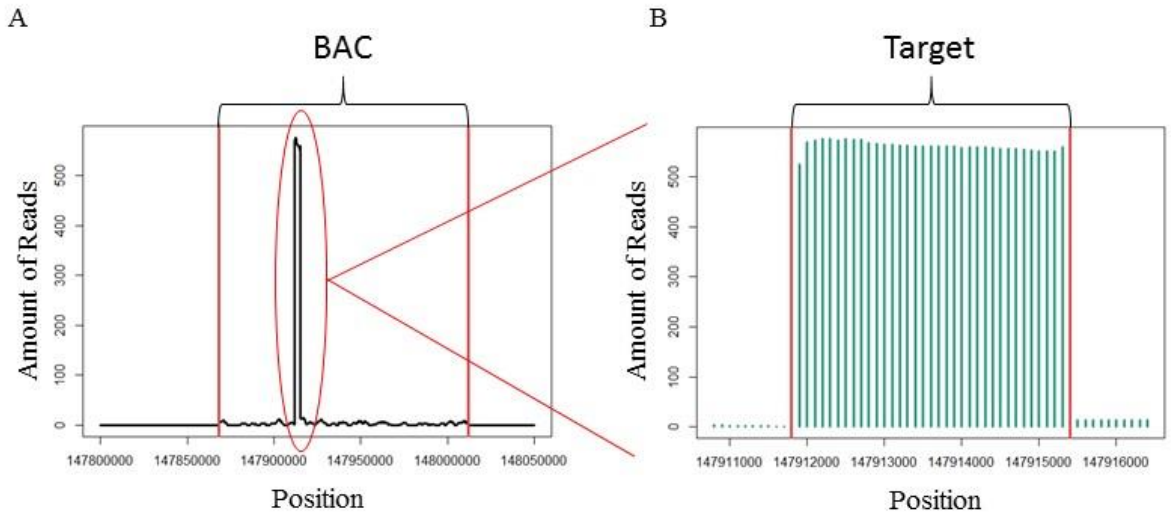


Figure 20: Distribution of the reads after SMRT sequencing of BAC DNA after CRISPR-CAS9 treatment and restriction digestion (A). In B the coverage of the target region is shown.

Chapter 4: Detecting AGG Interruptions in Male and Female FMR1 Premutation Carriers by Long-Read Sequencing

This chapter is a summary of 3 articles:

1. **Ardui, S.** [1]**,** Race, V. [1], Zablotskaya, A. [1], Hestand, M. [1], Van Esch, H. [1], Devriendt, K. [1], Matthijs, G. [1], Vermeesch, J.R.[1] (2017). Detecting AGG Interruptions in Male and Female *FMR1* Premutation Carriers by Single-Molecule Sequencing. Hum. Mutat. 38, 324–331. doi:10.1002/humu.23150.
2. **Ardui, S.** [1]**,** Race, V. [1], de Ravel, T. [1], Van Esch, H. [1], Devriendt, K. [1], Matthijs, G. [1], Vermeesch, J.R. [1] (2018). Detecting AGG Interruptions in Females With a *FMR1* Premutation by Long-Read Single-Molecule Sequencing: A 1 Year Clinical Experience. Front. Genet. 9, 150. doi:10.3389/fgene.2018.00150.
3. **Ardui, S.** [1]**,** Guillen J.J. [2], Tiscornia, G. [2], Vassena, R. [2], Matthijs, G. [2], Vermeesch, J.R[1]. Preliminary study of the influence of AGG interruptions and CGG repeat size on the stability of intermediate *FMR1* alleles (in preparation).

[1] Center for Human Genetics, University Hospitals Leuven, University of Leuven, Leuven, Belgium

[2] Clinica EUGIN, Travessera de les Corts 322, 08029, Barcelona, Spain

4.1. Abstract

The fragile X syndrome is the most frequent cause of inherited X-linked intellectual disability, which arises from a *FMR1* CGG expansion of a premutation [55-200 repeats] to a full mutation allele [> 200 repeats]. The risk for a premutation to expand to a full mutation allele mainly depends on the repeat length and AGG units interrupting this repeat. In genetic counseling it is important to have information on both these parameters to provide an accurate risk estimate to women carrying a premutation allele and having child-wish. For example, in case of a small risk a woman might opt for a natural pregnancy followed up by prenatal diagnosis while she might choose for preimplantation genetic diagnosis if the risk is high. Unfortunately, the detection of AGG interruptions is hampered by technical difficulties complicating the use of AGG interruptions in diagnostics. Therefore, we developed, validated and implemented a new methodology that uses single-molecule sequencing to identify AGG interruptions in males and females with a *FMR1* premutation. Here, we report on the benefits of AGG interruption detection by single-molecule sequencing and the impact of implementing the assay on genetic counseling. In addition, we also show how Single Molecule Real-Time sequencing can be used to assess repeat instability of intermediate alleles.

4.2. Introduction

The 5′ untranslated region (UTR) of the fragile X mental retardation gene (*FMR1*; MIM# 309550) contains a CGG tandem. The repeat is classified into different groups and associated with various disorders depending on the repeat size which can vary from 6 to more than 200 CGG's. Whereas most individuals in the general population have around 30 CGG repeats (<45 repeats), patients with fragile X syndrome (FXS; MIM# 300624) carry large, full expansions with more than 200 repeat units (Oberlé et al., 1991; Verkerk et al., 1991). These large expansions are usually epigenetically silenced thereby inhibiting the production of fragile X mental retardation protein (FMRP)(Pieretti et al., 1991). The absence of protein evokes FXS, a neurodevelopmental disorder characterized by intellectual disability, emotional problems, autism, hyperactivity, hypersensitivity and mild dysmorphic features (Penagarikano et al., 2007). Premutation carriers represent yet another group with repeat sizes varying between 55 and 200 repeats, and might be affected by fragile X-associated tremor/ataxia syndrome (FXTAS; MIM# 300623) or fragile X-associated primary ovarian insufficiency (FXPOI; MIM# 311360), amongst other medical problems (Hagerman and Hagerman, 2015; Sullivan et al., 2005; Van Esch, 2006). In between the premutation alleles (55-200 repeats) and normal alleles (<45 repeats) an intermediate zone (45-54 repeats) exists. Although carriers of intermediate alleles are generally believed to be healthy, some reports have shown that these alleles might be associated with parkinsonism (Loesch et al., 2009) and FXTAS, although with a milder phenotype, less frequently and at a later age-of-onset (Hall et al., 2012; Liu et al., 2013). Furthermore carriers of intermediate CGG alleles are also associated with FXPOI (Bodega et al., 2006; Bretherick et al., 2005) and an increased prevalence of *FMR1* gray zone alleles is present in Parkinson populations (Hall et al., 2011; Loesch et al., 2009; Zhang et al., 2012). Hence, depending on the size of the *FMR1* CGG repeat, an individual will be affected by different disorders varying both mechanistically and phenotypically.

The *FMR1* CGG repeat is susceptible to meiotic instability which is reflected by repeat size differences between parents and offspring. Whereas normal alleles (<45 repeats) are usually inherited stably, around 14% of intermediate alleles display repeat instability upon transmission (Nolin et al., 2015; Sullivan et al., 2005). Since these changes are only small (1 to 5 units), an intermediate allele might expand to a premutation allele, but will not expand into a full mutation during 1 generation (Levesque et al., 2009; Nolin et al., 2011). Nevertheless, it is possible that an intermediate allele evolves into a full mutation in only 2 generations (Fernandez-Carvajal et al., 2009; Terracciano et al., 2004; Zuñiga et al., 2005). With a frequency of 1 in 66, intermediate alleles are common in the general population (Levesque et al., 2009; Tassone et al., 2012b). The degree of instability is determined by the size of the repeat, the number of AGG interruptions and the parent of origin. These AGGs intersperse within the CGG repeat every 9 or 10 CGG repeats at the 5' end where their presence reduces repeat instability (Eichler et al., 1994; Nolin et al., 2015; Yrigollen et al., 2014a). For instance, a large intermediate allele without AGG's will be more unstable compared to smaller alleles with more AGG's. Besides, paternal alleles are more unstable than maternal alleles (Nolin et al., 2015; Sullivan et al., 2002; Yrigollen et al., 2014a). In both male and female premutation carriers, repeat expansions and contractions are common (Nolin et al., 2003b). Moreover, upon maternal transmission, a premutation allele has a risk of expanding into a full mutation.

With a reported frequency of around 1 in 200 females carrying a premutation, a significant fraction of the female population is at risk of having children with FXS (Cronister et al., 2008; Tassone et al., 2012a). This risk depends on the repeat size and the number of AGG triplets interrupting the repeat whereby larger repeats with fewer AGGs have the highest expansion risks (Nolin et al., 2015; Yrigollen et al., 2014a). The influence of AGG interruptions is the most profound for alleles ranging from 60 to 85 repeats. For instance, the risk of transmitting a full mutation for a woman with 75 repeats and 2 AGG triplets is 12%, but this increases to 77% if no AGG interruptions are present (Yrigollen et al., 2012). Some studies have reported that maternal age can also influence the expansion risk (Yrigollen et al., 2014a), but this could not be confirmed by others (Nolin et al., 2015). Hence, further large-scale studies are needed to solve this issue. For small (<60 repeats) or large (>85 repeats) premutation alleles the influence of AGG interruptions on the expansion risk is only minor. Alleles smaller than 60 repeats have only a full mutation expansion risk of 2.6%, even in the absence of AGG repeats while large alleles on the contrary have an expansion risk higher than 60%, even when 2 stabilizing AGGs are present (Yrigollen et al., 2012).

Due to the severity of the FXS, an accurate estimate of the risk that a woman with a premutation will transmit a full mutation to her offspring is crucial in genetic counseling because this influences her reproductive planning. When the expansion risk is high, women might opt for preimplantation genetic diagnosis (PGD) where one could select for unaffected males or noncarrier female embryos (Burlet et al., 2006; Sermon et al., 1999). Although the detection of large CGG alleles could be very challenging in a single cell picked from an early embryo, this can now be circumvented by making use of new haplotyping methods(Dimitriadou et al., 2017; Natesan et al., 2014; Zamani Esteki et al., 2015). If the risk of having a child with FXS is low, a women could choose for normal conception, optionally combined with invasive prenatal diagnosis to screen the fragile X status of their fetus (Biancalana et al., 2015). Although mapping the location and number of AGGs within the CGG repeat is essential for these risk prediction and genetic counseling, AGG measurement is not yet a standard feature of FRM1 diagnostic work-up in most laboratories worldwide (Biancalana et al., 2015; Jacquemont et al., 2011; Monaghan et al., 2013).

Measurement of AGG interruptions and its clinical uptake has been hampered by technical difficulties. The repeat size can easily be determined by PCR-based methods and/or Southern blot, but the detection of AGG interruptions is technically challenging. Traditional Southern blotting cannot localize the AGG interruptions. If determined, AGG interruptions are detected by a triplet-primed PCR (TP-PCR) (Filipovic-Sadic et al., 2010; Hayward and Usdin, 2017). This is an indirect method whereby the forward and reverse primer of a standard PCR are complemented with a third primer that will anneal right into the repeat. By adding the third primer, the PCR will produce a ladder of peaks that will be visible on an agarose gel or electropherogram as a smear. The main advantage of this technique is that it indicates if a full mutation is present in an individual, even if the full mutation cannot be completely amplified. An additional advantage of TP-PCR is that it also points out the presence of AGG units in premutation carriers: if an AGG unit is present in the repeat, the third primer will fail to anneal at that particular site and a gap will be present in the profile. TP-PCR readily identifies the number and location of AGG units within the CGG repeat in males because at every AGG the signal will drop to the baseline. In contrast, interpreting TP-PCR results in females remains challenging as they carry 2 X-chromosomes, each containing a different CGG repeat with a specific set of AGG units.

If the structure of those 2 repeats is different, TP-PCR does not allow to resolve the repeat structure (Chen et al., 2010). Further analysis requires 2 additional PCR reactions to decipher the exact repeat structure (Nolin et al., 2013). A disadvantage for both the diagnostic and scientific fragile X community is that those PCRs are intellectual property of Asuragen (TX) and can only be performed on site.

In order to circumvent the various limitations of TP-PCR, we explored single-molecule sequencing to determine the number and location of AGG interruptions in both males and females carrying *FMR1* premutation alleles. Single Molecule Real-Time (SMRT) sequencing from Pacific Biosciences was chosen since it is very well suited to sequence repeats. Firstly, this technology is able to sequence through large and very GC-rich repeats, including CGG repeats which was demonstrated in chapter 2 and in numerous publications (Chaisson et al., 2015; Loomis et al., 2013; Shin et al., 2013). Secondly, in chapter 2 it was shown that SMRT sequencing can identify AGG interruption in unamplified, human *FMR1* molecules. This finding is supplemented by Loomis et al. (2013) who detected an AGG interruption in a CGG repeat cloned in a plasmid. Finally, the long reads (>20 kb) generated by single-molecule sequencing allow to cross a circulated double-stranded template molecule multiple times. By making a consensus from all different passes, it is possible to eliminate sequencing errors that are randomly distributed across the reads and generate very accurate reads-of-insert (Carneiro et al., 2012; Hestand et al., 2016a; Jiao et al., 2013).

In chapter 2 we developed an amplification-free enrichment method which yields *FMR1* reads wherein AGG interruptions can be detected. However, this methodology requires a large DNA input, is not high-throughput and generates only little reads. Since in this chapter the aim is to develop an approach which yields a high coverage and can be applied on many patients, the approach developed in chapter 2 is not suited. Therefore, in this chapter we develop an amplicon-based strategy in combination with SMRT sequencing and we demonstrate that this enables the reconstruction of the complete repeat structure for gray zone and premutation alleles, not only for males, but also for females. Secondly, we implement AGG interruption detection by single-molecule sequencing for diagnostic use for all female carriers with an intermediate and small premutation alleles (45-100 repeats). We summarize our experience with the use of AGG triplets in the clinic after 1 year during which we analyzed 51 patients. This showed that using both repeat size and the number of AGG's in the genetic laboratory improved the risk assessments which positively impacted the management of the disorder. Thirdly, a preliminary study is presented where the influence of repeat size, number of AGG units and paternal versus maternal inheritance on the stability of 12 intermediate alleles during transmission was investigated. This data complements to the data on intermediate allele instability in literature that is still limited since most studies so far focused particularly on the instability of premutation and full mutation alleles.

4.3. Materials and Methods

4.3.1. DNA samples

For the validation of the AGG interruption detection by SMRT sequencing, the structure of the CGG repeat in the *FMR1* gene (Genebank Accession number NG 007529.2) was determined for 7 males (2 gray zone alleles, 5 premutation alleles) and 33 females (5 females with a normal and a gray zone allele and 28 females with a normal and a premutation allele).

After the uptake of the method for diagnostics, 51 female patients were ascertained from January to December 2017 at the Center for Human Genetics, KU Leuven, UZ Leuven, Belgium. Fifty patients carried a normal allele and an intermediate (26) or premutation allele (24). In addition, one female carrying 2 premutations was also included in this study. The patients were referred for diagnostic testing because of either FXTAS, POI, or because of a family history of FXS. The premutation alleles varied between 45 and 100 CGG units according to PCR.

The study was approved by the local Ethics Committee and consent was obtained from the patients, both informed and written.

To study intermediate allele instability, families carrying alleles with 45-54 CGG units were collected through the Eugin fertility clinic located in Barcelona (Spain). These carriers were recruited from women who presented as candidates for an oocyte donation program running in the clinic between September 2016 and December 2017. At least one parent had to be available, siblings of the proband were included only if they were available. After applying the inclusion criteria, 12 women with an intermediate allele were retained. From this group 9 (including 2 pairs of sisters) had both their father and mother available while for 3 women only the parent carrying the intermediate allele was available for testing. Hence, in total 29 individuals were included for the preliminary study of intermediate allele instability. The study was approved by Eugin's ethical committee (CEIC EUGIN) and consent was obtained from the patients, both informed and written.

4.3.2. DNA Extraction

For all participants DNA was isolated from peripheral white blood cells according to standard procedures. From one pregnant female patient a chorionic villi sample (CVS) sample was received. Villi from CVS were separated from maternal tissue under a microscope to minimize maternal contamination. Two to 4 villi were provided for DNA extraction

4.3.3. Amplicon Generation

First, the *FMR1* CGG repeat was amplified using the PCRX Enhancer System (Invitrogen, Carlsbad, CA) with 40 ng/µl DNA input and previously published specific primers (Figure 21A) (Filipovic-Sadic et al., 2010). In order to integrate barcodes, an M13 tail was attached at both the forward (M13-forward tail: TGTAAAACGACCCAGGGT) and the reverse primer (M13-reverse tail: CAAAGGACAGCTATGACC).

Next, a reaction mixture was prepared by combining 4 µl of 10X PCRX amplification buffer, 1.2 µl 50 mM MgSO4, 4 µl of 2mM dNTPs (Invitrogen), 16 µl of 10X PCR-X enhancer solution, 4 µl of a 2.5 µM mixture of both forward and reverse primer, 10 µl of DNA, and finally 0.5 µl Taq polymerase (Invitrogen). After gently mixing the reaction, the repeat was amplified starting with a heat denaturation at 95°C for 3 min, followed by 25 cycles of 95°C for 30 sec, 64°C for 60 sec, and 68°Cfor 90 sec, followed by a final extension step at 72°C for 5 min, whereafter the samples were stored at 4°C. After checking the efficiency of the PCR on a Fragment Analyser (Advanced Analytical, Ankeny, IA), the samples were purified with 1X washed Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) and eluted in 11 µl of water. Next, barcoded primers from PacBio's 96-well barcoding kit were attached to the amplicons by their M13 tail, which allowed pooling up to 24 different amplicons together. The reaction mixture was prepared as described above, but now the purified amplicons were used as DNA input together with PacBio's barcoded primers. The second PCR mixture was subsequently denatured at 95°C for 180 sec followed by 5 cycles of 95°C for 30 sec, 45°C for 60 sec, and 68°Cfor 90 sec, followed by another 5 cycles of 95°C for 30 sec, 65°Cfor 60 sec, and 68°C for 90 sec and a final elongation step at 72°Cfor 5 min. Afterwards, the amplicons were again purified by 1X washed Agencourt AMPure XP beads, visualized on the Fragment Analyser and pooled equimolar together.

4.3.4. Single-Molecule Real-Time Sequencing

The pooled amplicons were prepared for sequencing as described in PacBio's standard 500 bp Template Preparation and Sequencing protocol, using the Template Prep kit 3.0 (Pacific Biosciences, Menlo Park, CA). Hereafter, each library was sequenced on a PacBio RS II using the DNA/polymerase binding Kit P6 v2 (Pacific Biosciences) for a 360-min movie (Figure 21B). All runs used PacBio's DNA Sequencing Reagent Kit 4.0 v2.

4.3.5. De Novo Assembly of the CGG Repeat Structure

Generating reads-of-insert

The long reads generated by single-molecule cross each molecule multiple times (Figure 21C). Therefore, demultiplexed reads-of-insert were generated with the RS ReadsOfInsert.1 protocol from PacBio's SMRT portal (v2.3.0) with a minimum of 10 full passes, a minimum predicted accuracy of 90% and demultiplexing with symmetric barcodes (Figure 21D).

Selecting on-target reads

Next, only reads-of-insert derived from the *FMR1* CGG repeat were selected by aligning all reads-of-insert using BWA-SW v0.7.10 (Li and Durbin, 2009) against the human reference genome hg19 downloaded from UCSC (Karolchik et al., 2004), followed by conversion of SAM to BAM by Samtools v1.3.1 (Li et al., 2009). To finally turn the BAM files into BED format and select the on-target reads-of- inserts, BEDtools v2.20.1 was used (Quinlan and Hall, 2010).

Separation of the 2 alleles in females

In females, we separated the normal from the premutation allele by plotting the number of reads as a function of read size followed by separation of the normal from the premutation allele based on differences of the read size (Figure 21E). This is possible because normal alleles contain less CGG repeat units than premutation alleles and thus generate shorter reads-of-inserts. Reads-of-insert derived from the normal allele are called normal reads and reads-of-insert originating from the premutation alleles are called premutation reads.

De novo assembly

Subsequently, a *de novo* assembly (Figure 21F) was performed on the separated normal and premutation reads-of-insert using MIRA v4.0 (Chevreux et al., 2004). This specific assembler was chosen because it was conceived especially to resolve repeats, and it has been used before to perform *de novo* assembly on single-molecule sequencing data of large repetitive regions (Guo et al., 2014a). To perform an assembly on the normal reads, MIRA was run with custom-tuned parameters that can be found in the Supplementary Methods. Afterwards, only assemblies with the highest quality were selected. Ideally, this means the quality per base is 90.

A custom python script was used to extract the repeat size, the number of AGG interruptions and the repeat structure from the assembly (Figure 21G). To control the assembly process, the repeat structure extracted from the *de novo* assembly was compared with the repeat characteristics (repeat size, AGG interruptions) of the individual reads-of-insert. All variants are submitted to the *FMR1* locus- specific database that can be found at http://www.lovd.nl/*FMR1* (Fokkema et al., 2011).

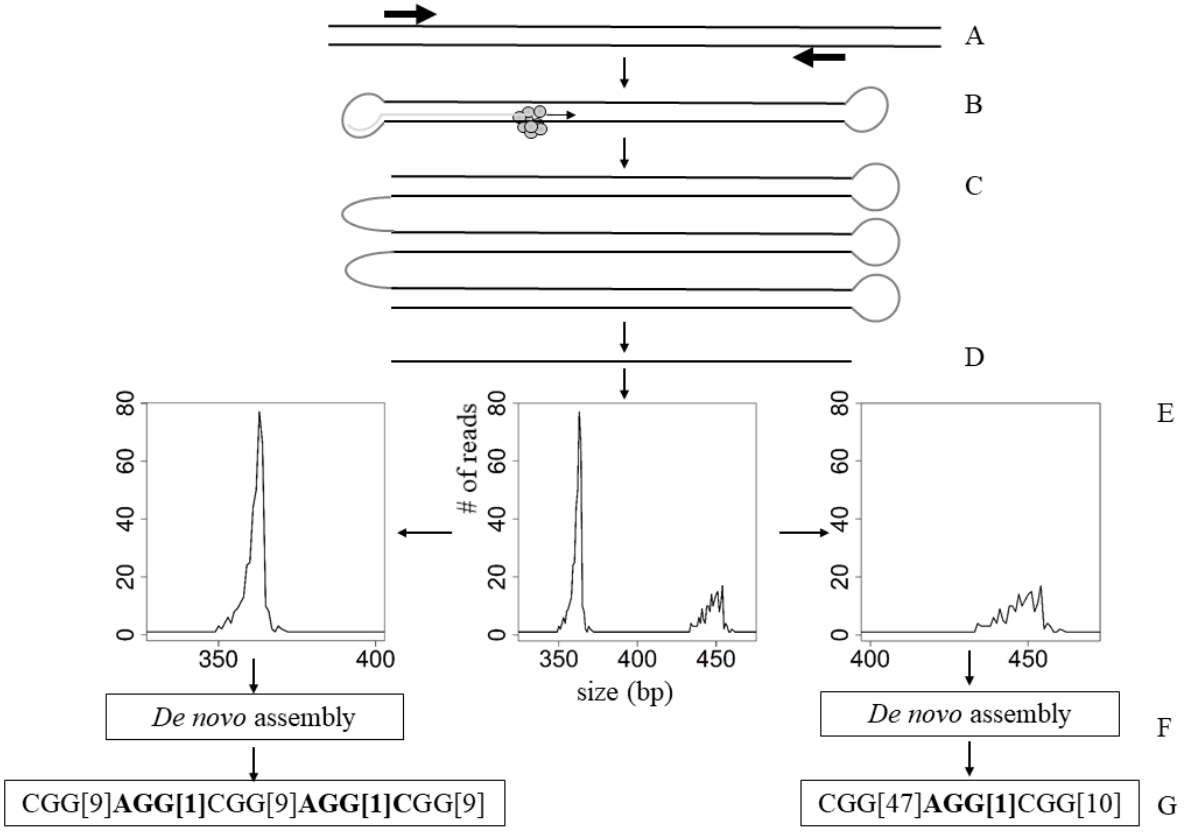


Figure 21: Overview of the workflow. First, PCR amplicons were generated whereby a different barcode is introduced for each amplicon (A). Next, different amplicons were pooled together and sequenced by PacBio single-molecule sequencing (B). The long reads generated by single-molecule sequencing allowed to cross a circulated double-stranded template molecule multiple times (C) from which reads-of-insert with a high quality are generated (D). After selecting the on-target reads, the distribution of the read sizes of those reads-of-insert was plotted (E), followed by separating reads-of-insert belonging to the normal allele from the premutation allele based on differences in read size (E). Finally, a *de novo* assembly was performed on the separated normal and premutation reads (F) from which the repeat structure was extracted (G).

4.3.6. Determination of the Precision and Robustness of AGG Interruption Detection

In order to describe the precision of the described AGG interruption detection method, we used the terminology proposed by Mattocks et al. (2010). Three females were randomly selected to determine the repeatability (within-run precision) and the intermediate precision (between-run precision). To determine the repeatability, 3 amplicons of each selected female were included in a single run. Next, to define the intermediate precision, amplicons of 3 females were included in 3 separate sequencing runs. Finally, the robustness of the method was tested by using 5, 40, and 100 ng/µl of input DNA for 1 female.

4.3.7. Validation of the Sequencing Results

The size and structure of the repeat determined by single-molecule sequencing was validated by an "in-house" PCR combined with an *FMR1* TP-PCR assay (Abbott, IL), carried out following the manufacturer's instructions. AGG information was extracted from the TP-PCR result if the electropherogram generated an interpretable result.

4.3.8. Genetic Counseling

Females with a premutation were offered genetic counseling. If the patient was considering having children, an accurate assessment of the risk that their premutation would expand to a full mutation was provided based on both the *FMR1* CGG repeat size and the number of AGG interruptions.

4.4. Results

4.4.1. Validation of AGG Detection by SMRT Sequencing

FMR1 CGG repeat structure determination

To determine the *FMR1* repeat structure we performed single-molecule sequencing of PCR amplicons from 33 different females and 7 males. Reads-of-insert were generated with a minimum of 10 full passes and a mean of 25 full passes. This ensured a high accuracy of the final reads-of-insert (Supplementary Figure 25). The CGG repeat of the 7 males was supported by a mean coverage of 261 [84-614] reads (Supplementary Figure 26A). In females, the premutation allele contains more CGG units than the normal allele and thus amplifies worse during PCR. Consequently, the premutation was covered by less reads. The female samples (1-33) had a mean coverage of 277 [83-458] and 158 [22-332] for the normal and premutation alleles respectively (Supplementary Figure 26B-C).

After sequencing, a *de novo* assembly was generated for the CGG repeats of the 7 males and the sizes of those assemblies were compared to the results from PCR. All assembled repeat sizes determined by single molecule sequencing fitted within the error range of PCR control

runs (± 1 repeat unit up to 54 units, ± 3 repeat units up until 80 CGG units, ±10% of repeat size starting from 80 repeats; Figure 22A).

Subsequently, the repeat structure was investigated (Table 8). For all male samples the number and position of AGG units observed by single molecule sequencing was 100% concordant with TP-PCR (Supplementary Figure 27A).

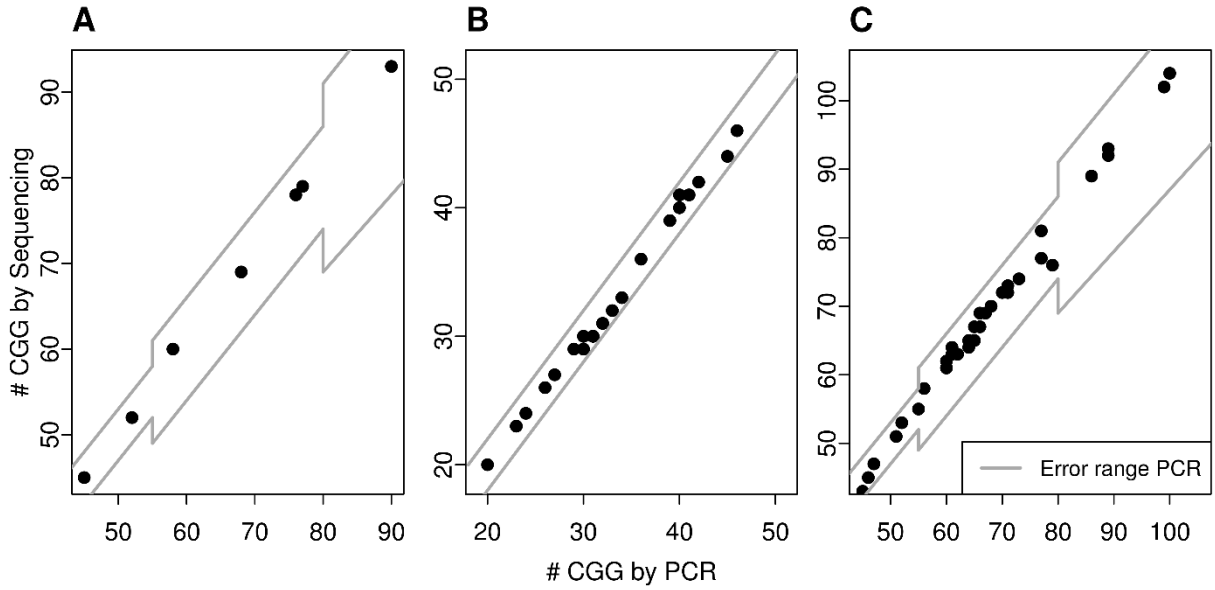


Figure 22: Correlation of the repeat size between single-molecule sequencing (Y axis) and PCR (X-axis) for: 7 male samples (A); the normal allele of all female samples (B); the gray-zone/premutation allele of all female samples (C). Samples are indicated by a dot. Gray lines show the error range of PCR that is ±1 repeat unit for repeats smaller than 54 repeat units, ±3 repeat units up until 80 CGG units, and ±10% of the repeat.

Table 8: Repeat characteristics for all male individuals.

	(TP)-PCR	SINGLE-MOLECULE SEQUENCING		
Male	**# Units**	**# Units**	**# AGG**	**Repeat structure***
1	45	45	0	CGG[45]
2	52	52	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[32]
3	58	60	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[40]
4	68	69	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[49]
5	76	78	1	CGG[9]AGG[1]CGG[68]
6	77	79	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[59]
7	90	93	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[73]

* Genomic DNA change relative to hg19/GRCh37 at g.146993570 (chrX)

Next, we investigated whether the repeat size and structure could also be determined for females. For 30 females the difference in repeat size between their 2 alleles was larger than 10 repeat units as previously determined by PCR. For this group we first separated the normal from the premutation allele based on differences in read length and subsequently performed a *de novo* assembly on each group of reads separately (Table 9). All assembled normal and premutation repeat sizes fell within the error range of PCR (Figure 22B-C). The repeat structure of single-molecule sequencing was validated by TP-PCR for 6 females with a similar repeat structure on each of their X chromosomes (Supplementary Figure 27B).

For 3 females (Table 10; female 31-33) the difference in repeat size between their 2 alleles was smaller than 10 repeats. Although it was still possible to recognize the presence of 2 different alleles in the size distribution of the reads-of-insert, it became difficult to separate the reads derived from the different alleles based on this characteristic (Figure 23A). Therefore, a *de novo* assembly was performed on the mixture of reads. To make sure both alleles were generated by the assembler, the distribution of the number of AGG interruptions as function of the repeat sizes of the individual reads-of-insert was also mapped (Figure 23B). In this figure all differences in repeat size and the number of AGG units are visualized, which allows to identify the most frequently occurring repeat structures representing the 2 female alleles. In figure 23B also smaller circles are present, flanking the most frequently occurring repeat sequences that represent stutter products inherent to PCR amplification of repeat rich regions. We also tested PacBio's Long Amplicon Analysis tool, but this performed overall worse than the MIRA assembly pipeline (see Supplemental Results and Supplemental Figure 28).

Table 10: Repeat characteristics for 4 females with similarly sized repeats.

	(TP)-PCR		SINGLE-MOLECULE SEQUENCING					
	ALLELE 1	ALLELE 2	ALLELE 1			ALLELE 2		
N°	**# Units**	**# Units**	**# Units**	**# AGG**	**Repeat structure***	**# Units**	**# AGG**	**Repeat structure***
31	41	46	41	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[20]	45	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[25]
32	42	47	42	1	CGG[20]AGG[1]CGG[21]	47	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[27]
33	40	45	41	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[21]	43	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[23]

* Genomic DNA change relative to hg19/GRCh37 at g.146993570 (chrX)

Table 9: Repeat characteristics for all females with a difference between normal and premutation allele >10 repeat units.

	(TP)-PCR		SINGLE-MOLECULE SEQUENCING					
	NORMAL	PREMUTATION	NORMAL			PREMUTATION		
N°	**# Units**	**# Units**	**# Units**	**# AGG**	**Repeat structure***	**# Units**	**# AGG**	**Repeat structure***
1	29	56	29	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[9]	58	1	CGG[10]AGG[1]CGG[47]
2	33	51	32	2	CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]	51	4	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]AGG[1]CGG[9]AGG[1]CGG[10]
3	40	68	40	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[20]	70	1	CGG[9]AGG[1]CGG[60]
4	30	65	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	67	0	CGG[67]
5	30	71	29	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[9]	73	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[53]
6	32	68	31	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[10]	70	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[50]
7	31	71	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	72	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[52]
8	26	89	26	1	CGG[9]AGG[1]CGG[16]	92	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[72]
9	32	60	31	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[10]	61	1	CGG[11]AGG[1]CGG[49]
10	31	61	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	64	1	CGG[9]AGG[1]CGG[54]
11	30	86	29	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[9]	89	1	CGG[9]AGG[1]CGG[79]
12	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[37]
13	31	79	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	76	0	CGG[76]
14	30	89	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	93	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[73]
15	31	70	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	72	1	CGG[9]AGG[1]CGG[62]
16	23	67	23	1	CGG[13]AGG[1]CGG[9]	69	1	CGG[9]AGG[1]CGG[59]
17	36	99	36	1	CGG[10]AGG[1]CGG[25]	102	1	CGG[9]AGG[1]CGG[92]
18	31	66	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	69	0	CGG[69]
19	24	61	24	1	CGG[14]AGG[1]CGG[9]	63	1	CGG[9]AGG[1]CGG[53]
20	32	77	31	2	CGG[9]AGG[1]CGG[11]AGG[1]CGG[9]	77	1	CGG[9]AGG[1]CGG[67]
21	32	100	31	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[10]	104	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[86]
22	31	64	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	65	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[45]
23	27	52	27	1	CGG[9]AGG[1]CGG[17]	53	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[33]
24	31	66	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	67	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[47]
25	30	62	29	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[9]	63	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[43]
26	23	60	23	1	CGG[12]AGG[1]CGG[10]	62	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[42]
27	34	77	33	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[13]	81	1	CGG[9]AGG[1]CGG[71]
28	39	64	39	0	CGG[39]	65	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[45]
29	30	60	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	61	1	CGG[9]AGG[1]CGG[51]
30	20	64	20	1	CGG[10]AGG[1]CGG[9]	64	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[44]

* Genomic DNA change relative to hg19/GRCh37 at g.146993570 (chrX)

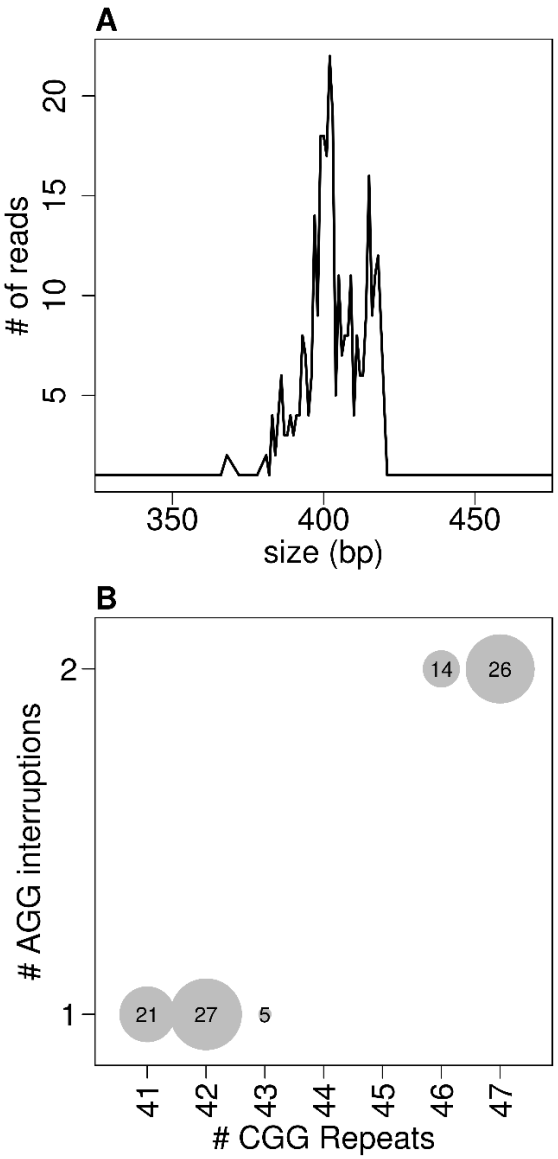


Figure 23. Distribution of the read sizes of the reads-of insert for female 32 (A). The 2 alleles only differ by 5 repeat units and hence are difficult to separate based on differences in the size of the reads-of- insert. (B) Relation between the number of CGG units (X axis) and AGG interruptions (Y axis) for the individual reads-of-inserts of female 32. The surface of the circles is relative to the number of supporting reads. Some minor circles with the same amount of AGG units but a different repeat size can also be observed and represent stutter products

Precision and Robustness of AGG Interruption Detection

The precision of AGG interruption detection was evaluated by investigating the repeatability and the intermediate precision by sequencing the PCR product of 3 females 3 times within the same sequencing run and spread over independent sequencing experiments. The number and position of AGG units in both the normal and premutation allele were always reproduced (Table 11). Except for female 17, the size of the repeats was also fully reproducible. In female 17, a difference of 1 and 2 CGG units for 2 within-run repetitions was found. This small variation is caused by the low coverage of the premutation allele of those 2 samples (7 and 19 reads, respectively; Supplemental Figure 26), and the presence of more stutter in this sample due to the large repeat size (99 units). Finally, varying the input DNA concentration before PCR did not influence the result (Table 11; female 12). Thus, single- molecule sequencing generates results with a high precision and robustness.

Table 11: Repeat characteristics for 3 females repeated both within and between different sequencing runs and with different DNA concentrations.

			(TP)-PCR		SINGLE-MOLECULE SEQUENCING					
			NORMAL	PREMUTATION	NORMAL			PREMUTATION		
N°	**Run**	**Input (ng/ul)**	**# Units**	**# Units**	**# Units**	**# AGG**	**Repeat structure***	**# Units**	**# AGG**	**Repeat structure***
2	1	40	33	51	32	2	CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]	51	4	CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]AGG[1]CGG[9]AGG[1]CGG[10]
	2	40	33	51	32	2	CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]	51	4	CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]AGG[1]CGG[9]AGG[1]CGG[10]
	2	40	33	51	32	2	CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]	51	4	CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]AGG[1]CGG[9]AGG[1]CGG[10]
	2	40	33	51	32	2	CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]	51	4	CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]AGG[1]CGG[9]AGG[1]CGG[10]
	3	40	33	51	32	2	CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]	51	4	CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]AGG[1]CGG[9]AGG[1]CGG[10]
12	1	40	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[37]AGG[1]CGG[7]AGG[1]CGG[9]
	2	40	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[37]
	2	40	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[37]
	2	40	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[37]
	3	40	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[37]
	3	5	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[37]
	3	100	31	55	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	55	2	CGG[9]AGG[1]CGG[7]AGG[1]CGG[37]
17	1	40	36	99	36	1	CGG[10]AGG[1]CGG[25]	104	1	CGG[9]AGG[1]CGG[94]
	2	40	36	99	36	1	CGG[10]AGG[1]CGG[25]	102	1	CGG[9]AGG[1]CGG[92]
	2	40	36	99	36	1	CGG[10]AGG[1]CGG[25]	103	1	CGG[9]AGG[1]CGG[93]
	2	40	36	99	36	1	CGG[10]AGG[1]CGG[25]	104	1	CGG[9]AGG[1]CGG[94]
	3	40	36	99	36	1	CGG[10]AGG[1]CGG[25]	104	1	CGG[9]AGG[1]CGG[94]

* Genomic DNA change relative to hg19/GRCh37 at g.146993570 (chrX)

4.4.2. Clinical Experience with AGG Interruption Detection

AGG analysis by SMRT sequencing was implemented diagnostically in order to more accurately assess the risk that offspring of premutation carriers will be affected by FXS and thus, to improve genetic counseling. We report the results of AGG interruption detection in 51 females with intermediate or premutation alleles.

The results of the *FMR1* CGG repeat analysis are summarized in Table 12. 50 females carried a normal allele and an intermediate (26) or a premutation allele (24), while one female carried 2 premutation alleles. Therefore, the total number of premutation alleles is also 26 (Table 12). The normal alleles of all 50 females ranged between 20 and 40 repeats and are interspersed with 0, 1, 2 or 3 AGGs. Two different clusters were identified within the structures of the normal alleles: a smaller group [20-24 repeats] interrupted by 1 AGG (11 patients) and a larger group [30-34 repeats] interrupted by 2 AGGs (29 patients), in line with previously published results (Chen et al., 2003; Eichler et al., 1996). The remaining 10 normal alleles are more distributed in size and number of AGGs (Table 12)(Eichler et al., 1996). From the 26 intermediate alleles and 26 premutation alleles (from 25 females), the majority (45) were interrupted by 1 or 2 AGG interruptions.

Information on AGG interruptions is vital for females carrying a premutation allele because of the risk of transmitting a full mutation to their children. For the 13 females with a premutation allele ranging between 60 and 84 repeats, knowing the AGG status was reassuring for these carrying 2 AGGs, while it alerted females without any AGG triplets. For example, female 1 (Table 13) presented with an allele of 69 CGG repeats without interruptions. This female opted for PGD because she has a relatively high expansion risk (23%) in combination with a reduced fertility. Female 2 (Table 13) carried 2 premutation alleles (65 and 73 repeats) and chose for PGD because she also carried a translocation. For this female, AGG interruption analysis was performed to determine which allele had the lowest expansion risk. As sequencing revealed that 2 AGGs were present in each allele, embryos carrying the smallest allele of 65 repeats with 2 AGG's could be prioritized for transfer during PGD. Female 3 (Table 13) carried a premutation allele of 69 repeats interrupted with 2 AGG triplets. This female only has a low expansion risk of 0.5% and hence decided to choose for a natural pregnancy. Invasive prenatal follow-up of the ensuing pregnancy showed that the fetus inherited the normal copy. Two females carried a repeat above 85 repeats and hence had a high expansion risk: 100% for the female with 95 repeats without AGGs and 60% for the female with 89 repeats and 2 AGGs (female 4, Table 13). Despite the high expansion risk of female 4, she conceived naturally. Fortunately, invasive prenatal follow-up showed the premutation allele even contracted in the female fetus (female 5, Table 13), most likely to the allele containing 66 repeats and 2 AGG's. A contraction to the allele with 44 repeats and 1 AGG is less likely but cannot be excluded because the DNA of the father was not available.

Sequencing of *FMR1* CGG repeats not only identified AGG triplets, but any sequence variation at this locus. The *FMR1* structure is most commonly built up from CGG[9]AGG and CGG[10]AGG building blocks. This was confirmed in most of the females of our cohort. However, 4 females (female 6 to 9, Table 13) carried a less common AGG interruption pattern. Female 10 carries an intermediate allele with 45 repeats and 2 AGG triplets, but also harbored a CTG interruption (Table 13).

Table 12: Overview of the repeat size, the number of interrupting AGG units and the corresponding expansion risk for 51 females included in fragile-X diagnostics.

	# AGG units				
Repeat Range	**0**	**1**	**2**	**3**	**# alleles**
Normal					**50**
20-24	1	11	0	0	12
25-29	1	0	3	0	4
30-34	0	2	29	0	31
35-39	0	0	2	0	2
40-44	0	0	0	1	1
Intermediate					**26**
45-49	1	3	14	1	19
50-54	0	4	3	0	7
Premutation					**26**
55-59	1 (1.6%)	3 (0.5%)	6 (0%)	0 (0%)	10
60-64	**0 (6.4%)**	**1 (2.2%)**	**0 (0.1%)**	**0 (0.1%)**	**1**
65-69	**2 (22.9%)**	**0 (9%)**	**3 (0.5%)**	**0 (0.5%)**	**5**
70-74	**1 (56.3%)**	**0 (30%)**	**3 (2.1%)**	**0 (2.1%)**	**4**
75-79	**0 (84.4%)**	**0 (65%)**	**2 (8.5%)**	**0 (8.5%)**	**2**
80-84	**0 (96%)**	**0 (89%)**	**2 (28.7%)**	**0 (28.7%)**	**2**
85-89	0 (99.1%)	0 (97.2%)	1 (63.3%)	0 (63.3%)	1
90-94	0 (99.8%)	0 (99.3%)	0 (88.2%)	0 (88.3%)	0
95-100	1 (99.9)	0 (99.8%)	0 (97%)	0 (97%)	1
# alleles	8	24	68	2	

Fifty females carry a normal allele (<45 repeats) and a grey-zone (45-54 repeats) or a premutation allele (55-200 repeats) and one female carries 2 premutation alleles. The repeat range where the number of AGG units has a major impact on the expansion risk is depicted in bold.

Table 13: Overview of the *FMR1* CGG repeat structure for patient 1-10.

			NORMAL			PREMUTATION
Female	**# Units**	**# AGG**	**Repeat Structure***	**# Units**	**# AGG**	**Repeat Structure***
1	22	1	CGG[9]AGG[1]CGG[12]	69	0	CGG[69]
2	65	2	CGG[9]AGG[1]CGG[9] AGG[1]CGG[45]	73	2	CGG[9]AGG[1]CGG[7] AGG[1]CGG[55]
3	33	2	CGG[9]AGG[1]CGG[9] AGG[1]CGG[13]	69	2	CGG[9]AGG[1]CGG[9] AGG[1]CGG[49]
4	30	2	CGG[10]AGG[1]CGG[9] AGG[1]CGG[9]	89	2	CGG[9]AGG[1]CGG[9] AGG[1]CGG[69]
5	44	1	CGG[9]AGG[1]CGG[34]	66	2	CGG[9]AGG[1]CGG[9] AGG[1]CGG[46]
6	32	2	CGG[9]AGG[1]**CGG[12**] AGG[1]CGG[9]	45	2	CGG[9]AGG[1]CGG[9] AGG[1]CGG[25]
7	34	2	CGG[9]AGG[1]**CGG[14**] AGG[1]CGG[9]	49	1	CGG[9]AGG[1]CGG[39]
8	30	2	CGG[10]AGG[1]CGG[9] AGG[1]CGG[9]	49	3	CGG[10]AGG[1]CGG[9]AGG[1] **CGG[18**]AGG[1CGG[9]
9	30	1	**CGG[20**]AGG[1]CGG[9]	69	2	CGG[9]AGG[1]CGG[7]AGG[1] CGG[51
10	40	3	CGG[10]AGG[1CGG[9]AGG[1] CGG[9]AGG[1]CGG[9]	45	2	CGG[7]**CTG**CGG[1]AGG[1] CGG[9]AGG[1]CGG[25]

* Genomic DNA change relative to hg19/GRCh37 at g.146993570 (chrX)

4.4.3. Preliminary study of intermediate allele instability

Stability of intermediate alleles is determined by the repeat size, the number of AGG interruptions and the parental origin of the allele (Nolin et al., 2015; Sullivan et al., 2002). To determine the repeat size and the number of AGG's, the entire cohort was subjected to SMRT sequencing. The characteristics of the 12 transmitted alleles are summarized in Table 14. Here, the alleles are grouped by repeat size (small intermediate alleles (45-49 units) versus large intermediate alleles (50-54 units)) and the number of AGG units (0, 1 or 2 interruptions). This table shows that most transmitted alleles are small (10), while only 2 individuals of our cohort carry a large intermediate allele. Eleven individuals carry repeats interrupted by AGG's: 5 women carried 1 AGG while 6 women carried an allele interrupted by 2 AGG's. Only one woman carried an allele without AGG units. The stability of intermediate alleles is larger if they are transmitted by males compared to females (Sullivan et al., 2002). From the 12 studied transmission, 9 alleles were transmitted via the father while 3 transmissions were transmitted through the mother (Supplementary Table 15).

Table 14: Summary of the repeat size and the number of AGG triplets for the 12 women carrying an intermediate allele.

	0 AGG	**1 AGG**	**2 AGG**	**Total**
45-49	1	3	6	10
50-54	0	2	0	2
Total	1	5	6	12

Finally, the stability of the intermediate alleles was studied. The sequencing results showed that all alleles were transmitted without any change in repeat size, i.e. no instability was detected in the included individuals (Supplementary Table 15). Interestingly, in individual 29 of family 10 mosaicism was detected with a cluster of alleles around 31 and 45 repeat units without AGG's on top of a normal allele (37 repeats – 3 AGG's). We hypothesized that the intermediate allele only became unstable after transmission because individual 29 carries the same allele as her father (45 CGG - 0 AGG's), and that this allele became unstable postzygogtically and contracted into a group of alleles clustered around 31 CGG units without AGG's. The normal allele with 37 repeats and 3 AGG's was inherited from her mother. A more detailed image of the different alleles of individual 29 is presented in Figure 24.

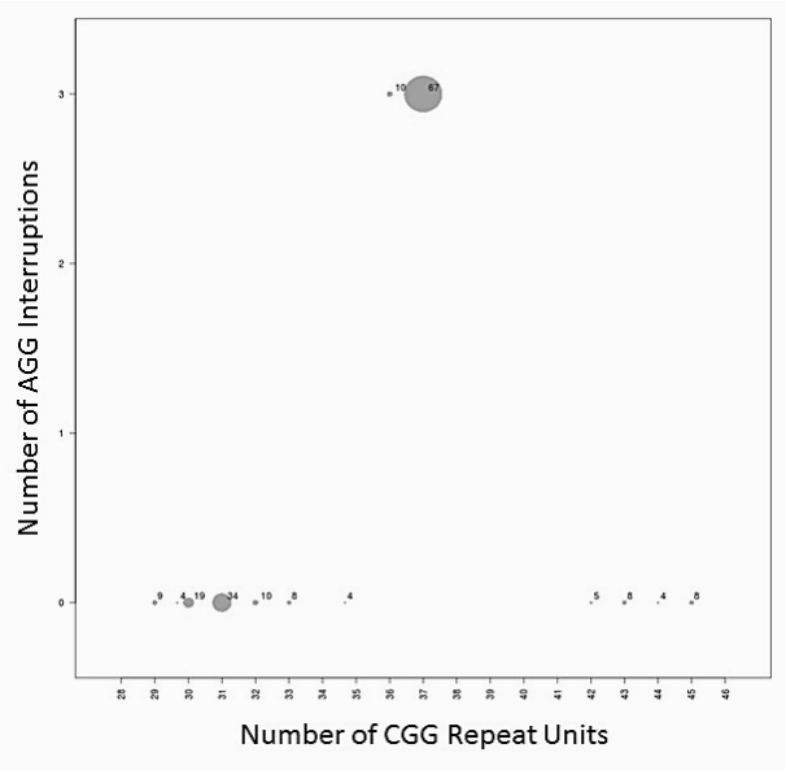


Figure 24: Overview of the different alleles of individual 29. The number of AGG interruptions is shown in function of the number of CGG repeat units. The size of the circles is in proportion with the number of supporting reads for each allele, which are also indicated in the figure. The normal allele (37 repeat units - 3 AGG's) was inherited from the mother of 29, while the intermediate allele (45 repeat units - 0 AGG's) was inherited from the father, became unstable postzygotically and contracted to a cluster of alleles centered around 31 CGG's without AGG's.

4.5. Discussion

Knowledge of the presence of AGG interruptions is of great value to determine the risk a female with a premutation allele will transmit a full mutation to her offspring, especially for small premutation alleles (55-85 CGG repeats)(Nolin et al., 2015; Yrigollen et al., 2014a). This is also increasingly recognized in international guidelines on *FMR1* genetic testing (Biancalana et al., 2015; Monaghan et al., 2013). Here, we demonstrated that single-molecule sequencing enabled not only the determination of the repeat sizes, but also the complete repeat structure in male and female gray zone and premutation alleles. The findings of all males and females were confirmed, whenever possible, by (TP-)PCR.

Single-molecule sequencing outperforms current strategies because it allows for an unambiguous separation of the normal from the expanded allele, which permits the determination of the repeat structure for each allele in every male or female. In addition, this method is significantly cheaper (± 15 euro/sample) compared to other methods, an advantage that will become even more strong thanks to PacBio's new Sequel system which has a higher throughput and is more cost-effective as compared to the PacBio RS II used in this study. Finally, single-molecule sequencing detects not only AGG interruptions, but any sequence variation at the repeat loci. For example, this technology will also identify duplications adjacent or within the repeat which are present in some individuals (Mononen et al., 2007) and avoid false-negative results sometimes generated by TP-PCR when interruptions are present inside a repeat (Braida et al., 2010; Radvansky et al., 2011).

In contrast to Loomis et al. (2013) who detected an AGG interruption in a plasmid, this is the first study to show the detection of AGG interruptions in males and the 2 alleles of females starting from genomic DNA.

After sequencing and *de novo* assembly, a main allele size is determined for each allele. However, the accuracy of SMRT sequencing allows not only to identify one main allele, but also to pick up any variation in repeat size centering around the main allele. This variation consists mainly of products which differ with ±1 repeat unit from the main allele. Since this is an amplicon-based method, this variation can originate from stutter products arising during the PCR reaction. The amount of stutter is influenced by different factors, including the repeat length and the number of AGG interruptions (Brookes et al., 2012; Mulero et al., 2006). In addition, the variation picked up by SMRT sequencing could also represent biological variation present in the DNA input material. Hence, during data interpretation it is important to be aware of the presence of these (stutter) products and carefully assess their influence on the determination of the repeat size of the main allele. However, since the repeat sizes called after de novo assembly fitted for 100% with the repeat sizes determined in the diagnostic laboratory (Figure 22), the presence of alleles with variable repeat sizes did not influence the accuracy of the assay. Hence, the impact of the presence of stutter on the determination of an expansion risk of an allele is negligible. In addition, even if stutter products present in a sample, both the number and the position of AGG interruptions is always completely concordant between the main allele size and the molecules with a variation in repeat size.

Knowledge of the risk for *FMR1* CGG expansion to occur has a profound impact on reproductive choices. Couples at risk of conceiving offspring with FXS can consciously choose for prenatal diagnosis with possible termination of an affected pregnancy (Burlet et al., 2006; Platteau et al., 2002; Sermon et al., 1999). This extremely difficult decision is often avoided by couples by not having children or choosing assisted reproduction associated with PGD to select only unaffected males or non-carrier female embryos. Unfortunately, PGD for this indication has always been difficult because female carriers are often affected by FXPOI which makes the retrieval of oocytes difficult (Burlet et al., 2006). Furthermore the expanded allele cannot be detected in a single cell, making PGD for FXS also technically a challenging task, although this can now be overcome by using new haplotyping methods (Natesan et al., 2014; Zamani Esteki et al., 2015). The risk of expansion will determine which reproductive choices will be made.

We implemented AGG analysis in *FMR1* diagnostic work-up since easy access to accurate AGG information is extremely valuable in guiding and reassuring couples to make the right decision. In this study we reported the results of AGG interruption analysis of the first 51 females with an intermediate or premutation allele that have been collected at the Center of Human Genetics, KU Leuven, UZ Leuven (Belgium) for 1 year. The impact of AGG interruptions is the most profound for females carrying a premutation sized between 60 and 84 repeats within which 13 females of our cohort fitted. From these 13 females, 3 females carried pure CGG repeats and hence have relatively high expansion risks ranging from 23 to 50% (Yrigollen et al., 2012, 2014a). The other 10 females had either 1 but most often 2 AGG interruptions and hence have more moderate expansion risks, except the 2 females with 80-84 repeats and 2 AGGs who also have around 30% chance their allele will expand into a full mutation (Yrigollen et al., 2012, 2014a).

Most of the normal, intermediate and premutation alleles are constructed with CGG_9AGG or $CGG_{10}AGG$ building blocks, concordant with previously published reports (Eichler et al., 1996; Yrigollen et al., 2014b). However, the repeats from 4 females deviated from these common building blocks and are more rare in the general population (Table 13)(Eichler et al., 1996; Yrigollen et al., 2014b). PCR-based assays might struggle to generate the correct repeat structure for these females as they use common haplotypes to infer the repeat structure of females whose X chromosomes camouflage each other's repeat structure (Chen et al., 2010). In another female a CTG interruption was detected within the CGG repeat, which has not been reported so far. Most interruptions are AGG triplets, although also a TGG interruption was discovered by Kunst and Warren (1994) in a male sample. Possibly, also these alternative interruptions might stabilize the CGG repeat. Systematic mapping and collection on the transmission of repeats carrying those rare interruptions would provide insights in the stability of such repeats. It remains unfortunate that TP-PCR cannot detect these novel interruptions which impedes further characterization of these unusual interruptions.

In rare cases where women carry 2 expanded alleles, selection can target the allele with the lowest risk. We already reported the CGG sequencing result obtained in this study to a female (Table 13, Female 2) carrying 2 premutations and who opted for PGD because she carried a translocation. The alleles of this woman have a size of 65 and 73 repeats and both carry 2 AGG interruptions. This knowledge influences the respective risks for expansion and allowed selection for the allele with the lowest risk, which is for this woman the allele of 65 repeats and 2 AGG interruptions.

Except for diagnostic use, single-molecule sequencing will also greatly facilitate large-scale studies which will be valuable to further fine tune risk estimates on the influence of AGG interruptions on the stability of the CGG repeat. We performed a small preliminary study where the influence of the repeat size, number of AGG interruptions and the influence of paternal versus maternal inheritance on the stability of *FMR1* intermediate alleles was investigated. Therefore, the stability of 12 intermediate alleles was determined based on 29 individuals in 10 families. All 12 intermediate alleles were inherited without any change in repeat size. According to literature, around 14% of intermediate alleles show small repeat changes upon transmission (Nolin et al., 2015). There are several reasons why this is not replicated in this study. First of all, this study is too small to produce statically significantly results. Secondly, the studied alleles are stable because they are mostly small (<50 repeat units) and interrupted by at least 1 AGG interruption. Interestingly, a mosaic individual was detected with a cluster of alleles around 45 and 31 CGG repeats with 0 AGGs in addition to a third cluster with 31 repeat units and 3 AGG units. We hypothesized that the intermediate allele was inherited stably from the father because the exact same repeat can be detected in both father and daughter, but that it became unstable postzygotically during mitotic divisions. This instability made the intermediate allele of 45 repeat units contract to 31 units. A similar case where a full mutation without AGG's contracted to an intermediate allele postzygotically has been reported already (Ferreira et al., 2013).

To conclude we have demonstrated that single-molecule sequencing correctly determines the repeat size of both the normal, gray zone and premutation alleles. Furthermore, we also detected the number and position of all AGG interruptions not only in males, but also in the 2 alleles of females. Single-molecule sequencing enables for the first time to separate unambiguously the 2 female repeats which enabled the generation of the exact repeat structure for both the normal and premutation allele. This technology was implemented in the *FMR1* diagnostic work-up and contributes to an accurate expansion risk for females with a premutation which simplifies choosing the most appropriate reproductive strategy. In addition, we show that SMRT sequencing is a good approach to investigate the influence of repeat size, AGG interruptions and the sex of the transmitting parent on intermediate allele instability. Since only 12 alleles were studied, more families will have to be collected in order to further fine-tune the predictions of expansion risks which are used today. This will facilitate the identification of individuals with unstable intermediate alleles after which these can be attentively followed up together with their relatives (Biancalana et al., 2015; Madrigal et al., 2011). In addition it seems likely this methodology can also be applied to study other tandem repeat expansion disorders where interruptions also influence the stability of the allele such as in Friedreich's ataxia (FRDA; MIM #229300), Myotonic dystrophy (DM1; MIM #160900) and different Spinocerebellar ataxia's (Braida et al., 2010; Gao et al., 2008; Holloway et al., 2011; Kroutil et al., 1996; Matsuura et al., 2006; Menon et al., 2013; Musova et al., 2009; Yu et al., 2011).

4.6. Data access

The raw data is available at the European Nucleotide Archive (ENA) (Leinonen et al. 2011) under study accession number PRJEB15075
(http://www.ebi.ac.uk/ena/data/view/PRJEB15075)

4.7. Supplementary Data

4.8.1. Supplemental Methods

De novo assembly on normal reads.

To perform an assembly on the normal reads, MIRA was run with the following settings: job = genome, denovo, accurate; technology=pcbiohq & parameters = -NW:cac=no.

De novo assembly on premutation reads and females with a small difference in repeat size.

For the *de novo* assembly on the premutation and mixture of reads the parameters were modified to: -NW:cac=no -HS:mnr=yes:nrr=50 -SK:mmhr=20 -PCBIOHQ_SETTINGS -CL:pec=yes:qcmq=34 & -AL:mo:400.

Long Amplicon Analysis.

As an alternative for the *de novo* assembly generated by MIRA 4.0, the Long Amplicon Analysis tool (LAA, SMRT portal v2.3.0) was evaluated. This algorithm was used to generate repeat structures only for the premutation alleles and thus was the protocol started with a minimum subread length of 370 bp, a maximum number of subreads of 2000, 10 bases trimmed from the ends of the sequences, clustering by gene family and phasing of the alleles. To demultiplex the amplicons, again symmetric barcodes were selected. Afterwards, only phased consensus sequences with an accuracy >90% were included in the analysis.

4.8.2. Supplemental Results

Long Amplicon Analysis.

In order to determine the repeat structure of the *FMR1*, except for the MIRA *de novo* assembler, also the Long Amplicon Analysis (LAA) pipeline included in Pacbio's analysis suite was evaluated. The LAA pipeline was used to analyze 33 female samples and 7 male samples. For 5 samples, no result was obtained because there was either no consensus generated (3) or only a consensus of one of both alleles in which case the premutation repeat was missing (2). For the group of females with a difference in repeat size >10 units (including 30 females), we compared the sizes generated by LAA and PCR (Supplemental Figure 28).

After LAA analysis 2 samples had a difference of 4 repeat units with the repeat size determined by PCR, while after MIRA assembly all sizes fitted within the error range of PCR (Supplemental Figure 28). The number of AGG interruptions matched 100% between the MIRA assembly and the LAA tool. For the 3 females with the difference in the allelic repeat size smaller than 10 units, the results were more difficult to interpret because the consensus was collapsed (2) or multiple repeat alleles were proposed (2). Overall, LAA performed worse than the MIRA assembly pipeline. Hence, MIRA assembly was chosen to perform all analysis in the manuscript.

4.8.3. Supplemental Figures and Tables

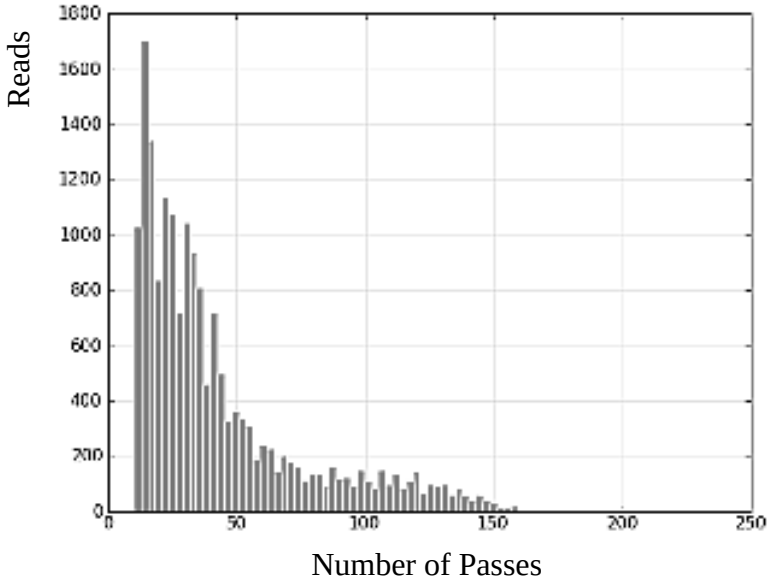


Figure 25: Distribution of the number of passes in function of the number of reads. All selected molecules have at least 10 full passes with half of the he molecules having at least 25 full passes. This ensures a high accuracy of the reads-of-insert.

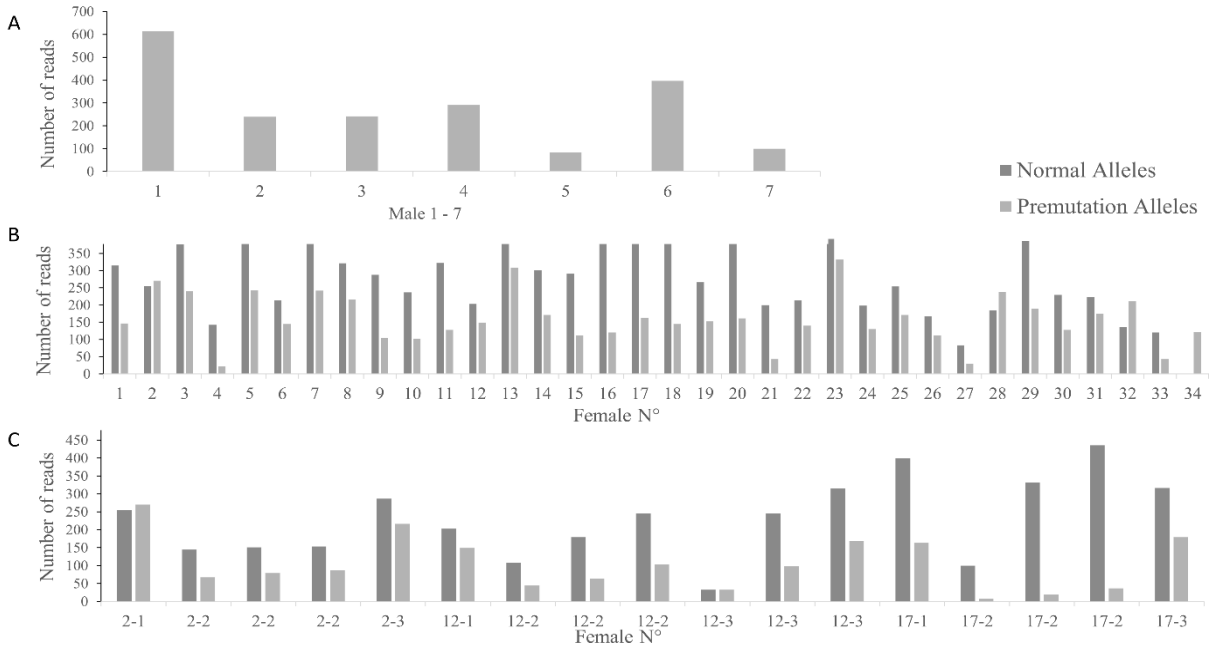


Figure 26: Overview of coverage of the different amplicons separated by the normal and the premutation allele. A: Coverage of the 7 male amplicons. B: Coverage of the normal and premutation allele for 34 female amplicons. Sample 1 – 33 are females with a normal and a grey zone/premutation allele, sample 34 is a female with 2 premutation alleles. C: Coverage for the repetitions of female 2, 12 and 17. The first number indicates the specific female, the second number indicates the sequencing run in which the amplicon was included. Tables 8, 9, 10 and 11 provide detailed information about the different samples.

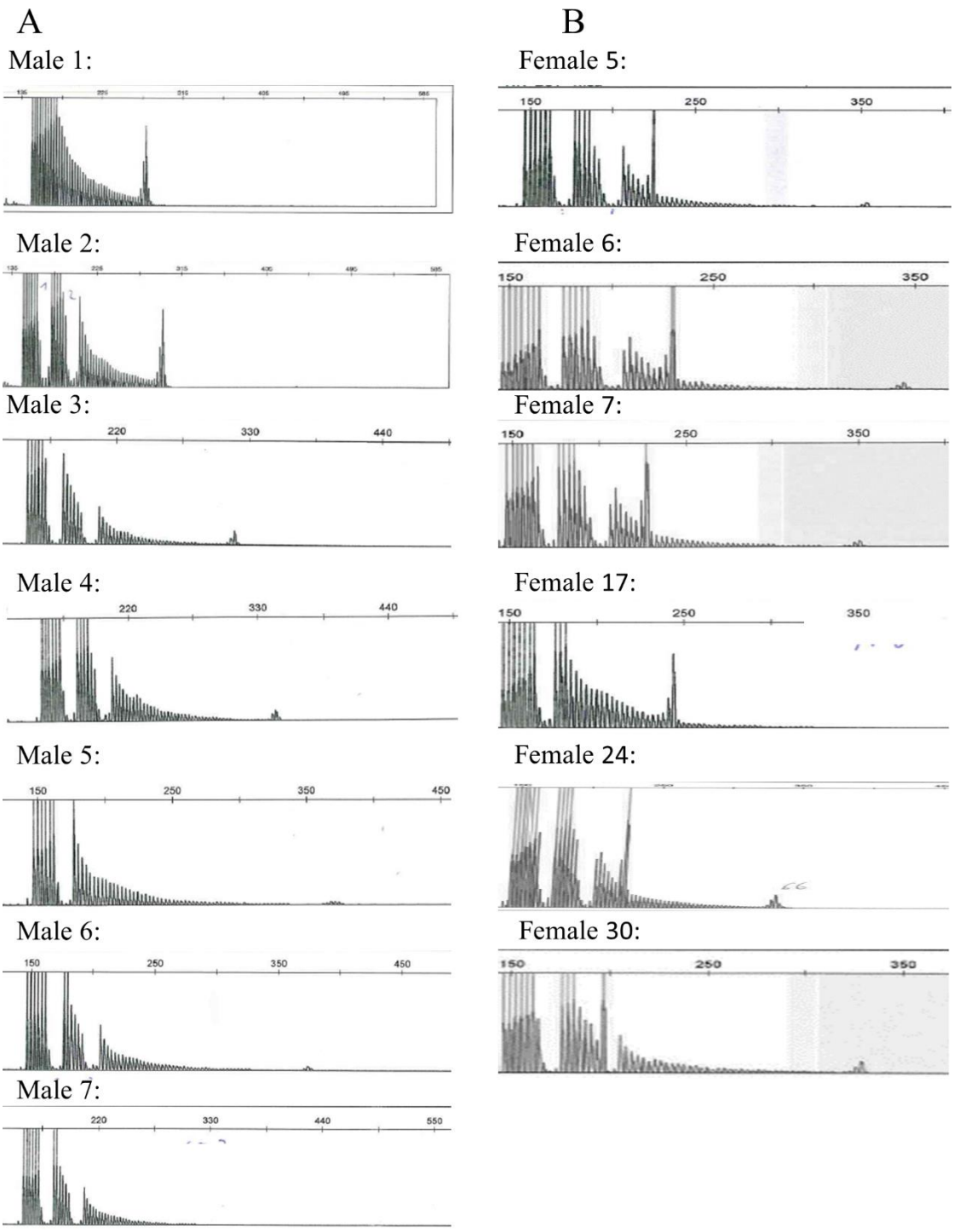


Figure 27: Validation of the of the FMR1 CGG repeat structure by TP-PCR for 7 males (A). Validation of the of the FMR1 CGG repeat structure by TP-PCR for 6 females. The repeat structure of the male and female examples can be found-in Table 8 and 9 (B). The number and position of all AGG units perfectly correlate between TP-PCR and single-molecule sequencing.

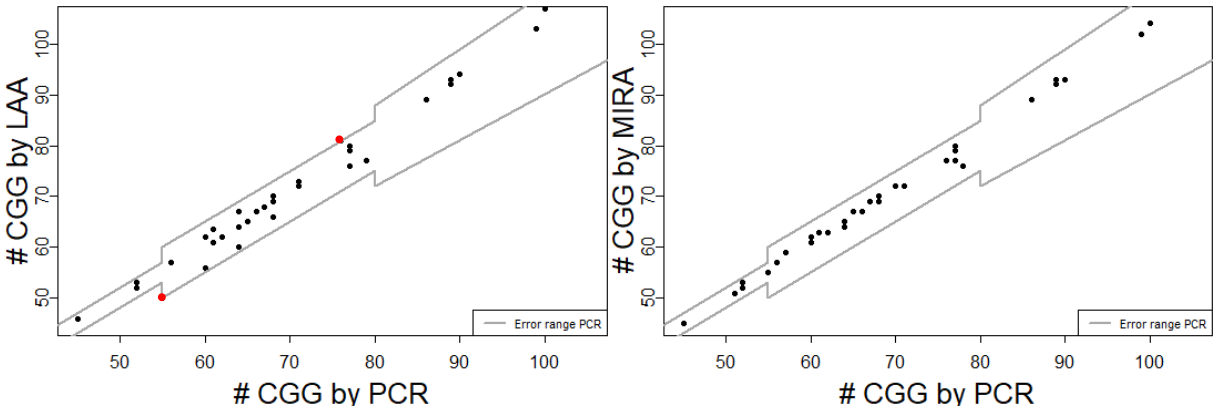


Figure 28: Comparison of the repeat size of 7 premutation males and 30 premutation females by PCR and sequencing followed by a *de novo* assembly by either Long Amplicon Analysis (LAA)(left panel) and MIRA assembly (right panel). Only females with a size difference larger than 10 units were included in this analysis. The LAA tool is included in PacBio's analysis suite but did not generate a result for 5 samples and in addition 2 samples have a difference of 4 repeat units and hence fall outside the error range of PCR (red dots, left panel)(see also 4.8.2 Supplemental results). The MIRA assembler was developed specially to handle repeats and generated a result for all samples which fitted always within the error range of PCR. Since MIRA assembly performed better than LAA, the former was chosen to execute the analysis included in the main manuscript.

Table 15: Overview of the repeat size, the number of AGG triplets and the repeat structure for the 29 individuals used to assess intermediate allele instability.

			ALLELE [1]			ALLELE 2		
		Relation	**# CGG**	**# AGG**	**Repeat structure**	**# CGG**	**# AGG**	**Repeat structure**
Family [1]	[1]	father of sample 3	32	2	CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]			
	2	mother of 3	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	47	[1]	CGG[9]AGG[1]CGG[37]
	3	**daughter of [1] & 2**	**32**	**2**	**CGG[9]AGG[1]CGG[12]AGG[1]CGG[9]**	**47**	**[1]**	**CGG[9]AGG[1]CGG[37]**
Family 2	4	mother of 5	40	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[20]	45	2	CGG[9]AGG[1]CGG[9]AGG[1] CGG[25]
	5	**daughter of 4**	**24**	**0**	**CGG[24]**	**45**	**2**	**CGG[9]AGG[1]CGG[9]AGG[1] CGG[25]**
Family 3	6	mother of 7	4[9]	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[29]			
	7	daughter of 6	**2[9]**	**2**	**CGG[9]AGG[1]CGG[9]AGG[1]CGG[9]**	**4[9]**	**2**	**CGG[9]AGG[1]CGG[9]AGG[1] CGG[29]**
Family 4	8	mother of [1]0	30	2	CGG[1]0AGG[1]CGG[9]AGG[1]CGG[9]	36	2	CGG[1]0AGG[1]CGG[9]AGG[1] CGG[15]
	[9]	father of [1]0	45	[1]	CGG[9]AGG[1]CGG[35]			
	[1]0	**daughter of 8 & [9]**	**30**	**2**	**CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]**	**45**	**[1]**	**CGG[9]AGG[1]CGG[35]**
Family	[1][1]	father of [1]2	4[9]	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[29]			
	[1]2	**daughter of [1][1]**	**30**	**2**	**CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]**	**4[9]**	**2**	**CGG[9]AGG[1]CGG[9]AGG[1] CGG[29]**
Family 6	[1]3	father of [1]5	45	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[24]			
	[1]4	mother of [1]5	[20]	[1]	CGG[10]AGG[1]CGG[9]	3[1]	[1]	CGG[10]AGG[1]CGG[20]
	[1]5	**daughter of [1]3 & [1]4**	**3[1]**	**[1]**	**CGG[10]AGG[1]CGG[20]**	**45**	**2**	**CGG[9]AGG[1]CGG[9]AGG[1] CGG[24]**
Family 7	[1]6	father of [1]8	45	[1]	CGG[10]AGG[1]CGG[34]			
	[1]7	mother of [1]8	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]			CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]
	[1]8	**daughter of [1]6 & [1]7**	**30**	**2**	**CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]**	**45**	**[1]**	**CGG[10]AGG[1]CGG34**
Family 8	[1][9]	father of 2[1] & 22	5[1]	[1]	CGG[9]AGG[1]CGG[41]			
	[20]	mother of 2[1] & 22	23	[1]	CGG[13]AGG[1]CGG[9]	30	2	CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]
	2[1]	**daughter of [1][9] & [20]**	**30**	**2**	**CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]**	**5[1]**	**[1]**	**CGG[9]AGG[1]CGG[41]**
	22	**daughter of [1][9] & [20]**	23	[1]	CGG[13]AGG[1]CGG[9]	5[1]	[1]	CGG[9]AGG[1]CGG[41]
Family [9]	23	father of 25 & 26	46	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[26]			
	24	mother of 25 & 26	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	30	2	CGG[10]AGG[1]CGG[9]AGG[1] CGG[9]
	25	**daughter of 23 & 24**	**30**	**2**	**CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]**	**46**	**2**	**CGG[9]AGG[1]CGG[9]AGG[1] CGG[26]**
	26	**daughter of 23 & 24**	30	2	CGG[10]AGG[1]CGG[9]AGG[1]CGG[9]	46	2	CGG[9]AGG[1]CGG[9]AGG[1] CGG[26]
Family [1]0	27	father of 2[9]	45	0	CGG[45]			
	28	mother of 2[9]	2[9]	2	CGG[9]AGG[1]CGG[9]AGG[1]CGG[9]	37	3	CGG[9]AGG[1]CGG[9]AGG[1] CGG[7]AGG[1]CGG[9]
	2[9]*	**daughter of 27 & 28**	**3[1]/45**	**0**	**CGG[31]/CGG[45]**	**37**	**3**	**CGG[9]AGG[1]CGG[9]AGG[1] CGG[7]AGG[1]CGG[9]**

*In woman 29 mosaicism of allele 1 is detected. Therefore, the 2 most occurring sequences of this allele are listed.

Chapter 5: Leveraging the power of SMRT sequencing to improve DMPK CTG repeat characterization

This chapter is partly based on:

1. Franck, S.*[1], Barbé, L.*[1], **Ardui, S.[2]**, Dziedzicka, D.[1], Vanroye, F.[1], Hilven, P.[1], Lanni, S.[3], Pearson, C. [3,4], Vermeesch, J.R.[2], Sermon, K.[1]. *MSH2* is the key driver of mismatch repair regulated repeat instability in myotonic dystrophy (in preparation).

[1]Department for Reproduction and Genetics, Vrije Universiteit Brussel, Brussels, 1090, Belgium;
[2]Department of Human Genetics, Katholieke Universiteit Leuven, Leuven, 3000, Belgium;
[3]Genetics & Genome Biology, The Hospital for Sick Children, Toronto, Ontario, M5G 0A4, Canada;
[4]Department of Pediatrics, University of Toronto, Toronto, Ontario, Canada M5S 1A1
*Authors equally contributed to this work

2. Dastidar,S. [1], **Ardui, S.[2]**, Singh, K.[1], Majumdar, D.[1], Nair, N.[1], Fu, Y.[3,4], Reyon, D.[3,4], Samara, E.[1], Gerli, M.F.M.[5], Klein, A. F.[6], De Schrijver, W.[1], Tipanee, J.[1], Seneca, S.[7], Tulalamba, W. [1], Wang, H.[1], Chai, Y. Ch.[1], In't Veld, P.[8], Furling, D.[6], Tedesco, F. S.[5],. Vermeesch, J. R.[2], Joung, J. K.[3,4], Chuah, M. K.[1,9] and VandenDriessche, T.[1,9] (2018). Efficient CRISPR / Cas9-mediated editing of trinucleotide repeat expansion in myotonic dystrophy patient-derived iPS and myogenic cells. Nucleic Acids Res. 1, 1–24.

[1]Department of Gene Therapy & Regenerative Medicine, Vrije Universiteit Brussel, Brussels 1090, Belgium,
[2]Department of Human Genetics, University of Leuven, Leuven 3000, Belgium
[3]Molecular Pathology Unit, Center for Cancer Research and Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA02129, USA
[4]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA,
[5]Department of Cell and Developmental Biology, University College London, London WC1E6DE, UK,
[6]Sorbonne Universités, INSERM, Association Institute de Myologie, Center de Recherche en Myologie, F-75013, France,
[7]Research Group Reproduction and Genetics (REGE), Center for Medical Genetics, UZ Brussels, Vrije Universiteit Brussel, Brussels 1090, Belgium,
[8]Department of Pathology, Vrije Universiteit Brussel, Brussels 1090, Belgium
[9]Center for Molecular & Vascular Biology, Department of Cardiovascular Sciences, University of Leuven, Leuven 3000, Belgium

5.1. Abstract

Myotonic Dystrophy type 1 (DM1) is caused by the expansion of a CTG repeat in the 3' untranslated region of the myotonic dystrophy protein kinase (*DMPK*) gene. In this study a novel methodology using Single Molecule Real-Time (SMRT) sequencing to determine CTG repeat variability was developed. This yielded a higher accuracy, higher throughput and less hands-on time compared to traditional techniques like Southern blot. This methodology is now applied to study the influence of the mismatch repair system on DM1 repeat instability. Besides, in a second project, SMRT sequencing was used to assess the efficiency of CRISPR/CAS9 excision of the *DMPK* CTG repeat region. Additionally, this allowed to grasp the complete picture of the molecular consequences of *DMPK* gene editing. In summary, this study shows that SMRT sequencing is a powerful and flexible tool able to contribute to different aspects of DM1 research.

5.2. Introduction

An unstable CTG tandem repeat is located in the 3' untranslated region (UTR) of the myotonic dystrophy protein kinase (*DMPK*; MIM# 605377) gene on chromosome 19 (Fu et al., 1992). Whereas moderate repeat numbers (5-37) are present in healthy individuals, expansion of these repeats to more than 50 units (up to more than 2000 units) causes myotonic dystrophy type 1 (DM1; MIM# 160900)(Mahadevan et al., 1992; Savić Pavićević et al., 2013). This is a multisystemic disorder where patients are typically affected with progressive myopathy and myotonia, cardiac conduction defects and cognitive impairments (Meola and Cardani, 2014). The disease is inherited in an autosomal dominant mode and has a prevalence of around 1:8000 (Theadom et al., 2014).

For this study we collaborated with Prof. Karen Sermon and prof. Thierry VandenDriessche from the Vrije Universiteit Brussel (VUB) to focus on 2 challenges in DM1 research. Prof. Karen Sermon and her team focus on the influence of the mismatch repair system (MMR) on tandem repeat instability of the *DMPK* CTG repeat. Although MMR plays a crucial role in safeguarding the integrity of the human genome, it has been shown that it drives the expansion of tandem repeats (Owen et al., 2005; Panigrahi et al., 2010; Schmidt and Pearson, 2016b). A proposed model to explain this is based on the inefficient removal of hairpin loops by MMR. The MutS homologue 2 (*MSH2*), a component of MMR, will recognize these hairpins and that should normally be followed by hairpin removal thereby leaving the STR unchanged. However, *MSH2* has an unusually high affinity for TR-containing hairpins that allows the hairpin loop to be incorporated causing expansion (Schmidt and Pearson, 2016b). Unfortunately, the role of *MSH2* has only been investigated in mouse models and human cell lines with *MSH2* knock-down, which did not completely reveal the functioning of *MSH2*. The mouse models do not faithfully reproduce the DM1 phenotypes and still a considerable amount of *MSH2* was expressed in the human knock-down models (Du et al., 2013; Hegan et al., 2006; Nakatani et al., 2015; Savouret et al., 2003; Seriola et al., 2011; Tomé et al., 2009). That is why the group of Karen Sermon developed a *MSH2* knock-out model with the CRISPR/Cas9 technology in DM1 affected human pluripotent stem cells (hPSC).

To make fully use of these valuable cell lines, it is crucial to be able to accurately assess the *DMPK* CTG repeat variability. An interesting option to assess this variability is to perform a small-pool PCR (SP-PCR) followed by Southern blot (Barbé et al., 2017; De Temmerman et al., 2008; Seriola et al., 2011). A SP-PCR consists of multiple PCR reactions on small pools of input DNA in the order of 0.5 to 200 genome equivalents. In contrast to normal PCR (5 - 100 ng of input DNA), SP-PCR allows a better determination of the repeat variability, including the detection of the common alleles and the rarer alleles only present in a minority of cells. Unfortunately, the detection of the PCR products by Southern blot is cumbersome, time-consuming and only resolves the size of the products with a low resolution. Hence, this is a limiting factor in the *MSH2* knock-out study where a high resolution and throughput is important. Since we showed in chapter 2 Single Molecule Real-Time (SMRT) sequencing can span long *FMR1* CGG repeats, here we want to explore if SMRT sequencing can also handle long *DMPK* CTG repeats. In addition, we also explore SMRT sequencing to determine *DMPK* CTG instability. In order to avoid variation introduced by PCR, ideally an amplification-free method enriching *DMPK* CTG repeats should be developed similar to chapter 3. However, for the sake of time limitations, this approach was not replicated for this TR.

Although the amplicon-based strategy developed in chapter 4 allows to determine the repeat structure of *FMR1* premutation alleles, it is not appropriate to determine repeat variability since after a PCR on bulk DNA stutter products cannot be distinguished from biological variation. Therefore, here SP-PCR in combination with SMRT sequencing was examined to determine *DMPK* CTG repeat variability.

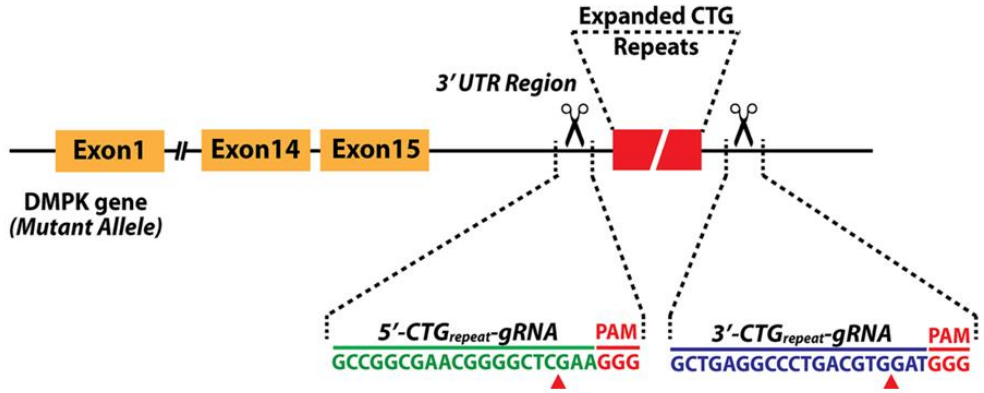


Figure 29: Excision of the DMPK CTG repeat from the 3'UTR (adapted from Dastidar et al., 2018).

Another challenge in DM1 research was tackled by the group of Prof. Thierry VandenDriessche. They explored the use of the CRISPR/CAS9 tool for gene-editing of the *DMPK* CTG repeat. To do so they examined the use of 2 sgRNA's located both up- and downstream of the CTG repeat. Similar to the approach developed in chapter 3, this strategy is aimed at excising the CTG repeat from the genome, albeit it is used here *in vivo* (Figure 29). Since muscle dysfunction is one of the most dominant phenotypes in DM1, they generated different DM1 patient-specific myogenic cell lines. These cell lines replicate many of the key features of DM1 *in vitro* and are therefore proven to be useful to validate gene editing. After CRISPR/CAS9 treatment of these cell line the efficiency and the molecular consequences of the gene-editing on the *DMPK* CTG region has to be determined. For this reason, an amplification-based approach similar to the approach used in chapter 4 is developed to determine the efficiency of CRISPR/CAS9 editing.

Thus, DM1 research is being hampered by the lack of an adequate technology to study the genetic architecture of the CTG repeat. SMRT sequencing proved already its value in the sequencing of long TRs (Loomis et al., 2013; McFarland et al., 2015), but is not yet used in DM1 research. Therefore, in this study 2 solutions based on SMRT sequencing and tailored to DM1 research were developed to:

1) Determine *DMPK* CTG variability of large repeats
2) Determine the efficiency and the impact of editing by CRISPR/CAS9 on the sequence composition of the *DMPK* CTG gene.

5.3. Materials and Methods

5.3.1. DNA samples

TR variability was determined in DM1 affected tissues, human Embryonic Stem Cells (hESCs) and human induced Pluripotent Stem Cells (hiPSCs) derived from fibroblasts. The CRISPR/CAS9 gene editing efficiency was performed in skeletal myotubes and myocytes derived from primary myoblasts and fibroblasts from DM1 patients that were reprogrammed into iPSCs. DNA was extracted from these cell lines using the Qiagen DNeasy Blood and Tissue kit (Qiagen) according to the manufacturer's protocol.

This study was approved by the local ethical committee and the Commission for Medical Ethics of the UZ Brussel and informed consent was obtained from all patients.

5.3.2. PCR Amplification

The *DMPK* CTG repeat was amplified with the LongAmp Taq polymerase kit (New England Biolabs, Ipswich, MA) and 2 specific primers (Table 16). In a 25 µL reaction mix containing 2.5 units LongAmp Taq DNA polymerase, 1x LongAmp buffer (New England Biolabs), 0.2 mM dNTPs (Illustra DNA polymerization mix, GE Healthcare) and 2.5% dimethyl sulphoxide were mixed together with 0.4 µM primers listed in Table 16.

Afterwards, the mixture was incubated as followed: 4 minutes of initial denaturation at 94°C, 35 cycles of 30 seconds denaturation at 94°C, 8 min annealing and extension at 65°C and a final extension step at 65°C for 10 minutes. This approach can be used for standard DNA input amounts (≈ 50 ng) and for small-pool PCR's where the input DNA amount is only 20 pg.

Table 16: Forward and reverse primer to amplify the *DMPK* CTG variability

	CTG variability determination	**CRISPR/CAS9 editing efficiency**
Forward	CTTCCCAGGCCTGCAGTTTGCCCATCCA	ATCTTCGGGCAGCCAATCAAC
Reverse	GAACGGGGCTCGAAGGGTCCTTGT	CGTGGAGGATGGAACACGGAC

5.3.3. SMRT sequencing

The generated amplicons were prepared for sequencing as described in PacBio's guide for Preparing SMRTbell Libraries using PacBio® Barcoded Adapters for Multiplex SMRT® Sequencing. Up to 40 different PCR products were annealed with a different barcoded adaptor and multiplexed together in one library preparation. When the total DNA input amount of a library was only low (< 10 ng), 500 ng of PUC19 plasmid was added before exonuclease treatment, this to avoid degradation of intact SMRTbells. Hereafter, each library was sequenced completely on a single SMRT cell by a PacBio RS II using the DNA/Polymerase binding Kit P6 v2 (Pacific Biosciences) for a 360 minutes movie. We used PacBio's DNA Sequencing Reagent Kit 4.0 v2 for all runs.

5.3.4. Sequencing Analysis

The long reads generated by SMRT sequencing allow to pass each input molecule multiple times, yielding an accurate, circular consensus (CCS) read. These reads were obtained with the RS_ReadsOfInsert.1 protocol from PacBio's SMRT portal (v2.3.0) with a minimum of 1 full pass, a minimum predicted accuracy of 90% and demultiplexing with symmetric barcodes.

Afterwards an in-house python script was developed to determine for each PCR product the distribution of the repeat sizes and the repeat content for all individual reads. The repeat size was determined by measuring the distance between 2 unique regions flanking the CTG repeat. Hence, this analysis does not require alignment to a reference. This is an advantage since these long CTG repeats do not map efficiently to their original location. Repeat variability could subsequently be determined by comparing the repeat size within or between PCR products. The CRISPR/CAS9 gene editing efficiency was determined by studying the distribution of the read sizes. Before editing, only the wild type (≈ 723 bp) and the expanded allele (> 4000 bp) are present. After excision with CRISPR/CAS9, fragments with an excised CTG repeat (≈ 633 bp) or an incomplete excised repeat (723-4000) will be present. Besides, wild type and expanded alleles may also be present in the sample if CRISPR/CAS9 editing was not 100% efficient (Figure 30). Based on the proportions of the different read sizes the efficiency of CRISPR/CAS9 excision (cutting efficiency) could be determined by:

$$\text{Cutting Efficiency (\%)} = 1 - \frac{\#\text{ Wild type alleles}}{\#\text{Wild type alleles} + \frac{\#\text{ Cut alleles}}{2}}$$

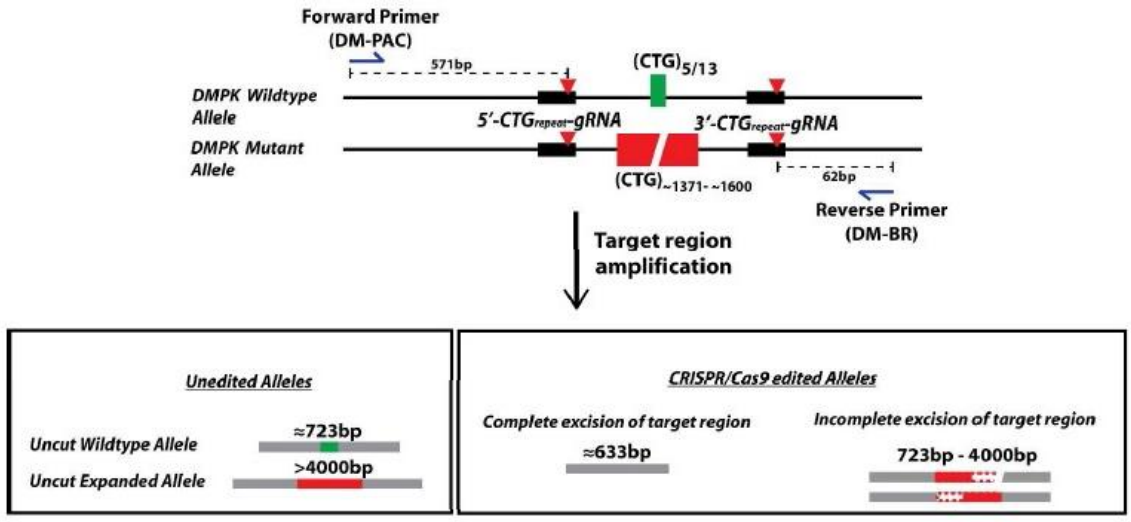


Figure 30: Overview of the theoretical read sizes before (unedited alleles) and after CRISPR-CAS9 editing (CRISPR/CAS9 edited alleles)(adapted from Dastidar et al., 2018).

5.4. Results

5.4.1. Determination of DMPK CTG variability

In order to determine the repeat variability of long CTG repeats, DNA from 4 DM1-affected tissues (heart, kidney, liver and muscle) was amplified and analyzed by SMRT sequencing.

The results were compared to the results obtained by SP-PCR followed by Southern blot, a technique that is traditionally used to determine the variability (Figure 31)(Barbé et al., 2017; De Temmerman et al., 2008; Seriola et al., 2011). Strikingly, SMRT sequencing can pass long CTG repeats with up to 1500 units (≈ 4.5 kb of repeated sequence). However, there was no concordance between the variability determined by sequencing and Southern blot. The size distribution of the sequencing library was analyzed before and after sequencing revealed a shift in the size distribution (Figure 32). Before sequencing, the mean size of the library was around 2500 bp, but this decreased to 1000 bp after sequencing. This shift in size distribution is explained by the way a library is loaded on a SMRT Cell: the molecules move by diffusion into the zero-mode waveguides (ZMWs). Since small molecules move easier inside ZMWs then large molecules, they will preferentially be sequenced. This explains the reduction in the library size and the discordance between Southern blot and SMRT sequencing.

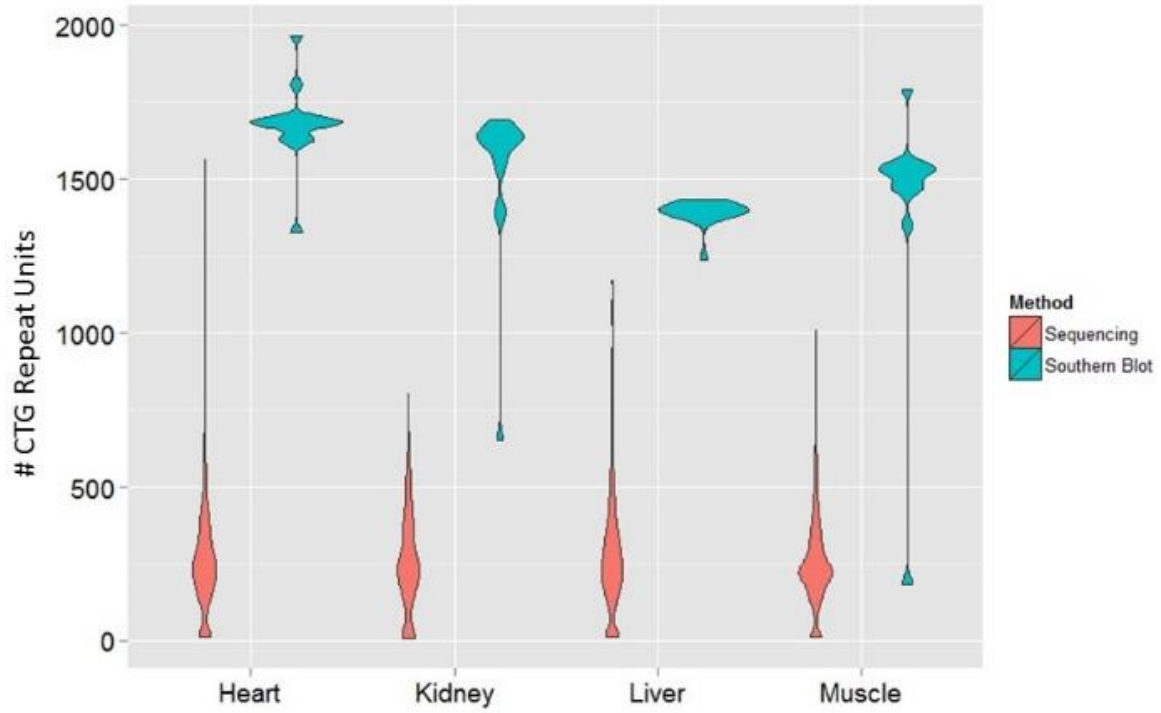


Figure 31: *DMPK* CTG repeat variability of 4 DM1 affected tissues (heart, kidney, liver and muscle) determined by SMRT sequencing and Southern blot. There was no concordance between both methods.

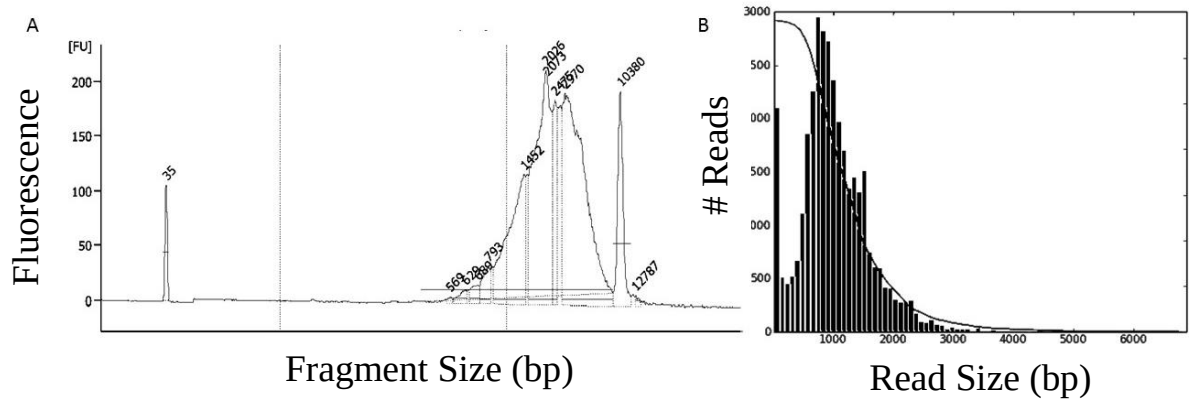


Figure 32: Size distribution of the sequencing library before sequencing (A) and after sequencing (B). SMRT sequencing preferentially sequences the smaller molecules in a library causing a reduction in the library size from 2500 bp before sequencing, to 1000 bp after sequencing. This biases the determination of TR variability.

To circumvent the size bias introduced by PCR and sequencing, DNA was diluted into small pools of 20 pg (≈ 5 genome equivalents) before amplification (Figure 33A). For each sample 20 SP-PCRs were performed followed by barcoding, multiplexing and sequencing (Figure 33B-C). After a consensus was generated for each read (Figure 33D), the repeat size distribution was determined within each PCR product. Based on this distribution, for each PCR product a plot was generated (Figure 33E) and the median repeat size was determined. Finally, The TR variability of a sample could be determined by combining the median repeat sizes of all 20 PCR products (Figure 33F). By using SP-PCR, the alleles with a high abundance will no longer overwhelm the minor alleles since a PCR started from only 5 genomes. Also, the bias introduced by sequencing can be avoided because all generated molecules of one PCR product will have a similar size.

In order to validate the assessment of SP-PCR by SMRT sequencing, the TR variability was determined in a DM1 affected muscle tissue and compared to SP-PCR followed by Southern blot (Figure 34). This analysis showed that there was no statistical difference between the TR variability determined by SMRT sequencing and Southern blot. Hence, this approach can now be used to for TR variability determination.

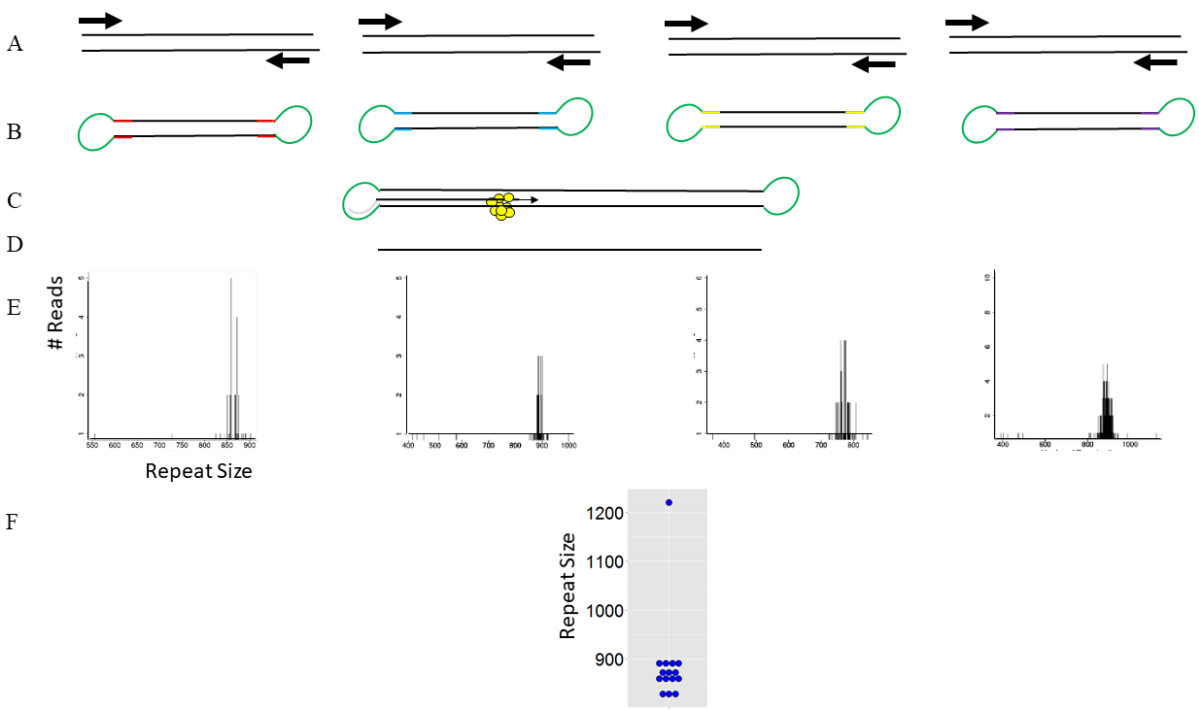


Figure 33: Overview of SP-PCR followed by SMRT sequencing. Twenty PCR reactions on 20 pg of DNA were performed (A), followed by barcoding (B) which allowed to multiplex the different PCR products in one library. After sequencing (C) and consensus generation (D), the distribution of the repeat sizes was determined for each PCR product (E). In order to retrieve the repeat size of the original molecule(s), the median of each PCR product was determined. By combining the median repeat sizes of all 20 PCR products (F), an accurate view on the CTG repeat variability was generated.

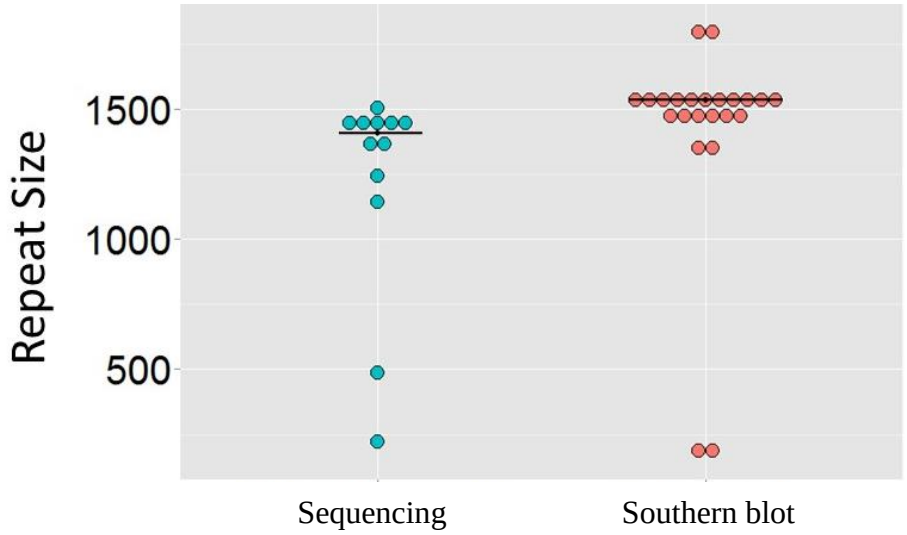


Figure 34: Comparison of the CTG repeat variability of DM1 affected muscle tissue determined by SMRT sequencing and Southern blot, both after SP-PCR. There is no statistical difference in TR variability between both methods.

Subsequently, SMRT sequencing was applied to different passages of cell lines with (*MSH2* +/+) and without (*MSH2* -/-) *MSH2* knock-out. This showed that the size of the CTG repeat increased and was more unstable with increasing culture times in the presence of *MSH2* (Figure 35). By contrast, after *MSH2* knock-out the CTG repeat becomes more stable and even contracts over time (Figure 35).

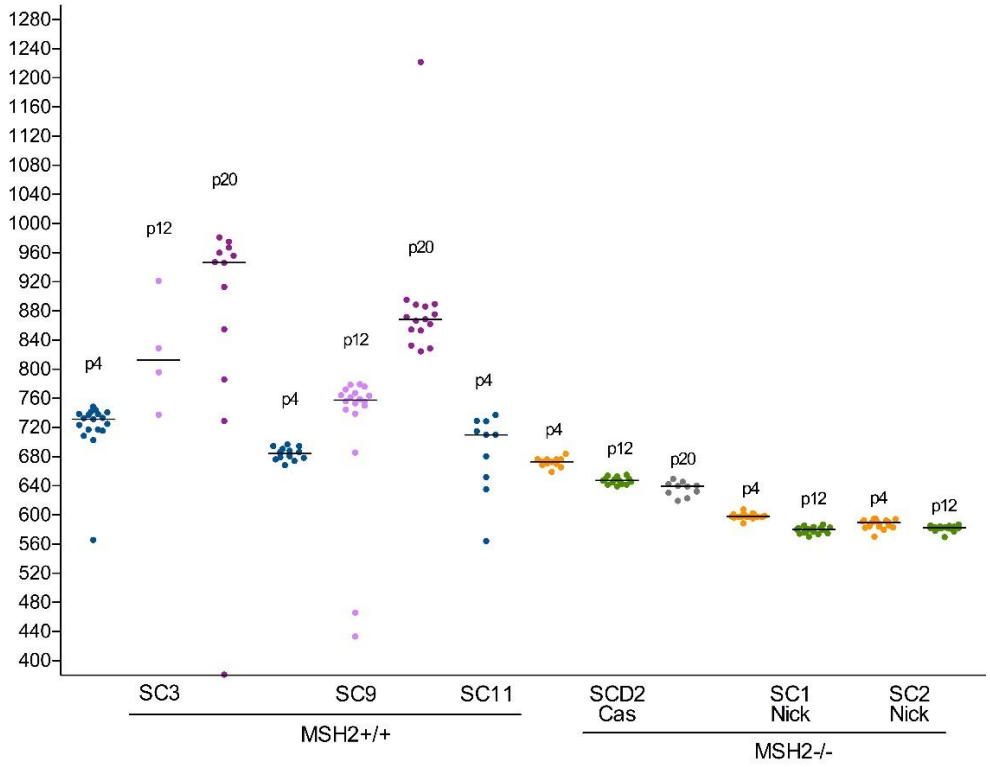


Figure 35: Repeat variability determined by SP-PCR and SMRT sequencing for different cell lines with *MSH2* (*MSH2* +/+) and without (*MSH2* -/-) and different passages (4, 12 and 20). The presence of *MSH2* stimulates repeat exansion and instability, while its knock-out stabilizes the repeats and even induces contraction.

5.4.2. Determination of the efficiency of DMPK CTG excision by CRISPR/CAS9

In part 2 of this study the effects of CRISPR/CAS9 gene editing of the *DMPK* CTG repeat were studied. Two primers flanking the cutting sites of CRISPR/CAS9 were used to amplify the repeat followed by sequencing (Figure 30). In an unedited cell line, this will generate 2 amplicons: one from the wild type allele (≈ 723 bp) and one from the expanded allele (≈ 4000 bp). In a cell line edited with CRISPR/CAS9, amplicons with a complete excision of the repeat (≈ 630 bp) are detected next to unedited alleles if the gene editing efficiency did not reach 100%. In addition, there was also a possibility that only a single gRNA cut up- or downstream of the repeat resulting in amplicons with only a partial repeat deletion (723 – 4000 bp) (Figure 36). A typical example of the read size distribution of a sample edited with CRISPR/CAS9 and some representative reads are shown in Figure 36.

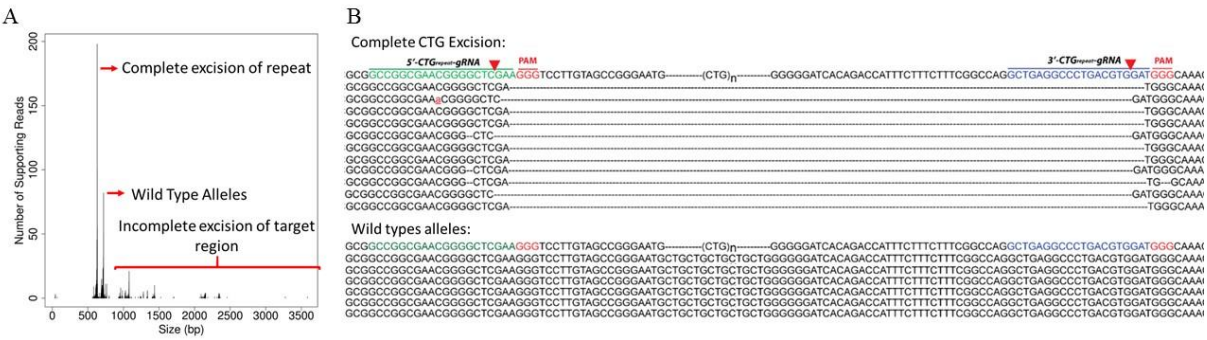


Figure 36: SMRT sequencing of the *DMPK* CTG repeat after gene editing by CRISPR/CAS9. (A) Distribution of the read sizes of a cell line edited with CRISPR/CAS9. Except for the wild type allele (≈ 723 bp), also alleles with a complete excision (≈ 630 bp) or a partial excision (723 - 4000 bp). (B) Representative reads from the wild type and CTG excised alleles.

After sequencing, the efficiency of CRISPR/CAS9 gene editing was calculated based on the number of wild type and CTG excised alleles. Long, unedited alleles were not incorporated into the calculations because they were underrepresented in the data due to their large size compared to the normal/unedited alleles. SMRT sequencing showed that a robust excision (10 – 46%) of the CTG repeat from the *DMPK* 3'UTR was achieved in 4 myogenic differentiated DM1-iPSC cell lines (Figure 37). In addition, SMRT sequencing was able to unambiguously demonstrate targeting of the wild-type allele due to the presence of indels in proximity of the respective PAM sites located upstream or downstream of the normal CTG repeat with the characteristic low number of repeats. Similarly, it was also possible to demonstrate targeting of the mutant allele due to the presence of indels in proximity of the respective PAM sites located upstream or downstream of the CTG repeat expansion.

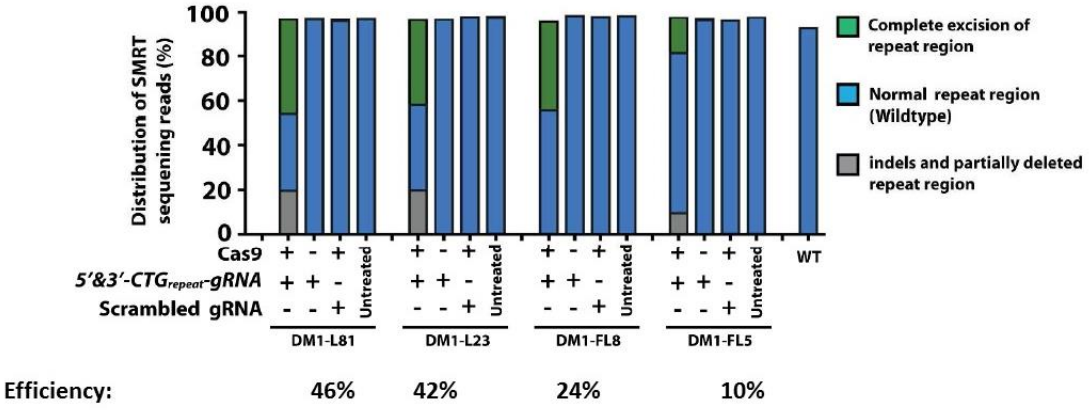


Figure 37: Overview of the efficiency of the CTG excision from the *DMPK* 3'UTR region. For each cell line, 2 controls were generated where the cell line was treated only with functional guide RNA's or with CAS9 and scrambled gRNA. No editing was found in the control samples.

5.5. Discussion

SMRT sequencing has already a strong track record of the successful sequencing of tandem repeats (Guo et al., 2014b; Loomis et al., 2013), but to our knowledge it has not yet been used to study the *DMPK* CTG repeat. Up until today, long CTG repeats can only be characterized by Southern blot, a labor-intensive method with only a limited resolution. Therefore, in this study the use of SMRT sequencing was explored to investigate the *DMPK* CTG repeat.

We showed that SP-PCR followed by SMRT sequencing provides a detailed and accurate determination of the variability of long CTG repeats. This was used to characterize different cell lines with and without *MSH2* knock-out which contributed to the evidence for the role of *MSH2* in driving TR expansion. In a second project, SMRT sequencing was used to define the efficiency of CRISPR/CAS9 excision of the *DMPK* CTG repeat in 4 myogenic differentiated DM1-iPSC cell lines. These cell lines represent a very valuable model since skeletal dysfunction is the main clinical manifestation of DM1 and because they are non-transformed cell lines. Except for the gene editing efficiency, sequencing also allows to zoom in on the DNA composition of the target site. This permits to control CRISPR/CAS cutting sites and to screen for induced mutations. A drawback of the approach is that both the wild type and the mutated allele are targeted by CRISPR/CAS. However, it is reassuring that excision of the CTG repeat in either the wild type or mutant DMPK locus does not result in any mutation of the DMPK protein itself, since the repeat is located in the 3'UTR. Altogether, this proof-of-concept study validates the use of CRISPR/Cas9 to genetically correct nucleotide repeat expansions associated with dominant genetic disorders like DM1. This opens possibilities for therapeutic options since in these genetically corrected cells also the cellular phenotype and the downstream pathways were restored (data not included).

SMRT sequencing not only outperforms Southern blot in accuracy and resolution, it is also less cumbersome and requires significantly less hands-on time. The latter is especially important in this study where a high number of cell lines need to be processed. Similar questions as we addressed in this paper (TR variability & gene editing efficiency) are also important for other TRs, and hence we foresee that these strategies will also be transferred to these research fields.

SP-PCR is an interesting solution to circumvent the bias introduced by PCR when amplifying a broad distribution of tandem repeats. However, performing SP-PCR entails that multiple PCR have to be done in parallel. In addition, these PCR products have to be handled individually during part of the library preparation. Indeed, this increases the work load and the reagent cost of the amplification and library preparation. Therefore, novel approaches avoiding SP-PCR will be explored in the future to tackle the *DMPK* CTG repeat variability. Already the size bias of the Sequel (the newest instrument from Pacific Biosciences) is significantly less compared to the older RSII platform which was used in this study. Ultimately, a targeted amplification-free enrichment method similar to chapter 3 would remove the need for PCR completely. Such methods have already been used for other tandem repeats (*FMR1, HTT, SCA10,…),* but are not yet being developed for DM1 (See chapter 3 and Höijer et al., 2018; Schüle et al., 2017; Tsai et al., 2017).

Sequencing allowed to determine the CTG repeat variability and the CRISPR/Cas9 activity at the target region of the DM1-iPSC-Myo cells. The developed methodologies are now intensively used in DM1 research and can be easily applied to other TRs like FXS.

Chapter 6: Discussion

Partly based on:

Ardui, S.[1], Ameur, A.[2], Vermeesch, J. [1], Hestand, M.[3] (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 46, 2159–2168.

[1]Department of Human Genetics, KU Leuven, Leuven 3000, Belgium
[2]Department of Immunology, Genetics and Pathology, Uppsala University, Science for Life Laboratory, Uppsala 75108, Sweden,
[3]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia and 4Department of Clinical Genetics, VU University Medical Center, Amsterdam 1081 BT, The Netherlands

6.1. Sequencing Analysis of STRs

Modern medical genomic research and diagnostics relies heavily on DNA sequencing. These technologies are used in a wide range of applications at different stages during the entire human lifespan. Applications range from prenatal diagnostics and newborn screening, to diagnosing rare diseases, hereditary forms of cancer, pharmacogenetics testing, predisposition testing for a plethora of diseases and testing for future generations in terms of carrier screening and pre-implantation genetic diagnoses (Katsanis and Katsanis, 2013; Vermeesch et al., 2016). DNA sequencing even has even the potential to determine the complete genetic and epigenetic signature of STRs, which remained largely unknown hitherto (Bornman et al., 2012).

The widespread use of sequencing is fueled by its rapid evolution during the last decades. It started 40 years ago with Sanger sequencing. Although it provides high quality reads of 1 kb, Sanger sequencing has only a low-throughout (Heather and Chain, 2016). The first decade of the 21st century brought forth the development of multiple new methods of DNA sequencing. As opposed to first-generation platforms, these new second-generation technologies have considerably shorter reads, but at massively higher throughput. Though these more recent short-read platforms have permitted scientists to quickly hunt for causative mutations in a panel of disease genes, the exome, or even the entire human genome (Koboldt et al., 2013), they all share common drawbacks. For example, the short read lengths hinder resolving repeat regions (McFarland et al., 2015), and the amplification steps during library preparation and/or the actual sequencing reaction also introduce chimeric reads, variation in repeat size, and an underrepresentation of GC-rich/poor regions (Guan and Sung, 2016).

The rise of long-read sequencing (or third generation sequencing) overcomes many of these problems by generating long reads and a real-time base read-out, allowing base modification detection. Two long-read sequencers are nowadays available: the Single Molecule Real-Time (SMRT) sequencing developed by Pacific Biosciences and nanopore sequencing developed by Oxford Nanopore Technologies (ONT). Nanopore based technologies are catching on more and more and they are likely to represent future platforms. Though nanopore based technologies are simple and low-cost (reviewed in (Deamer et al., 2016; Jain et al., 2016; Lu et al., 2016), SMRT sequencing is currently more matured. For example, SMRT sequencing has hitherto closed more gaps in the human genome, improved structural variation and can characterize STRs more accurately compared to nanopore sequencing (Chaisson et al., 2015; Ebbert et al., 2018; Ishiura et al., 2018; Loomis et al., 2013; McFarland et al., 2015; Seo et al., 2016).

6.2. Leveraging the Power of SMRT Sequencing for STR analysis

Repeat expansion disorders are influenced by the size of the repeat, the presence of methylation, interruptions and the degree of mosaicism. (Pretto et al., 2014b). Interestingly, all these aspects can be grasped by SMRT sequencing. Here we discuss how the benefits of SMRT sequencing can be used to improve both research and the clinical management of these disorders.

6.2.1. The Power of SMRT Sequencing in STR Research

Applications based on amplification

Repeatome research, including STRs underlying repeat expansion disorders, significantly benefits from the advent of SMRT sequencing. The first STRs studied by SMRT sequencing were an expanded *FMR1* CGG repeat and an ATTCT repeat embedded in intron 9 of *SCA10*. The repeats were amplified and completely sequenced. This allowed the construction of the repeat sequences and the detection of both known and novel interruptions (Loomis et al., 2013; McFarland et al., 2015). We used SMRT sequencing to study the *DMPK* CTG repeat (Chapter 5). This approach allows to determine the *DMPK* CTG variability and is less cumbersome than Southern blot. By using this method to analyze multiple hESC lines with- and without *MSH2*-knock out, it was shown that *MSH2* propels *DMPK* CTG variability in these cell lines. These findings corroborate with earlier findings in mouse models and human knock-down systems (Hegan et al., 2006; Nakatani et al., 2015). In future, the technology will be applied to study the influence of methylation of a CCCTC-binding factor site neighboring the *DMPK* CTG repeat on somatic instability of the CTG repeat. Secondly, SMRT sequencing was also used to show that CRISPR/CAS9 can excise the *DMPK* CTG region in myogenic differentiated DM1-iPSC cell lines (Chapter 5). Interestingly, excision of the expanded repeat restored the normal cellular phenotype which shows the potential of this approach as a possible therapeutic strategy in the future. To our knowledge, this is the first study addressing the genetics of DM1 with SMRT sequencing. In the future can these approaches also be expanded to other disease-causing STRs.

Amplification-Free Applications

Novel amplification-free enrichment methods are currently being developed. Methods using amplification are very error-prone, especially when amplifying (tandem) repeats, and furthermore remove all epigenetic marks (Loomis et al., 2013). Thus, using amplification impedes a complete genetic and epigenetic characterization of tandem repeats. Since the PacBio instrument suite only has a limited throughput, until recently it was a huge economic burden for a laboratory to sequence an entire human genome if one is only interested in a particular locus. In order to rein this limited throughput, whilst still taking advantage of the strengths of SMRT sequencing, we developed an amplification-free enrichment method for the *FMR1* CGG repeat based on excision of the region by CRISPR/CAS9 (Chapter 3). This method achieved significant enrichment in both BAC molecules (30X enrichment, 32,832 on-target reads) and the human genome (102X enrichment, 5-on-target reads). In BAC molecules, the true underlying biological repeat size variability and DNA modifications were detected, showing the potential of the enrichment method. In human DNA, no variation in repeat size nor any DNA modifications could be detected due to the low coverage. For this reason, further improvements will have to be explored to make this method more powerful and economically stable. The developed enrichment method is also easily transferable to other laboratories since only mainstream laboratory equipment and reagents are needed and to other loci since the design is easily adaptable.

Simultaneous with the development of our method, currently 2 other amplification-free enrichment strategies are under development as well (Pham et al., 2016; Tsai et al., 2017). Both methods target the *FMR1* CGG repeat. The method developed by Tsai et al. (2017) can in addition also enrich the *C9ORF72* G_4C_2 repeat, the *HTT* CAG repeat and the *Sca10* ATTCT repeats simultaneously with the *FMR1* CGG repeat. Already a few applications were explored with their methodologies. For example, they studied *HTT* CAG variability in blood and sequenced expanded *Sca10* ATTCT repeats (Höijer et al., 2018; Schüle et al., 2017). The fact that multiple research groups are developing amplification-free methods highlights the need of the sequencing community for such methods.

The future of SMRT sequencing in STR research

SMRT sequencing will increase our understanding of the genetics of STRs. Besides avoiding amplification biases, SMRT sequencing permits native DNA capture that could allow the direct detection of epigenetics. Altogether, this will shed more light on the influence of the (epi)genetic variability on the penetrance, complexity and phenotypical variation of repeat expansion disorders like fragile X syndrome (FXS). Nevertheless, before reaching these targets, several improvements must be realized. For example, DNA modification detection in human genomes by SMRT sequencing is nowadays still at an early stage of development. Unfortunately, the influence of 5mC on the kinetics of the sequencing polymerase is only subtle, and thus can they only be detected with a high coverage (≈ 100 -250 X). In the future, a higher throughput and improved detection algorithms could facilitate the read-out of base-modification. Also, more matured amplification-free enrichment technologies will facilitate obtaining a high coverage of specific loci in an affordable manner. Since the (epi)genetic variability of STRs is important for all repeat expansion disorders, the enrichment technologies should be expanded to all disease-causing STRs. For example, in chapter 5 the variability of the *DMPK* CTG repeat was still determined with a PCR based approach. The advent of an amplification-free method enriching the *DMPK* CTG repeat would make the determination of the CTG variability more straightforward. In addition, this may also shed light on the presence of up- and downstream regions of the CTG repeat. Interestingly, it has already been suggested that methylation of these regions may account for the maternal inheritance bias in DM1, the larger maternal expansions, age-of-onset and disease severity (Barbé et al., 2017). Furthermore, it would not only be interesting to enrich (disease-causing) STRs, but also other genomic loci like tumor suppressor genes prone to epigenetic silencing.

SMRT sequencing will also facilitate the discovery of new disease-causing STRs. Many of these elements escaped detection by second-generation sequencing apparatuses, and hence remain to be discovered. For instance, once it was known that chromosome 9p was involved in both ALS and FTD, it took 5 years until the C9orf72 G4C2 repeat was finally discovered in 2011 (Renton et al., 2011). The presence of SMRT sequencing could have greatly speeded up this process. Moreover, it will simultaneously reveal novel interruptions which are known to influence the stability of STRs and the phenotype of the patient (Nolin et al., 2015; Schüle et al., 2017).

6.2.2. The Power of SMRT Sequencing in Diagnostics

SMRT sequencing has the potential to uncover unexplored diagnostic opportunities. In addition to the strengths of this technology, which we already discussed extensively, SMRT sequencing also has also a very short turnaround time making it appealing for diagnostic use.

SMRT sequencing is already in use for the clinical management of FXS and tissue transplantations. In FXS it is specifically used to detect the number of interrupting AGG units, which together with the repeat size correlates with the risk that a premutation allele expands to a full mutation in future generations (Chapter 4). Since these full mutations cause FXS, the AGG information is subsequently used in the genetic counselling of women weighing the risk of having a child with FXS. We anticipate that this method will shortly replace competing technologies for AGG interruption detection for both research and clinical applications.

On a long-term perspective, SMRT sequencing has the potential to play an even more important role in fragile-X diagnostics. Once the amplification- free enrichment methods become more reproducible, cheaper and streamlined, they could be used diagnostically to determine the size of full mutations of the *FMR1* CGG repeats and simultaneously assess the presence of mosaicism and methylation, all of which influence the phenotype of FXS. Traditionally, this is determined by a combination of error-prone PCR-based assays and tedious Southern blots. Consequently, replacing Southern blots with faster and more direct SMRT sequencing will greatly enhance *FMR1* and additional repeat disorder diagnostics. Another application of SMRT sequencing that is already diagnostically applied is the genotyping of the human major histocompatibility complex (HLA) to select donors in organ- and stem cell transplantation. The long reads permit a better characterization of this highly polymorphic region since it can sequence entire genes, as opposed to exons with second-generation sequencing (Gabriel et al., 2009; Trowsdale and Knight, 2013; Turner et al., 2018).

SMRT sequencing improves insights in many clinically relevant regions. Hence, more and more applications are nowadays on the verge of diagnostics. Firstly, SMRT sequencing has revolutionized viral and microbial sequencing since a *de novo* assembly of a bacterium can be constructed from only one SMRT cell, while a viral genome can even be contained in one single read (Bull et al., 2016; Miyoshi-Akiyama et al., 2015). It has already been used to sequence several viruses (influenza virus, hepatitis B virus & human immunodeficiency virus) and bacteria (*Myobacterium tuberculosis* and *Salmonella enterica*), where it is used to identify the cause of the infectious disease and the presence of mutations and to indicate the virulence of the invading bacteria or viruses by revealing the epigenetic modifications (Bergfors et al., 2016; Bull et al., 2016; Miyoshi-Akiyama et al., 2015; Nakano et al., 2017b; Satou et al., 2015; Yao et al., 2016). Since the genome of bacteria and viruses is only small, they easily meet the coverage requirements necessary to detect these modifications. Secondly, SMRT sequencing can also be used to differentiate between a pseudogene and its homologous functional gene. The drug metabolism gene *CYP2D6* has multiple homologous pseudogenes and thus SMRT sequencing can be used to identify variants specific to the gene. These variants influence the metabolizing potential of the gene and hence this will allow to identify metabolizer phenotypes (Buermans et al., 2017; Qiao et al., 2016). Thirdly, in cancer genes, SMRT sequencing can be used to identify and phase mutations leading to drug resistance (Cavelier et al., 2015).

In addition, whole genome and transcriptome SMRT sequencing of cancer cell models already revealed novel fusion genes and splicing isoforms (Nattestad et al., 2018). Although whole genome sequencing is at the moment only affordable in a research setting, this might soon become available for diagnostics as well. Interestingly, since whole genome sequencing is performed on non-amplified material, this would also detect DNA modifications which is crucial in the development of cancer. Moreover, whole-genome SMRT sequencing can be used to *de novo* assemble and phase high quality genomes in which one can hunt for causative mutations and novel genes (Chaisson et al., 2015; Chin et al., 2016; Shi et al., 2016). This opens possibilities for truly personalized genomics in the near future.

6.2.3. Long-read sequencing and beyond

SMRT sequencing and nanopore sequencing are currently the only third generation sequencers that are commercially available and used in the research field. Although nanopore sequencing can generate longer reads and requires less investment, the reads from SMRT sequencing are more accurate than these of nanopore sequencing. Whereas this might not be crucial for whole genome sequencing, this is nevertheless critical to accurately assess the genetics of STRs (Ebbert et al., 2018). Remarkably, novel technologies are currently at the horizon and can be used in combination with SMRT sequencing or even have the possibility to replace them. For example, in order to generate better genome assemblies, SMRT sequencing can be combined with optical and chromatin interaction mapping. Optical mapping incorporates fluorescent nucleotides at nicks created by endonucleases. This way they generate fluorescent patterns on extremely long molecules (> 2 Mb), which can be assembled into a genome (Moll et al., 2017). Chromatin interaction mapping (Hi-C) is based on crosslinking DNA loci which are in proximity in the 3D space, but not per se in the reference genome. These techniques generate the bigger picture of the genome architecture that can be used to scaffold contigs generated by SMRT sequencing (Bickhart et al., 2017; Shi et al., 2016). Novel third-generation sequencers like Genia (Roche) and Hyb & Seq method (Nanostring) are currently being developed and might replace SMRT sequencing or ONT nanopore sequencing in the future. A novel, fourth generation of sequencers is also being developed at the moment. These technologies promise to generate *in situ* or spatially resolved genomics and transcriptomics, which will improve the study of tissue heterogeneity (Ke et al., 2016). When applied on STRs, they could increase our understanding of the impact of (epi)genetic mosaicism on RNA and protein level at a specific location within a tissue. One could wonder if, one day, it will be possible to determine the genome, transcriptome and proteome of spatially resolved, living, single cells.

6.3. STR research: What's next?

Up until today different aspects of STRs remain opaque because they have been understudied and existing technologies are often not suited for their analysis. Consequently, up until today the degree of STR variability within a tissue or individual and between different individuals is not yet characterized. In addition, it is unclear which biological consequences these variations may have. Upcoming third generation sequencing technologies can grasp STR variability at a much higher level and could contribute to an increased understanding of STR biology in the future.

Though STRs are very polymorphic genetic elements, they have often been missed in the quest for causative mutations. Therefore, the list of the 40 repeat expansion disorders which have been described hitherto is probably only the tip of the iceberg. A first step to unravel more causative STRs is the supplementation of the existing STR databases that are currently limited to STRs with a limited repeat length and complexity. This can be improved by inquiring long-read genome assemblies for novel and polymorphic STRs. A recent study of 5 genome assemblies already produced 200.000 novel polymorphic STRs (Genovese et al., 2018). These STR catalogues can subsequently be used, for example, in the deciphering of the underlying causes of X-linked intellectual disability.

Although arrays, whole-exome and whole-genome sequencing with second-generation sequencers has been performed on this disease, the origin remains unknown for 1/3 of these patients (Hu et al., 2016). Hence, it is expected that novel technologies and analysis tools with an increased sensitivity for STR polymorphisms will reveal novel, disease-causing STRs in these patients with X-linked intellectual disability (Duitama et al., 2014). This hypothesis is further strengthened by the fact that many STRs are residing in genes involved in neurodevelopmental and brain function (Legendre et al., 2007; Riley and Krieger, 2009). Furthermore, STR polymorphism is part of the missing heritability which is currently observed in many association studies (Hannan, 2018b). Many of these studies have been executed for complex polygenic disorders, but often yielded fewer causative mutations than expected. These studies are relying on SNP arrays, and hence do not grasp STR polymorphism. Besides, these SNPs cannot be associated with STRs since they are often more variable than the interrogated SNPs. Since STRs do have a functional impact, novel approaches assessing STRs in association studies will increase the detection of causative mutations contributing to polygenic disorders.

Characterizing STRs at a nucleotide level will potentially lead to novel therapeutic approaches in the future. An especially alluring therapeutic approach is gene editing, whereby the causative genetic mutation is blocked or even restored. Indeed, this has also the potential to replace an expanded STR with an allele containing a normal number of repeat units. In chapter 5 SMRT sequencing was used to assess the efficiency of CRISPR/CAS9 excision of expanded *DMPK* CTG repeats. CRISPR/CAS9 excision reached a robust efficiency and in addition also the cellular phenotype on RNA- and protein level was restored. One of the major advantages of the proof-of-concept study is that it is based on DM1 patient-derived myogenic cells. Interestingly, these cells are capable of muscle tissue regeneration in vitro. Hence, they have the potential to be used to replace dysfunctional and degrading muscle tissue in an autologous setting (Dastidar et al., 2018). Only targeting the disease allele or replacing the excised expanded repeat with a normal allele would even further improve this method. Nevertheless, the emergence of CRISPR/CAS9 gene editing opens new perspectives for correcting genetic disorders including large nucleotide expansion disorders like FXS.

Chapter 7: Summary - Samenvatting

7.1. Summary

Around 1.5 million short tandem repeats (STRs) are spread across the entire human genome. STRs are functionally important elements that are able to modulate the phenotype of an individual. They can modify cellular biology by influencing the genome, transcriptome and proteome of a cell. The most extreme examples of the functional impact of STRs are the more than 40 repeat expansion disorders like fragile X syndrome (FXS) and myotonic dystrophy 1 (DM1).

Up to now, many aspects of STRs remain illusive. Due to an historical underestimation of their importance and the lack of adequate technologies they remain understudied. The rise of Single Molecule Real-Time (SMRT) sequencing from Pacific Biosciences changes this paradigm and arms researchers with better tools to investigate STRs. In this thesis we studied different aspects of STRs, thereby exploring different assets of SMRT sequencing.

SMRT sequencing can span long, GC-rich repeats, whilst simultaneously revealing DNA modifications in the sequenced region. Unfortunately, the throughput of the technology is limited, making it economically unfeasible to sequence an entire genome if one is only interested in a single locus or a subset of the genome. Therefore, enrichment strategies like PCR are commonly being used. These strategies aver nevertheless very error-prone, especially when amplifying repeats. Furthermore, they remove all epigenetic marks. Thus, amplification impedes the complete genetic and epigenetic characterization of STRs. To tackle this, we developed a CRISPR-CAS9 approach to excise the *FMR1* CGG repeat in combination with restriction enzymes to remove off-target genomic DNA. This generated a very accurate picture of the *FMR1* CGG repeat variability of the BAC molecule and made it possible to identify DNA methylation. Indeed, besides avoiding amplification biases, this method permits native DNA capture and, hence, allows for direct detection of base modifications. On human DNA, enrichment factors over 100X were achieved while up to 5 reads covering the *FMR1* CGG repeat could be retrieved from one SMRT cell. Albeit further improvements are necessary to allow wide spread implementation, this method has the potential to significantly further unravel the complex genotype-phenotype correlations in FXS. In addition, it could be used to screen for long, methylated CGG alleles in diagnostics where it could replace complicated and laborious Southern blots.

FXS arises from the *FMR1* CGG expansion of a premutation (55–200 repeats) to a full mutation allele (>200 repeats) in females. This type of expansion is the most frequent cause of inherited X-linked intellectual disability. The risk for a premutation to expand to a full mutation allele depends on the repeat length and AGG triplets interrupting this repeat. Therefore, it is necessary to map these AGG interruptions in order to study the stability of the *FMR1* allele. Additionally, easy access to accurate size estimates and AGG information is also of great importance in genetic counseling since they allow for women carrying a premutation allele to estimate the risk for expansion. Unfortunately, the detection of AGG interruptions is hampered by technical difficulties. We demonstrated that single-molecule sequencing enables the determination of not only the repeat size, but also of the complete repeat sequence, including AGG interruptions in male and female alleles. This approach outperforms current strategies because it allows for an unambiguous separation of the normal allele from the expanded one.

This permits the determination of the repeat structure for each allele in every male or female. Hence, we implemented SMRT sequencing as a diagnostic tool to identify AGG interruptions in females with a *FMR1* premutation. By doing so, we improved the risk assessments for genetic counseling and positively impacted the management of the disorder. Except for diagnostic use, single-molecule sequencing will also facilitate large-scale studies assessing the influence of AGG interruptions on the stability of the CGG repeat. We performed already a proof-of-principle study to investigate the influence of AGG's on the stability of intermediate *FMR1* CGG alleles (45-54 repeat units).

SMRT sequencing was also explored to study the *DMPK* CTG repeat underlying DM1. Firstly, the variability of long CTG repeats was determined by small-pool PCR followed by long-read sequencing. This approach resulted in a higher accuracy, higher throughput and less hands-on time compared to Southern blots. Therefore, this methodology is now used to study the influence of the mismatch repair system on DM1 repeat instability. Undoubtedly, this approach will be implemented more broadly in the future. Besides, long-read sequencing was also used to assess the efficiency of CRISPR/CAS9 excision of the *DMPK* CTG repeat region. To our knowledge, this is the first study tackling the *DMPK* CTG repeat by single-molecule long-read sequencing. Ultimately, a targeted amplification-free enrichment method for the *DMPK* CTG repeat would remove the need for PCR completely and could further improve the analysis of this repeat.

To conclude, SMRT sequencing is a powerful tool forging ahead STR research and diagnostics. In this thesis novel methodologies were developed to make maximal use of the advantages of SMRT sequencing (high accuracy, long reads & detection of base modifications). It will be interesting to see how novel methodologies employing long-read sequencing developed in this thesis and by other research groups will move the STR field ahead in the future.

7.2. Samenvatting

Er bevinden zich ongeveer 1,5 miljoen korte tandemherhalingen (bv. CGGCGGCGG) verspreid over het hele humane genoom: van exonen en intronen tot promotors, transcriptiefactoren en zelfs regio's zonder coderende functie. Deze korte tandemherhalingen zijn functionele elementen die een impact hebben op het fenotype van een individu. Hun aanwezigheid heeft immers effect op de biologie van een cel doordat ze het genoom, het transcriptoom en het proteoom kunnen beïnvloeden. Dat tandemherhalingen functionele elementen zijn, komt het meest duidelijk tot uiting in de verschillende ziektes die veroorzaakt worden door grote expansies van deze herhalingen. Vandaag de dag zijn er meer dan 40 van deze ziektes gekend waaronder het fragiele-X syndroom (FXS) en myotone dystrofie type 1 (DM1).

Vele aspecten van deze korte tandemherhalingen zijn tot nu toe onduidelijk doordat ze in het verleden niet voldoende bestudeerd werden. Dit heeft twee belangrijke redenen: enerzijds onderschatten wetenschappers vaak de impact die deze elementen kunnen hebben, anderzijds worstelen de meeste technologieën tot nu toe met hun repetitieve karakter. Dankzij de komst van een nieuwe technologie, 'Single Molecule Real-Time (SMRT) sequencing' ontwikkeld door Pacific Biosciences, kunnen deze tandemherhalingen meer gedetailleerd geanalyseerd worden.

In deze doctoraatsthesis wordt gebruik gemaakt van verschillende voordelen van SMRT sequencing om verschillende aspecten van tandemherhalingen te bestuderen. Twee voordelen van SMRT sequencing zijn ten eerste, de lange leeslengte die toelaat om grote, GC-rijke tandemherhalingen te overspannen en ten tweede, de detectie van epigenetische modificaties. Een nadeel is echter het beperkt aantal moleculen dat per experiment kan gelezen worden. Hierdoor is het financieel onhaalbaar om een volledig humaan genoom te sequeneren als men slechts geïnteresseerd is in één bepaalde regio. Daarom worden er vaak aanrijkingsmethoden zoals PCR gebruikt om één bepaalde regio te amplificeren. Voor deze studie is dit echter niet interessant aangezien PCR erg onnauwkeurig is wanneer het een tandemherhaling moet amplificeren. Bovendien verwijdert het alle epigenetische markeringen. Mét PCR is een volledige genetische en epigenetische karakterisering van tandemherhalingen dus onmogelijk. Daarom ontwikkelden we in dit onderzoek een aanrijkingsmethode zonder amplificatie waarbij de moleculaire schaar CRISPR/CAS9 werd gebruikt om de *FMR1* CGG herhaling uit het genoom te knippen. Door het toevoegen van restrictie-enzymen, konden off-target genomische DNA fragmenten weggeknipt worden. Door deze amplificatie-vrije aanrijkingsmethode toe te passen op BAC DNA kon de variabiliteit van de *FMR1* CGG herhaling en de epigenetische modificaties accuraat bepaald worden. Doordat met deze methode het originele DNA molecule wordt afgelezen, kan niet alleen de originele, biologische variabiliteit van tandemherhalingen bepaald worden, maar kan ook om de aanwezigheid van DNA modificaties gedetecteerd worden. In humaan DNA kon een aanrijking van meer dan 100X bekomen worden, waarbij in één SMRT cel tot 5 reads de *FMR1* CGG herhaling bevatten. Deze methode heeft het potentieel om complexe genotype-fenotype correlaties in FXS te verbeteren, mits het aantal reads met de *FMR1* CGG herhaling verder verhoogd kan worden in de toekomst.

Daarnaast zou deze amplificatie-vrije aanrijking van de *FMR1* CGG herhaling in de toekomst ook in fragiele-X diagnostiek geïmplementeerd kunnen worden om lange, gemethyleerde CGG allelen op te sporen. Het kan daarbij Southern blot vervangen, een arbeidsintensieve en inaccurate methode die vandaag in gebruik is.

FXS, de meest voorkomende erfelijke vorm van een X-gebonden verstandelijke beperking, ontstaat nadat een premutatie (55-200 CGG herhalingen) expandeert tot een volledige mutatie (>200 CGG herhalingen). Het risico dat een premutatie overgaat in een volledige mutatie hangt af van de grootte van de premutatie en het aantal AGG eenheden die de CGG herhalingen onderbreken (CGGCGGCGG**AGG**CGGCGG). Het is daarom belangrijk om deze AGG eenheden in kaart te brengen wanneer men de stabiliteit van het *FMR1* CGG allel wil bestuderen. Daarnaast is een eenvoudige toegang tot accurate AGG informatie ook belangrijk in genetische counseling, waar de informatie samen met de grootte van de tandemherhaling gebruikt wordt om het risico dat een premutatie expandeert te bepalen. Deze risicoanalyse kan dan gebruikt worden door vrouwen met een premutatie om de kans in te schatten dat ze een kind met FXS. De detectie van deze AGG's is echter zeer moeilijk met de huidige technieken. Daarom werd in dit onderzoek aangetoond dat SMRT sequencing kan gebruikt worden om AGG eenheden te detecteren in zowel mannelijke als vrouwelijke stalen. Deze methode is uniek aangezien het de enige methode is die het normale allel van het premutatie allel kan scheiden en vervolgens de structuur kan bepalen in beide allelen. Daarom werd SMRT sequencing na de ontwikkeling en validatie geïmplementeerd als een diagnostische test voor vrouwen met een *FMR1* premutatie. Hierdoor konden we de risicoanalyse verfijnen wat de genetische counseling van vrouwen met een premutatie kon verbeteren. Daarnaast werd deze methode ook gebruikt in een kleinschalige studie waarbij de invloed van AGG onderbrekingen op de stabiliteit van intermediaire *FMR1* allelen (45-54 CGG eenheden) werd bestudeerd.

SMRT sequencing werd ook gebruikt om de *DMPK* CTG herhaling geassocieerd met DM1 te bestuderen. Ten eerste werd de variabiliteit van de herhaling bepaald door small-pool PCR te combineren met SMRT sequencing. Met behulp van deze methode kon de variabiliteit van de CTG herhaling sneller en met een hogere accuraatheid geanalyseerd worden, in vergelijking met Southern blots. Daarom wordt dit nu toegepast om de invloed van DNA-herstelmechanismen op de variabiliteit van de CTG herhaling te bestuderen. Daarnaast werd SMRT sequencing ook ingezet om de efficiëntie van het uitknippen van de *DMPK* CTG herhaling door CRISPR/CAS9 te bepalen. Doordat hiervoor een sequeneringsmethode gebruikt werd, kon de invloed van CRISPR/CAS9 tot op nucleotide niveau bepaald worden. Dit is de eerste studie waarbij SMRT sequencing toegepast werd voor de studie van de *DMPK* CTG herhaling. Deze analyse zou in de toekomst nog verder verbeterd kunnen worden als er ook voor deze herhaling een amplificatie-vrije aanrijkingsmethode zou ontwikkeld worden.

Uit deze studie kunnen we concluderen dat SMRT sequencing een krachtig instrument is dat de studie en diagnostiek van tandemherhalingen verbetert. Nieuwe methodologieën werden ontwikkeld om de voordelen van deze technologie maximaal te gebruiken (lange 'reads', een hoge accuraatheid en de detectie van DNA modificaties). Het is erg interessant om op te volgen hoe nieuwe methodes, zowel ontwikkeld in deze thesis als door andere onderzoeksgroepen, zullen gebruikt worden om onze kennis over tandemherhalingen verder te vergroten in de toekomst.

Bibliography

Akogwu, I., Wang, N., Zhang, C., and Gong, P. (2016). A comparative study of k-spectrum-based error correction methods for nextgeneration sequencing data analysis. *Hum. Genomics* 10, 20. doi:10.1186/s40246-016-0068-0.

Albrecht, A., and Mundlos, S. (2005). The other trinucleotide repeat: Polyalanine expansion disorders. *Curr. Opin. Genet. Dev.* 15, 285–293. doi:10.1016/j.gde.2005.04.003.

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. 8, 61–65. doi:10.1038/NMETH.1527.

Ameur, A., Kloosterman, W. P., and Hestand, M. S. (2018). Single-Molecule Sequencing: Towards Clinical Applications. *Trends Biotechnol.* doi:10.1016/j.tibtech.2018.07.013.

Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* 46, 2159–2168. doi:10.1093/nar/gky066.

Asagoshi, K., Liu, Y., Masaoka, A., Lan, L., Prasad, R., Horton, J. K., et al. (2010). DNA polymerase β-dependent long patch base excision repair in living cells. *DNA Repair (Amst).* 9, 109–119. doi:10.1016/j.dnarep.2009.11.002.

Bagni, C., and Oostra, B. a (2013). Fragile X syndrome: From protein function to therapy. *Am. J. Med. Genet. A.* doi:10.1002/ajmg.a.36241.

Bahlo, M., Bennett, M. F., Degorski, P., Tankard, R. M., Delatycki, M. B., and Lockhart, P. J. (2018). Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Research* 7, 736. doi:10.12688/f1000research.13980.1.

Barbé, L., Lanni, S., López-Castel, A., Franck, S., Spits, C., Keymolen, K., et al. (2017). CpG Methylation, a Parent-of-Origin Effect for Maternal-Biased Transmission of Congenital Myotonic Dystrophy. *Am. J. Hum. Genet.* 100, 488–505. doi:10.1016/j.ajhg.2017.01.033.

Bergfors, A., Leenheer, D., Bergqvist, A., Ameur, A., and Lennerstrand, J. (2016). Analysis of hepatitis C NS5A resistance associated polymorphisms using ultra deep single molecule real time (SMRT) sequencing. *Antiviral Res.* 126, 81–89. doi:10.1016/j.antiviral.2015.12.005.

Biancalana, V., Glaeser, D., McQuaid, S., and Steinbach, P. (2015). EMQN best practice guidelines for the molecular genetic testing and reporting of fragile X syndrome and other fragile X-associated disorders. *Eur. J. Hum. Genet.* 23, 417–425. doi:10.1038/ejhg.2014.185.

Biasiotto, G., Archetti, S., Di Lorenzo, D., Merola, F., Paiardi, G., Borroni, B., et al. (2017). A PCR-based protocol to accurately size C9orf72 intermediate-length alleles. *Mol. Cell. Probes* 32, 60–64. doi:10.1016/j.mcp.2016.10.008.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., et al. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* 49, 643–650. doi:10.1038/ng.3802.

Biesecker, L. G., and Green, R. C. (2014). Diagnostic Clinical Genome and Exome Sequencing. *N. Engl. J. Med.* 370, 2418–2425. doi:10.1056/NEJMra1312543.

Birge, L. M., Pitts, M. L., Richard, B. H., and Wilkinson, G. S. (2010). Length polymorphism and head shape association among genes with polyglutamine repeats in the stalk-eyed fly, Teleopsis dalmanni. *BMC Evol. Biol.* 10. doi:10.1186/1471-2148-10-227.

Bjelland, S., and Seeberg, E. (2003). Mutagenicity, toxicity and repair of DNA base damage induced by oxidation. in *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 37–80. doi:10.1016/j.mrfmmm.2003.07.002.

Bodega, B., Bione, S., Dalprà, L., Toniolo, D., Ornaghi, F., Vegetti, W., et al. (2006). Influence of intermediate and uninterrupted FMR1 CGG expansions in premature ovarian failure manifestation. *Hum. Reprod.* 21, 952–957. doi:10.1093/humrep/dei432.

Bornman, D., Hester, M., Schuetter, J., Kasoji, M., Minard-Smith, A., Barden, C., et al. (2012). Short-read, high-throughput sequencing technology for STR genotyping. *Biotechniques* Apr, 1–6. doi:10.2144/000113857.

Børsting, C., and Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. *Forensic Sci. Int. Genet.* 18, 78–89. doi:10.1016/j.fsigen.2015.02.002.

Braida, C., Stefanatos, R. K. A., Adam, B., Mahajan, N., Smeets, H. J. M., Niel, F., et al. (2010). Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Hum. Mol. Genet.* 19, 1399–1412. doi:10.1093/hmg/ddq015.

Brasa, S., Mueller, A., Jacquemont, S., Hahne, F., Rozenberg, I., Peters, T., et al. (2016). Reciprocal changes in DNA methylation and hydroxymethylation and a broad repressive epigenetic switch characterize FMR1 transcriptional silencing in fragile X syndrome. *Clin. Epigenetics* 8, 15. doi:10.1186/s13148-016-0181-x.

Bretherick, K. L., Fluker, M. R., and Robinson, W. P. (2005). FMR1 repeat sizes in the gray zone and high end of the normal range are associated with premature ovarian failure. *Hum. Genet.* 117, 376–382. doi:10.1007/s00439-005-1326-8.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat. *Am. J. Hum. Genet.* 62, 1408–1415. doi:10.1086/301869.

Brookes, C., Bright, J. A., Harbison, S., and Buckleton, J. (2012). Characterising stutter in forensic STR multiplexes. *Forensic Sci. Int. Genet.* 6, 58–63. doi:10.1016/j.fsigen.2011.02.001.

Buermans, H. P. J., Vossen, R. H. A. M., Anvar, S. Y., Allard, W. G., Guchelaar, H. J., White, S. J., et al. (2017). Flexible and Scalable Full-Length CYP2D6 Long Amplicon PacBio Sequencing. *Hum. Mutat.* 38, 310–316. doi:10.1002/humu.23166.

Bull, R. A., Eltahla, A. A., Rodrigo, C., Koekkoek, S. M., Walker, M., Pirozyan, M. R., et al. (2016). A method for near full-length amplification and sequencing for six hepatitis C virus genotypes. *BMC Genomics* 17, 247. doi:10.1186/s12864-016-2575-8.

Burlet, P., Frydman, N., Gigarel, N., Kerbrat, V., Tachdjian, G., Feyereisen, E., et al. (2006). Multiple displacement amplification improves PGD for fragile X syndrome. *Mol. Hum. Reprod.* 12, 647–652. doi:10.1093/molehr/gal069.

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., and DePristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* 13, 375. doi:10.1186/1471-2164-13-375.

Cavelier, L., Ameur, A., Häggqvist, S., Höijer, I., Cahill, N., Olsson-Strömberg, U., et al. (2015). Clonal distribution of BCR-ABL1 mutations and splice isoforms by single-molecule long-read RNA sequencing. *BMC Cancer* 15, 45. doi:10.1186/s12885-015-1046-y.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517, 608–611. doi:10.1002/cphc.200.

Chakraborty, S., Vatta, M., Bachinski, L. L., Krahe, R., Dlouhy, S., and Bai, S. (2016). Molecular diagnosis of myotonic dystrophy. *Curr. Protoc. Hum. Genet.* 91, 9.29.1-9.29.19. doi:10.1002/cphg.22.

Chen, L., Hadd, A., Sah, S., Filipovic-Sadic, S., Krosting, J., Sekinger, E., et al. (2010). An information-rich CGG repeat primed PCR that detects the full range of fragile X expanded alleles and minimizes the need for southern blot analysis. *J. Mol. Diagn.* 12, 589–600. doi:10.2353/jmoldx.2010.090227.

Chen, L. S., Tassone, F., Sahota, P., and Hagerman, P. J. (2003). The (CGG)n repeat element within the 5' untranslated region of the FMR1 message provides both positive and negative cis effects on in vivo translation of a downstream reporter. *Hum. Mol. Genet.* 12, 3067–3074. doi:10.1093/hmg/ddg331.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E. G., Wetter, T., et al. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14, 1147–1159. doi:10.1101/gr.1917404.

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi:10.1038/nmeth.4035.

Cho, D. H., Thienes, C. P., Mahoney, S. E., Analau, E., Filippova, G. N., and Tapscott, S. J. (2005). Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. *Mol. Cell* 20, 483–489. doi:10.1016/j.molcel.2005.09.002.

Clark, T. a, Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S. W., et al. (2013). Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.* 11, 4. doi:10.1186/1741-7007-11-4.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4, 265–270. doi:10.1038/nnano.2009.12.

Colak, D., Zaninovic, N., Cohen, M. S., Rosenwaks, Z., Yang, W.-Y., Gerhardt, J., et al. (2014). Promoter-Bound Trinucleotide Repeat mRNA Drives Epigenetic Silencing in Fragile X Syndrome. *Science (80-.).* 343, 1002–1005. doi:10.1126/science.1245831.

Cronister, A., Teicher, J., Rohlfs, E. M., Donnenfeld, A., and Hallam, S. (2008). Prevalence and instability of fragile X alleles: Implications for offering fragile X prenatal diagnosis. *Obstet. Gynecol.* 111, 596–601. doi:10.1097/AOG.0b013e318163be0b.

Dastidar, S., Ardui, S., Singh, K., Majumdar, D., Nair, N., Fu, Y., et al. (2018). Efficient CRISPR / Cas9-mediated editing of trinucleotide repeat expansion in myotonic dystrophy patient-derived iPS and myogenic cells. *Nucleic Acids Res.* 1, 1–24. doi:10.1093/nar/gky548.

De Temmerman, N., Seneca, S., Van Steirteghem, A., Haentjens, P., Van der Elst, J., Liebaers, I., et al. (2008). CTG repeat instability in a human embryonic stem cell line carrying the myotonic dystrophy type 1 mutation. *Mol. Hum. Reprod.* 14, 405–412. doi:10.1093/molehr/gan034.

De Temmerman, N., Sermon, K., Seneca, S., De Rycke, M., Hilven, P., Lissens, W., et al. (2004). Intergenerational instability of the expanded CTG repeat in the DMPK gene: studies in human gametes and preimplantation embryos. *Am. J. Hum. Genet.* 75, 325–329. doi:10.1086/422762.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nat. Biotechnol.* 518. doi:10.1038/nbt.3423.

Devys, D., Biancalana, V., Rousseau, F., Boue, J., Mandel, J. L., and Oberle, I. (1992). Analysis of full fragile X mutations in fetal tissues and monozygotic twins indicate that abnormal methylation and somatic heterogeneity are established early in development. *Am. J. Med. Genet.* 43, 208–216. doi:10.1002/ajmg.1320430134.

Dimitriadou, E., Melotte, C., Debrock, S., Esteki, M. Z., Dierickx, K., Voet, T., et al. (2017). Principles guiding embryo selection following genome-wide haplotyping of preimplantation embryos. *Hum. Reprod.* 32, 687–697. doi:10.1093/humrep/dex011.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., et al. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat. Biotechnol.* 32, 1262–7. doi:10.1038/nbt.3026.

Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., Van Blitterswijk, M., Narzisi, G., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* 27, 1895–1903. doi:10.1101/gr.225672.117.

Doolittle, W. F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284, 601–603. doi:10.1038/284601a0.

Du, J., Campau, E., Soragni, E., Jespersen, C., and Gottesfeld, J. M. (2013). Length-dependent CTG.CAG triplet-repeat expansion in myotonic dystrophy patient-derived induced pluripotent stem cells. *Hum. Mol. Genet.* 22, 5276–5287. doi:10.1093/hmg/ddt386.

Duitama, J., Zablotskaya, A., Gemayel, R., Jansen, A., Belet, S., Verstrepen, K. J., et al. (2014). Large - scale analysis of tandem repeat variability in the human genome. *Nucleic Acids Res.* 42, 5728–5741. doi:10.1093/nar/gku212.

Ebbert, M. T. W., Farrugia, S. L., Sens, J. P., Jansen-West, K., Gendron, T. F., Prudencio, M., et al. (2018). Long-read sequencing across the C9orf72 “GGGGCC” repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol. Neurodegener.* 13, 46. doi:10.1186/s13024-018-0274-4.

Eichler, E. E., Holden, J. J., Popovich, B. W., Reiss, A. L., Snow, K., Thibodeau, S. N., et al. (1994). Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat. Genet.* 8, 88–94. doi:10.1038/ng0994-88.

Eichler, E. E., Macpherson, J. N., Murray, A., Jacobs, P. A., Chakravarti, A., and Nelson, D. L. (1996). Haplotype and interspersion analysis of the FMR1 CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome. *Hum. Mol. Genet.* 5, 319–330. doi:10.1093/hmg/5.3.319.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (80-.).* 323, 133–138. doi:10.1126/science.1162986.

Elbashir, S. M., Harborth, J., Weber, K., and Tuschl, T. (2002). Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods* 26, 199–213. doi:10.1016/S1046-2023(02)00023-3.

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* 5, 435–445. doi:10.1038/nrg1348.

Evans-Galea, M. V., Hannan, A. J., Carrodus, N., Delatycki, M. B., and Saffery, R. (2013a). Epigenetic modifications in trinucleotide repeat diseases. *Trends Mol. Med.* 19, 655–663. doi:10.1016/j.molmed.2013.07.007.

Evans-Galea, M. V, Hannan, A. J., Carrodus, N., Delatycki, M. B., and Saffery, R. (2013b). Epigenetic modifications in trinucleotide repeat diseases. *Trends Mol. Med.*, 1–9. doi:10.1016/j.molmed.2013.07.007.

Feng, Z., Fang, G., Korlach, J., Clark, T., Luong, K., Zhang, X., et al. (2013). Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput. Biol.* 9, e1002935. doi:10.1371/journal.pcbi.1002935.

Fernandez-Carvajal, I., Lopez Posadas, B., Pan, R., Raske, C., Hagerman, P. J., and Tassone, F. (2009). Expansion of an FMR1 grey-zone allele to a full mutation in two generations. *J. Mol. Diagn.* 11, 306–310. doi:10.2353/jmoldx.2009.080174.

Ferreira, S. I., Pires, L. M., Ferrão, J., Sá, J., Serra, A., and Carreira, I. M. (2013). Mosaicism for FMR1 gene full mutation and intermediate allele in a female foetus: A postzygotic retraction event. *Gene* 527, 421–425. doi:10.1016/j.gene.2013.05.079.

Field, A. E., Robertson, N. A., Wang, T., Havas, A., Ideker, T., and Adams, P. D. (2018). DNA Methylation Clocks: Categories, Causes, and Consequences. *Mol. Cell* 71, 882–895. doi:10.1016/j.molcel.2018.08.008.

Filipovic-Sadic, S., Sah, S., Chen, L., Krosting, J., Sekinger, E., Zhang, W., et al. (2010). A novel FMR1 PCR method for the routine detection of low abundance expanded alleles and full mutations in fragile X syndrome. *Clin. Chem.* 56, 399–408. doi:10.1373/clinchem.2009.136101.

Fokkema, I. F. A. C., Taschner, P. E. M., Schaafsma, G. C. P., Celli, J., Laros, J. F. J., and den Dunnen, J. T. (2011). LOVD v.2.0: The next generation in gene variant databases. *Hum. Mutat.* 32, 557–563. doi:10.1002/humu.21438.

Fondon, J. W., and Garner, H. R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci.* 101, 18058–18063. doi:10.1073/pnas.0408118101.

Fondon, J. W., Hammock, E. A. D., Hannan, A. J., and King, D. G. (2008). Simple sequence repeats: genetic modulators of brain function and behavior. *Trends Neurosci.* 31, 328–334. doi:10.1016/j.tins.2008.03.006.

Frenkel, Z. M., and Trifonov, E. N. (2012). Origin and evolution of genes and genomes. Crucial role of triplet expansions. *J. Biomol. Struct. Dyn.* 30, 201–210. doi:10.1080/07391102.2012.677771.

Fu, Y. H., Kuhl, D. P., Pizzuti, a, Pieretti, M., Sutcliffe, J. S., Richards, S., et al. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* 67, 1047–58. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1760838.

Fu, Y. H., Pizzuti, A., Fenwick, R. G., King, J., Rajnarayan, S., Dunne, P. W., et al. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science (80-.).* 255, 1256–1258. doi:10.1126/science.1546326.

Gabriel, C., Danzer, M., Hackl, C., Kopal, G., Hufnagl, P., Hofer, K., et al. (2009). Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. *Hum. Immunol.* 70, 960–964. doi:10.1016/j.humimm.2009.08.009.

Gaj, T., Gersbach, C. a, and Barbas, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* 31, 397–405. doi:10.1016/j.tibtech.2013.04.004.

Gao, R., Matsuura, T., Coolbaugh, M., Zühlke, C., Nakamura, K., Rasmussen, A., et al. (2008). Instability of expanded CAG/CAA repeats in spinocerebellar ataxia type 17. *Eur. J. Hum. Genet.* 16, 215–222. doi:10.1038/sj.ejhg.5201954.

Gao, Z., and Cooper, T. A. (2013). Antisense Oligonucleotides: Rising Stars in Eliminating RNA Toxicity in Myotonic Dystrophy. *Hum. Gene Ther.* 24, 499–507. doi:10.1089/hum.2012.212.

Gemayel, R., Vinces, M. D., Legendre, M., and Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* 44, 445–77. doi:10.1146/annurev-genet-072610-155046.

Genovese, L. M., Geraci, F., Corrado, L., Mangano, E., D'Aurizio, R., Bordoni, R., et al. (2018). A census of tandemly repeated polymorphic loci in genic regions through the comparative integration of human genome assemblies. *Front. Genet.* 9, 155. doi:10.3389/fgene.2018.00155.

Gonitel, R., Moffitt, H., Sathasivam, K., Woodman, B., Detloff, P. J., Faull, R. L. M., et al. (2008). DNA instability in postmitotic neurons. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3467–3472. doi:10.1073/pnas.0800048105.

Green, K. M., Linsalata, A. E., and Todd, P. K. (2016). RAN translation—What makes it run? *Brain Res.* 1647, 30–42. doi:10.1016/j.brainres.2016.04.003.

Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. *Nat. Rev. Genet.* doi:10.1038/nrg1674.

Guan, P., and Sung, W.-K. (2016). Structural variation detection using next-generation sequencing data: A comparative technical review. *Methods* 102, 36–49. doi:http://dx.doi.org/10.1016/j.ymeth.2016.01.020.

Guo, X., Zheng, S., Dang, H., Pace, R. G., Stonebraker, J. R., Jones, C. D., et al. (2014a). Genome Reference and Sequence Variation in the Large Repetitive Central Exon of Human MUC5AC. *Am. J. Respir. Cell Mol. Biol.* 50, 223–232. doi:10.1165/rcmb.2013-0235OC.

Guo, X., Zheng, S., Dang, H., Pace, R. G., Stonebraker, J. R., Jones, C. D., et al. (2014b). Genome reference and sequence variation in the large repetitive central Exon of human MUC5AC. *Am. J. Respir. Cell Mol. Biol.* 50, 223–232. doi:10.1165/rcmb.2013-0235OC.

Hagerman, P. J., and Hagerman, R. J. (2015). Fragile X-associated tremor/ataxia syndrome. *Ann. N. Y. Acad. Sci.* 1338, 58–70. doi:10.1111/nyas.12693.

Hall, D. A., Berry-Kravis, E., Zhang, W., Tassone, F., Spector, E., Zerbe, G., et al. (2011). FMR1 gray-zone alleles: Association with Parkinson's disease in women? *Mov. Disord.* 26, 1900–1906. doi:10.1002/mds.23755.

Hall, D. A., Tassone, F., Klepitskaya, O., and Leehey, M. (2012). Fragile X-Associated Tremor Ataxia Syndrome in FMR1 Gray Zone Allele Carriers. *Mov. Disord.* 27, 296–300. doi:10.1002/mds.24021.

Hamada, H., Seidman, M., Howard, B. H., and Gorman, C. M. (1984). Enhanced Gene Expression by the Poly(dT-dG) Poly(dC-dA) Sequence. *Mol. Cell. Biol.* 4, 2622–2630. Available at: http://mcb.asm.org/ [Accessed December 15, 2018].

Hannan, A. J. (2018a). Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* 19, 286–298. doi:10.1038/nrg.2017.115.

Hannan, A. J. (2018b). Tandem repeats mediating genetic plasticity in health and disease. doi:10.1038/nrg.2017.115.

Hayward, B. E., and Usdin, K. (2017). Improved Assays for AGG Interruptions in Fragile X Premutation Carriers. *J. Mol. Diagnostics* 19, 828–835. doi:10.1016/j.jmoldx.2017.06.008.

He, F., and Todd, P. (2011). Epigenetic Mechanisms in Repeat Expansion Disorders. *Semin. Neurol.* 31, 470–483. doi:10.1055/s-0031-1299786.Epigenetic.

Heather, J. M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* 107, 1–8. doi:10.1016/j.ygeno.2015.11.003.

Hegan, D. C., Narayanan, L., Jirik, F. R., Edelmann, W., Liskay, R. M., and Glazer, P. M. (2006). Differing patterns of genetic instability in mice deficient in the mismatch repair genes Pms2, Mlh1, Msh2, Msh3 and Msh6. *Carcinogenesis* 27, 2402–2408. doi:10.1093/carcin/bgl079.

Hestand, M. S., Houdt, J. Van, Cristofoli, F., and Vermeesch, J. R. (2016a). Polymerase Specific Error Rates and Profiles Identified by Single Molecule Sequencing. *Mutat. Res.* 784–785, 39–45. doi:10.1016/j.mrfmmm.2016.01.003.

Hestand, M. S., Van Houdt, J., Cristofoli, F., and Vermeesch, J. R. (2016b). Polymerase specific error rates and profiles identified by single molecule sequencing. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* 784–785, 39–45. doi:10.1016/j.mrfmmm.2016.01.003.

Höijer, I., Tsai, Y. C., Clark, T. A., Kotturi, P., Dahl, N., Stattin, E. L., et al. (2018). Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. *Hum. Mutat.* 39, 1262–1272. doi:10.1002/humu.23580.

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

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of Bacteria and Archaea. *Science (80-.).* 327, 167–170. doi:10.1126/science.1179555.

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* doi:10.1038/nbt.2647.

Hu, H., Haas, S. A., Chelly, J., Van Esch, H., Raynaud, M., De Brouwer, A. P. M., et al. (2016). X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Mol. Psychiatry* 21, 133–148. doi:10.1038/mp.2014.193.

Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M. K., Fujiyama, A., et al. (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* 50, 581–590. doi:10.1038/s41588-018-0067-2.

Jacquemont, S., Birnbaum, S., Redler, S., Steinbach, P., and Biancalana, V. (2011). Clinical utility gene card for: fragile X mental retardation syndrome, fragile X-associated tremor/ataxia syndrome and fragile X-associated primary ovarian insufficiency. *Eur. J. Hum. Genet.* 19, doi:10.1038/ejhg.2011.55. doi:10.1038/ejhg.2011.55.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 239. doi:10.1186/s13059-016-1122-x.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Hypervariable “minisatellite” regions in human DNA. *Nature* 314, 67–73. doi:10.1038/314067a0.

Jelinek, W. R., Toomey, T. P., Leinwand, L. A., Duncan, C. H., Biro, P. A., Choudary, P. V, et al. (1980). Ubiquitous, interspersed repeated sequences in mammalian genomes. *Proc. Natl. Acad. Sci. U. S. A.* 77, 1398–402. doi:10.1073/pnas.77.3.1398.

Jia, H., Guo, Y., Zhao, W., and Wang, K. (2014). Long-range PCR in next-generation sequencing: Comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci. Rep.* 4, 5737. doi:10.1038/srep05737.

Jiao, X., Zheng, X., Ma, L., Kutty, G., Gogineni, E., Sun, Q., et al. (2013). A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. *J. Data Min. Genomics Proteomics* 4, pii16008. doi:10.4172/2153-0602.1000136.A.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. a, and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–21. doi:10.1126/science.1225829.

Jiraanont, P., Kumar, M., Tang, H.-T., Espinal, G., Hagerman, P. J., Hagerman, R. J., et al. (2017). Size and methylation mosaicism in males with Fragile X syndrome. *Expert Rev. Mol. Diagn.* 17, 1023–1032. doi:10.1080/14737159.2017.1377612.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32, 493–496. doi:10.1093/nar/gkh103.

Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P., and Siksnys, V. (2013a). crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. *RNA Biol.* 10, 841–51. doi:10.4161/rna.24203.

Karvelis, T., Gasiunas, G., and Siksnys, V. (2013b). Programmable DNA cleavage in vitro by Cas9. *Biochem. Soc. Trans.* 41, 1401–1406. doi:10.1042/BST20130164.

Kashi, Y., and King, D. G. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* 22, 253–259. doi:10.1016/j.tig.2006.03.005.

Katsanis, S. H., and Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* 14, 415–426. doi:10.1038/nrg3493.

Ke, R., Mignardi, M., Hauling, T., and Nilsson, M. (2016). Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Hum. Mutat.* 37, 1363–1367. doi:10.1002/humu.23051.

Kennedy, L., Evans, E., Chen, C. M., Craven, L., Detloff, P. J., Ennis, M., et al. (2003). Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum. Mol. Genet.* 12, 3359–3367. doi:10.1093/hmg/ddg352.

King, D. G., Soller, M., and Kashi, Y. (1997). Evolutionary tuning knobs. *Endeavour* 21, 36–40. doi:10.1016/S0160-9327(97)01005-3.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). XThe next-generation sequencing revolution and its impact on genomics. *Cell* 155, 27–38. doi:10.1016/j.cell.2013.09.006.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 693–700. doi:10.1038/nbt.2280.

Kroutil, L. C., Register, K., Bebenek, K., and Kunkel, T. A. (1996). Exonucleolytic proofreading during replication of repetitive DNA. *Biochemistry* 35, 1046–1053. doi:10.1021/bi952178h.

Kumari, D., and Usdin, K. (2009). Chromatin remodeling in the noncoding repeat expansion diseases. *J. Biol. Chem.* 284, 7413–7417. doi:10.1074/jbc.R800026200.

Kunst, C. B., and Warren, S. T. (1994). Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* 77, 853–861. doi:10.1016/0092-8674(94)90134-1.

Lander, E. S. (2016). The Heroes of CRISPR. *Cell* 164, 18–28. doi:10.1016/j.cell.2015.12.041.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062.

Lang, W. H., Coats, J. E., Majka, J., Hura, G. L., Lin, Y., Rasnik, I., et al. (2011). Conformational trapping of mismatch recognition complex MSH2/MSH3 on repair-resistant DNA loops. *Proc. Natl. Acad. Sci. U. S. A.* 108, E837-44. doi:10.1073/pnas.1105461108.

Lee, D. Y., and McMurray, C. T. (2014). Trinucleotide expansion in disease: Why is there a length threshold? *Curr. Opin. Genet. Dev.* 26, 131–140. doi:10.1016/j.gde.2014.07.003.

Lee, J. K., Conrad, A., Epping, E., Mathews, K., Magnotta, V., Dawson, J. D., et al. (2018). Effect of Trinucleotide Repeats in the Huntington's Gene on Intelligence. *EBioMedicine* 31, 47–53. doi:10.1016/j.ebiom.2018.03.031.

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K. J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.* 17, 1787–1796. doi:10.1101/gr.6554007.

Levesque, S., Dombrowski, C., Morel, M. L., Rehel, R., Côté, J. S., Bussìeres, J., et al. (2009). Screening and instability of FMR1 alleles in a prospective sample of 24,449 mother– newborn pairs from the general population. *Clin. Genet.* 76, 511–523. doi:10.1111/j.1399-0004.2009.01237.x.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Liljegren, M. M., de Muinck, E. J., and Trosvik, P. (2016). Microsatellite Length Scoring by Single Molecule Real Time Sequencing – Effects of Sequence Structure and PCR Regime. *PLoS One* 11, e0159232. doi:10.1371/journal.pone.0159232.

Lin, Y., and Wilson, J. H. (2007). Transcription-Induced CAG Repeat Contraction in Human Cells Is Mediated in Part by Transcription-Coupled Nucleotide Excision Repair. *Mol. Cell. Biol.* 27, 6209–6217. doi:10.1128/MCB.00739-07.

Liu, Q., Zhang, P., Wang, D., Gu, W., and Wang, K. (2017a). Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med.* 9, 65. doi:10.1186/s13073-017-0456-7.

Liu, Q., Zhang, P., Wang, D., Gu, W., and Wang, K. (2017b). Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med.* 9. doi:10.1186/s13073-017-0456-7.

Liu, Y., Tao, W., Wen, S., Li, Z., Yang, A., Deng, Z., et al. (2015). In vitro CRISPR/cas9 system for efficient targeted DNA editing. *MBio* 6, 1714–1729. doi:10.1128/mBio.01714-15.

Liu, Y., and Wilson, S. H. (2012). DNA base excision repair: a mechanism of trinucleotide repeat expansion. *Trends Biochem. Sci.* 37, 162–72. doi:10.1016/j.tibs.2011.12.002.

Liu, Y., Winarni, T., Zhang, L., Tassone, F., and Hagerman, R. (2013). Fragile X-associated tremor/ataxia syndrome (FXTAS) in grey zone carriers. *Clin. Genet.* 84, 74–77. doi:10.1111/cge.12026.

Loesch, D. Z., Khaniani, M. S., Slater, H. R., Rubio, J. P., Bui, Q. M., Kotschet, K., et al. (2009). Small CGG repeat expansion alleles of FMR1 gene are associated with parkinsonism. *Clin. Genet.* 76, 471–476. doi:10.1111/j.1399-0004.2009.01275.x.

Loomis, E. W., Eid, J. S., Peluso, P., Yin, J., Hickey, L., Rank, D., et al. (2013). Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* 23, 121–128. doi:10.1101/gr.141705.112.

López-Morató, M., Brook, J. D., and Wojciechowska, M. (2018). Small molecules which improve pathogenesis of myotonic dystrophy type 1. *Front. Neurol.* 9, 349. doi:10.3389/fneur.2018.00349.

López Castel, A., Cleary, J. D., and Pearson, C. E. (2010). Repeat instability as the basis for human diseases and as a potential target for therapy. 11. Available at: http://dx.doi.org/10.1038/nrm2854.

Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics. Proteomics Bioinformatics* 14, 265–279. doi:10.1016/j.gpb.2016.05.004.

Madrigal, I., Xunclà, M., Tejada, M. I., Martínez, F., Fernández-Carvajal, I., Pérez-Jurado, A., et al. (2011). Intermediate FMR1 alleles and cognitive and/or behavioural phenotypes. *Eur. J. Hum. Genet.* 19, 921–923. doi:10.1038/ejhg.2011.41.

Maffioletti, S. M., Gerli, M. F. M., Ragazzi, M., Dastidar, S., Benedetti, S., Loperfido, M., et al. (2015). Efficient derivation and inducible differentiation of expandable skeletal myogenic cells from human ES and patient-specific iPS cells. *Nat. Protoc.* 10, 941–958. doi:10.1038/nprot.2015.057.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., et al. (1992). Myotonic dystrophy mutation: An unstable CTG repeat in the 3′ untranslated region of the gene. *Science (80-.).* 255, 1253–1255. doi:10.1126/science.1546325.

Mali, P., Esvelt, K. M., and Church, G. M. (2013). Cas9 as a versatile tool for engineering biology. *Nat. Methods* 10, 957–63. doi:10.1038/nmeth.2649.

Martin, L. J. (2008). DNA damage and repair: Relevance to mechanisms of neurodegeneration. *J. Neuropathol. Exp. Neurol.* 67, 377–387. doi:10.1097/NEN.0b013e31816ff780.

Martorell, L., Johnson, K., Boucher, C. A., and Baiget, M. (1997). Somatic instability of the myotonic dystrophy (CTG), repeat during human fetal development. *Hum. Mol. Genet.* 6, 877–880.

Matsuura, T., Fang, P., Pearson, C. E., Jayakar, P., Ashizawa, T., Roa, B. B., et al. (2006). Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: repeat purity as a disease modifier? *Am J Hum Genet* 78, 125–129. doi:10.1086/498654.

McFarland, K. N., Liu, J., Landrian, I., Godiska, R., Shanker, S., Yu, F., et al. (2015). SMRT sequencing of long tandem nucleotide repeats in SCA10 reveals unique insight of repeat expansion structure. *PLoS One* 10, 1–13. doi:10.1371/journal.pone.0135906.

McMurray, C. T. (2010). Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.* 11, 786–799. doi:10.1038/nrg2828.

Menon, R. P., Nethisinghe, S., Faggiano, S., Vannocci, T., Rezaei, H., Pemble, S., et al. (2013). The Role of Interruptions in polyQ in the Pathology of SCA1. *PLoS Genet.* 9, e1003648. doi:10.1371/journal.pgen.1003648. doi:10.1371/journal.pgen.1003648.

Meola, G., and Cardani, R. (2014). Myotonic dystrophies: An update on clinical aspects, genetic, pathology, and molecular pathomechanisms ☆. doi:10.1016/j.bbadis.2014.05.019.

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

Mitsuhashi, S., Nakagawa, S., Takahashi Ueda, M., Imanishi, T., Frith, M. C., and Mitsuhashi, H. (2017). Nanopore-based single molecule sequencing of the D4Z4 array responsible for facioscapulohumeral muscular dystrophy. *Sci. Rep.* 7, 14789. doi:10.1038/s41598-017-13712-6.

Miyoshi-Akiyama, T., Satou, K., Kato, M., Shiroma, A., Matsumura, K., Tamotsu, H., et al. (2015). Complete annotated genome sequence of Mycobacterium tuberculosis (Zopf) Lehmann and Neumann (ATCC35812) (Kurono). *Tuberculosis* 95, 37–39. doi:10.1016/j.tube.2014.10.007.

Moll, K. M., Zhou, P., Ramaraj, T., Fajardo, D., Devitt, N. P., Sadowsky, M. J., et al. (2017). Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. *BMC Genomics* 18, 578. doi:10.1186/s12864-017-3971-4.

Monaghan, K. G., Lyon, E., and Spector, E. B. (2013). ACMG Standards and Guidelines for fragile X testing: a revision to the disease-specific supplements to the Standards and Guidelines for Clinical Genetics Laboratories of the American College of Medical Genetics and Genomics. *Genet. Med.* 15, 575–586. doi:10.1038/gim.2013.61.

Mononen, T., Von Koskull, H., Airaksinen, R. L., and Juvonen, V. (2007). A novel duplication in the FMR1 gene: Implications for molecular analysis in fragile X syndrome and repeat instability. *Clin. Genet.* 72, 528–531. doi:10.1111/j.1399-0004.2007.00903.x.

Morales, F., Couto, J. M., Higham, C. F., Hogg, G., Cuenca, P., Braida, C., et al. (2012). Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Hum. Mol. Genet.* 21, 3558–3567. doi:10.1093/hmg/dds185.

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

Mulero, J. J., Chang, C. W., and Hennessy, L. K. (2006). Characterization of the N+3 stutter product in the trinucleotide repeat locus DYS392. *J. Forensic Sci.* 51, 1069–1073. doi:10.1111/j.1556-4029.2006.00227.x.

Musova, Z., Mazanec, R., Krepelova, A., Ehler, E., Vales, J., Jaklova, R., et al. (2009). Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. *Am. J. Med. Genet. Part A* 149A, 1365–1369. doi:10.1002/ajmg.a.32987.

Nakamori, M., and Thornton, C. (2010). Epigenetic changes and non-coding expanded repeats. *Neurobiol. Dis.* 39, 21–27. doi:10.1016/j.nbd.2010.02.004.

Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., et al. (2017a). Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell*. doi:10.1007/s13577-017-0168-8.

Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., et al. (2017b). Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* 30, 1–13. doi:10.1007/s13577-017-0168-8.

Nakatani, R., Nakamori, M., Fujimura, H., and Mochizuki, H. (2015). Large expansion of CTG • CAG repeats is exacerbated by MutS β in human cells. *Nat. Publ. Gr.*, 1–11. doi:10.1038/srep11020.

Natesan, S. A., Bladon, A. J., Coskun, S., Qubbaj, W., Prates, R., Munne, S., et al. (2014). Genome-wide karyomapping accurately identifies the inheritance of single-gene defects in human preimplantation embryos in vitro. *Genet. Med.* 16, 838–845. doi:10.1038/gim.2014.45.

Nattestad, M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F. J., Rescheneder, P., et al. (2018). Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* 28, 1126–1135. doi:10.1101/gr.231100.117.

Nithianantharajah, J., and Hannan, A. J. (2007). Dynamic mutations as digital genetic modulators of brain development, function and dysfunction. *BioEssays* 29, 525–535. doi:10.1002/bies.20589.

Nolin, S. L., Brown, W. T., Glicksman, A., Houck, G. E., Gargano, A. D., Sullivan, A., et al. (2003a). Expansion of the fragile X CGG repeat in females with premutation or intermediate alleles. *Am. J. Hum. Genet.* 72, 454–464. doi:10.1086/367713.

Nolin, S. L., Brown, W. T., Glicksman, A., Houck, G. E., Gargano, A. D., Sullivan, A., et al. (2003b). Expansion of the fragile X CGG repeat in females with premutation or intermediate alleles. *Am. J. Hum. Genet.* 72, 454–464. doi:10.1086/367713.

Nolin, S. L., Glicksman, A., Ding, X., Ersalesi, N., Brown, W. T., Sherman, S. L., et al. (2011). Fragile X analysis of 1112 prenatal samples from 1991 to 2010. 925–931. doi:10.1002/pd.

Nolin, S. L., Glicksman, A., Ersalesi, N., Dobkin, C., Brown, W. T., Cao, R., et al. (2015). Fragile X full mutation expansions are inhibited by one or more AGG interruptions in premutation carriers. *Genet. Med.* 17, 358–364. doi:10.1038/gim.2014.106.

Nolin, S. L., Houck, G. E., Gargano, a D., Blumstein, H., Dobkin, C. S., and Brown, W. T. (1999). FMR1 CGG-repeat instability in single sperm and lymphocytes of fragile-X premutation males. *Am. J. Hum. Genet.* 65, 680–688. doi:10.1086/302543.

Nolin, S. L., Sah, S., Glicksman, A., Sherman, S. L., Allen, E., Berry-Kravis, E., et al. (2013). Fragile X AGG analysis provides new risk predictions for 45-69 repeat alleles. *Am. J. Med. Genet. Part A* 161, 771–778. doi:10.1002/ajmg.a.35833.

Oberlé, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., et al. (1991). Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* 252, 1097–1102. doi:10.1126/science.252.5009.1097.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symp. Biol.* 23, 366–370. doi:citeulike-article-id:3483106.

Orr, H. T., and Zoghbi, H. Y. (2007). Trinucleotide Repeat Disorders - annurev.neuro.29.051605.113042. *Annu. Rev. Neurosci.* 30, 575–623. doi:10.1146/annurev.neuro.29.051605.113042.

Owen, B. A. L., Yang, Z., Lai, M., Gajec, M., Gajek, M., Badger, J. D., et al. (2005). (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nat. Struct. Mol. Biol.* 12, 663–70. doi:10.1038/nsmb965.

Panigrahi, G. B., Slean, M. M., Simard, J. P., Gileadi, O., and Pearson, C. E. (2010). Isolated short CTG/CAG DNA slip-outs are repaired efficiently by hMutSbeta, but clustered slip-outs are poorly repaired. *Proc Natl Acad Sci U S A* 107, 12593–12598. doi:0909087107 [pii]\r10.1073/pnas.0909087107.

Paques, F., Leung, W.-Y., and Haber, J. E. (1998). Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair Downloaded from. *Mol. Cell. Biol.* 18, 2045–2054. Available at: http://mcb.asm.org/ [Accessed December 15, 2018].

Park, C. Y., Halevy, T., Lee, D. R., Sung, J. J., Lee, J. S., Yanuka, O., et al. (2015). Reversion of FMR1 Methylation and Silencing by Editing the Triplet Repeats in Fragile X iPSC-Derived Neurons. *Cell Rep.* 13, 234–241. doi:10.1016/j.celrep.2015.08.084.

Pearson, C. E. (2011). Repeat associated non-ATG translation initiation: one DNA, two transcripts, seven reading frames, potentially nine toxic entities! *PLoS Genet.* 7, e1002018. doi:10.1371/journal.pgen.1002018.

Pearson, C. E., Edamura, K. N., and Cleary, J. D. (2005a). Repeat instability: Mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6, 729–742. doi:10.1038/nrg1689.

Pearson, C. E., Nichol Edamura, K., and Cleary, J. D. (2005b). Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6, 729–42. doi:10.1038/nrg1689.

Penagarikano, O., Mulle, J. G., and Warren, S. T. (2007). The pathophysiology of fragile x syndrome. *Annu. Rev. Genomics Hum. Genet.* 8, 109–129. doi:10.1146/annurev.genom.8.080706.092249.

Pham, T. T., Yin, J., Eid, J. S., Adams, E., Lam, R., Turner, S. W., et al. (2016). Single-locus enrichment without amplification for sequencing and direct detection of epigenetic modifications. *Mol. Genet. Genomics* 291, 1–14. doi:10.1007/s00438-016-1167-2.

Pieretti, M., Zhang, F., Fu, Y.-H., Warren, S. T., Oostra, B. A., Caskey, C. T., et al. (1991). Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* 66, 817–822. doi:10.1016/0092-8674(91)90125-I.

Platteau, P., Sermon, K., Seneca, S., Van Steirteghem, A., Devroey, P., and Liebaers, I. (2002). Preimplantation genetic diagnosis for fragile Xa syndrome: Difficult but not impossible. *Hum. Reprod.* 17, 2807–2812. doi:10.1093/humrep/17.11.2807.

Pratte, A., Prévost, C., Puymirat, J., and Mathieu, J. (2015). Anticipation in myotonic dystrophy type 1 parents with small CTG expansions. *Am. J. Med. Genet. Part A* 167, 708–714. doi:10.1002/ajmg.a.36950.

Pretto, D. I., Mendoza-Morales, G., Lo, J., Cao, R., Hadd, A., Latham, G. J., et al. (2014a). CGG allele size somatic mosaicism and methylation in FMR1 premutation alleles. *J. Med. Genet.* 51, 309–18. doi:10.1136/jmedgenet-2013-102021.

Pretto, D., Yrigollen, C. M., Tang, H.-T., Williamson, J., Espinal, G., Iwahashi, C. K., et al. (2014b). Clinical and molecular implications of mosaicism in FMR1 full mutations. *Front. Genet.* 5, 1–11. doi:10.3389/fgene.2014.00318.

Putiri, E. L., and Robertson, K. D. (2011). Epigenetic mechanisms and genome stability. *Clin. Epigenetics* 2, 299–314. doi:10.1007/s13148-010-0017-z.

Qiao, W., Yang, Y., Sebra, R., Mendiratta, G., Gaedigk, A., Desnick, R. J., et al. (2016). Long-read single-molecule real-time (SMRT) full gene sequencing of cytochrome P450-2D6 (CYP2D6) HHS Public Access. *Hum Mutat* 37, 315–323. doi:10.1002/humu.22936.

Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., et al. (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* 44, 3750–3762. doi:10.1093/nar/gkw219.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033.

Radvansky, J., Ficek, A., Minarik, G., Palffy, R., and Kadasi, L. (2011). Effect of unexpected sequence interruptions to conventional PCR and repeat primed PCR in myotonic dystrophy type 1 testing. *Diagn. Mol. Pathol.* 20, 48–51. doi:10.1097/PDM.0b013e3181efe290.

Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M., et al. (2017). Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* 14, 411–413. doi:10.1038/nmeth.4189.

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., et al. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 72, 257–68. doi:10.1016/j.neuron.2011.09.010.

Rhoads, A., and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* 13, 278–289. doi:10.1016/j.gpb.2015.08.002.

Richard, G. F., and Pâques, F. (2000). Mini- and microsatellite expansions: The recombination connection. *EMBO Rep.* 1, 122–126. doi:10.1093/embo-reports/kvd031.

Rifé, M., Badenas, C., Quintó, L., Puigoriol, E., Tazón, B., Rodriguez-Revenga, L., et al. (2004). Analysis of CGG variation through 642 meioses in Fragile X families. *Mol. Hum. Reprod.* 10, 773–776. doi:10.1093/molehr/gah102.

Riley, D., and Krieger, J. (2009). Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. *Gene* 15, 997–1003. doi:10.1016/j.biotechadv.2011.08.021.Secreted.

Roeck, A. De, Coster, W. De, Bossaerts, L., Cacace, R., Pooter, T. De, Dongen, J. Van, et al. (2018). Accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *bioRxiv*. doi:10.1101/439026.

Santa María, L., Pugin, a, Alliende, M., Aliaga, S., Curotto, B., Aravena, T., et al. (2013). FXTAS in an unmethylated mosaic male with fragile X syndrome from Chile. *Clin. Genet.*, 1–5. doi:10.1111/cge.12278.

Satou, K., Shimoji, M., Tamotsu, H., Juan, A., Ashimine, N., Shinzato, M., et al. (2015). Complete Genome Sequences of Low-Passage Virulent and High-Passage Avirulent Variants of Pathogenic Leptospira interrogans Serovar Manilae Strain UP-MMC-NIID, Originally Isolated from a Patient with Severe Leptospirosis, Determined Using PacBio Single-Mol. *Genome Announc.* 3, e00882-15. doi:10.1128/genomeA.00882-15.

Saverio Tedesco, F., M Gerli, M. F., Perani, L., Benedetti, S., Ungaro, F., Cassano, M., et al. (2012). Transplantation of Genetically Corrected Human iPSC-Derived Progenitors in Mice with Limb-Girdle Muscular Dystrophy. *Transl. Med.* 4, 140ra189. Available at: www.ScienceTranslationalMedicine.org [Accessed September 28, 2018].

Savić Pavićević, D., Miladinović, J., Brkušanin, M., Šviković, S., Djurica, S., Brajušković, G., et al. (2013). Molecular genetics and genetic testing in myotonic dystrophy type 1. *Biomed Res. Int.* 2013, 391821. doi:10.1155/2013/391821.

Savouret, C., Brisson, E., Essers, J., Kanaar, R., Pastink, A., Te Riele, H., et al. (2003). CTG repeat instability and size variation timing in DNA repair-deficient mice. *EMBO J.* 22, 2264–2273. doi:10.1093/emboj/cdg202.

Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M. A., et al. (2013). Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. *PLoS One* 8, 54710. doi:10.1371/journal.pone.0054710.

Schadt, E. E., Banerjee, O., Fang, G., Feng, Z., Wong, W. H., Zhang, X., et al. (2013). Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.* 23, 129–141. doi:10.1101/gr.136739.111.

Schmidt, M. H. M., and Pearson, C. E. (2016a). Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst).* 38, 117–126. doi:10.1016/j.dnarep.2015.11.008.

Schmidt, M. H. M., and Pearson, C. E. (2016b). Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst).* 38, 117–126. doi:10.1016/j.dnarep.2015.11.008.

Schüle, B., McFarland, K. N., Lee, K., Tsai, Y.-C., Nguyen, K.-D., Sun, C., et al. (2017). Parkinson's disease associated with pure ATXN10 repeat expansion. *npj Park. Dis.* 3, 27. doi:10.1038/s41531-017-0029-x.

Seisenberger, S., Peat, J. R., Hore, T. A., Santos, F., Dean, W., and Reik, W. (2013). Reprogramming DNA methylation in the mammalian life cycle: Building and breaking epigenetic barriers. *Philos. Trans. R. Soc. B Biol. Sci.* 368. doi:10.1098/rstb.2011.0330.

Seneca, S., Lissens, W., Endels, K., Caljon, B., Bonduelle, M., Keymolen, K., et al. (2012). Reliable and sensitive detection of fragile X (expanded) alleles in clinical prenatal DNA samples with a fast turnaround time. *J. Mol. Diagn.* 14, 560–8. doi:10.1016/j.jmoldx.2012.05.003.

Seo, J., Rhie, A., Kim, J., Lee, S., Sohn, M., Kim, C.-U., et al. (2016). De novo assembly and phasing of a Korean human genome. *Nature* 538, 243–247. doi:10.1038/nature20098.

Seriola, A., Spits, C., Simard, J. P., Hilven, P., Haentjens, P., Pearson, C. E., et al. (2011). Huntington's and myotonic dystrophy hESCs: Down-regulated trinucleotide repeat instability and mismatch repair machinery expression bupon differentiation. *Hum. Mol. Genet.* 20, 176–185. doi:10.1093/hmg/ddq456.

Sermon, K., Seneca, S., Vanderfaeillie, A., Lissens, W., Joris, H., Vandervorst, M., et al. (1999). Preimplantation diagnosis for fragile X syndrome based on the detection of the non-expanded paternal and maternal CGG. *Prenat. Diagn.* 19, 1223–1230. doi:10.1002/(SICI)1097-0223(199912)19:13<1223::AID-PD724>3.0.CO;2-0.

Sharma, V., Chow, H. Y., Siegel, D., and Wurmbach, E. (2017). Qualitative and quantitative assessment of Illumina's forensic STR and SNP kits on MiSeq FGx^{TM}. *PLoS One* 12, e0187932. doi:10.1371/journal.pone.0187932.

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* 7, 12065. doi:10.1038/ncomms12065.

Shin, J. W., and Lee, J.-M. (2018). The prospects of CRISPR-based genome engineering in the treatment of neurodegenerative disorders. *Ther. Adv. Neurol. Disord. Rev.* 11, 1–11. doi:10.1177/https.

Shin, S. C., Ahn, D. H., Kim, S. J., Lee, H., Oh, T.-J., Lee, J. E., et al. (2013). Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. *PLoS One* 8, e68824. doi:10.1371/journal.pone.0068824. doi:10.1371/journal.pone.0068824.

Sia, E. A., Jinks-Robertson, S., and Petes, T. D. (1997). Genetic control of microsatellite stability. *Mutat. Res. - DNA Repair* 383, 61–70. doi:10.1016/S0921-8777(96)00046-8.

Singh, S., Zhang, A., Dlouhy, S., and Bai, S. (2014). Detection of large expansions in myotonic dystrophy type 1 using triplet primed PCR. *Front. Genet.* 5, 1–6. doi:10.3389/fgene.2014.00094.

Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science (80-.).* 191, 528–535. doi:10.1126/science.1251186.

Sofola, O. A., Jin, P., Qin, Y., Duan, R., Liu, H., de Haro, M., et al. (2007). RNA binding proteins hnRNP A2/B1 and CUGBP1 suppress Fragile X CGG premutation repeat-induced neurodegeneration in a Drosophila model of FXTAS. *Neuron* 55, 565–571. doi:10.1016/j.dci.2009.07.003.Characterization.

Spiro, C., Pelletier, R., Rolfsmeier, M. L., Dixon, M. J., Lahue, R. S., Gupta, G., et al. (1999). Inhibition of FEN-1 Processing by DNA Secondary Structure at Trinucleotide Repeats at least in part, from improper DNA secondary structure during replication and/or re-pair. Recent studies have shown that trinucleotide re-peats associated with human disea. Available at: https://ac.els-cdn.com/S1097276500802361/1-s2.0-S1097276500802361-main.pdf?_tid=cc11f82b-6a28-402b-afd3-6a24a6a8ccd0&acdnat=1538042876_6cccc0771ba77b7be276dd9024d398c2 [Accessed September 27, 2018].

Staresincic, L., Fagbemi, A. F., Enzlin, J. H., Gourdin, A. M., Wijgers, N., Dunand-Sauthier, I., et al. (2009). Coordination of dual incision and repair synthesis in human nucleotide excision repair. *EMBO J.* 28, 1111–1120. doi:10.1038/emboj.2009.49.

Stern, A., Brown, M., Nickel, P., and Meyer, T. F. (1986). Opacity genes in Neisseria gonorrhoeae: Control of phase and antigenic variation. *Cell* 47, 61–71. doi:10.1016/0092-8674(86)90366-1.

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., and Doudna, J. a (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* 507, 62–67. doi:10.1038/nature13011.

Sullivan, A. K., Crawford, D. C., Scott, E. H., Leslie, M. L., and Sherman, S. L. (2002). Paternally transmitted FMR1 alleles are less stable than maternally transmitted alleles in the common and intermediate size range. *Am. J. Hum. Genet.* 70, 1532–1544. doi:10.1086/340846.

Sullivan, A. K., Marcus, M., Epstein, M. P., Allen, E. G., Anido, A. E., Paquin, J. J., et al. (2005). Association of FMR1 repeat size with ovarian dysfunction. *Hum. Reprod.* 20, 402–412. doi:10.1093/humrep/deh635.

Swami, M., Hendricks, A. E., Gillis, T., Massood, T., Mysore, J., Myers, R. H., et al. (2009). Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.* 18, 3039–3047. doi:10.1093/hmg/ddp242.

Tachida, H., and Iizuka, M. (1992). Persistence of repeated sequences that evolve by replication slippage. *Genetics* 131, 471–478. Available at: http://www.genetics.org/content/genetics/131/2/471.full.pdf [Accessed December 15, 2018].

Tang, H., Kirkness, E. F., Lippert, C., Biggs, W. H., Fabani, M., Guzman, E., et al. (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am. J. Hum. Genet.* 101, 700–715. doi:10.1016/j.ajhg.2017.09.013.

Tassone, F., Beilina, A., Carosi, C., Albertosi, S., Bagni, C., Li, L., et al. (2007). Elevated FMR1 mRNA in premutation carriers is due to increased transcription. *RNA* 13, 555–562. doi:10.1261/rna.280807.

Tassone, F., Iong, K. P., Tong, T.-H., Lo, J., Gane, L. W., Berry-Kravis, E., et al. (2012a). FMR1 CGG allele size and prevalence ascertained through newborn screening in the United States. *Genome Med.* 4, 100. doi:10.1186/gm401.

Tassone, F., Iong, K. P., Tong, T. H., Lo, J., Gane, L. W., Berry-Kravis, E., et al. (2012b). FMR1 CGG allele size and prevalence ascertained through newborn screening in the United States. *Genome Med.* 4, 1–13. doi:10.1186/gm401.

Terracciano, A., Pomponi, M. G., Maria, G., Marino, E., Chiurazzi, P., Rinaldi, M. M., et al. (2004). Expansion to full mutation of a FMR1 intermediate allele over two generations. *Eur. J. Hum. Genet.* 12, 333–336. doi:10.1038/sj.ejhg.5201154.

Theadom, A., Rodrigues, M., Roxburgh, R., Balalla, S., Higgins, C., Bhattacharjee, R., et al. (2014). Prevalence of Muscular Dystrophies: A Systematic Literature Review. *Neuroepidemiology* 43, 259–268. doi:10.1159/000369343.

Tirosh, I., Barkai, N., and Verstrepen, K. J. (2009). Promoter architecture and the evolvability of gene expression. *J. Biol.* 8, 95. doi:10.1186/jbiol204.

Tomé, S., Holt, I., Edelmann, W., Morris, G. E., Munnich, A., Pearson, C. E., et al. (2009). MSH2 ATPase Domain Mutation Affects CTG N CAG Repeat Instability in Transgenic Mice. 5. doi:10.1371/journal.pgen.1000482.

Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S., and Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38. doi:10.1093/nar/gkq543.

Trowsdale, J., and Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annu. Rev. Genomics Hum. Genet.* 14, 301–323. doi:10.1146/annurev-genom-091212-153455.

Tsai, Y., Greenberg, D., Powell, J., Höijer, I., Ameur, A., Strahl, M., et al. (2017). Amplification-free , CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions. *bioRxiv*, 1–26. doi:10.1101/203919.

Tsujimoto, Y., Takakuwa, T., Takayama, H., Nishimura, K., Okuyama, A., Aozasa, K., et al. (2004). In Situ Shortening of CAG Repeat Length within the Androgen Receptor Gene in Prostatic Cancer and Its Possible Precursors. *Prostate* 58, 283–290. doi:10.1002/pros.10333.

Turner, T. R., Hayhurst, J. D., Hayward, D. R., Bultitude, W. P., Barker, D. J., Robinson, J., et al. (2018). Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 International HLA and Immunogenetics Workshop cell lines. *Hla* 91, 88–101. doi:10.1111/tan.13184.

Usdin, K. (2008). The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Res.* 18, 1011–1019. doi:10.1101/gr.070409.107.

Usdin, K., Hayward, B. E., Kumari, D., Lokanga, R. A., Sciascia, N., and Zhao, X. N. (2014). Repeat-mediated genetic and epigenetic changes at the FMR1 locus in the Fragile X-related disorders. *Front. Genet.* 5, 226. doi:10.3389/fgene.2014.00226.

Usdin, K., House, N. C. M., and Freudenreich, C. H. (2015). Repeat instability during DNA repair: Insights from model systems. *Crit. Rev. Biochem. Mol. Biol.* 50, 142–167. doi:10.3109/10409238.2014.999192.

Van Blitterswijk, M., Dejesus-Hernandez, M., and Rademakers, R. (2012). How do C9ORF72 repeat expansions cause amyotrophic lateral sclerosis and frontotemporal dementia: Can we learn from other noncoding repeat expansion disorders? *Curr. Opin. Neurol.* 25, 689–700. doi:10.1097/WCO.0b013e32835a3efb.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet.* 34, 666–681. doi:10.1016/j.tig.2018.05.008.

Van Esch, H. (2006). The Fragile X premutation: New insights and:clinical consequences. *Eur. J. Med. Genet.* 49, 1–8. doi:10.1016/j.ejmg.2005.11.001.

Verkerk, A. J., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P., Pizzuti, A., et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* 65, 905–914. doi:10.1016/0092-8674(91)90397-H.

Verma, M., Rogers, S., Divi, R. L., Schully, S. D., Nelson, S., Joseph Su, L., et al. (2014). Epigenetic research in cancer epidemiology: trends, opportunities, and challenges. *Cancer Epidemiol. Biomarkers Prev.* 23, 223–33. doi:10.1158/1055-9965.EPI-13-0573.

Vermeesch, J. R., Voet, T., and Devriendt, K. (2016). Prenatal and pre-implantation genetic diagnosis. *Nat. Rev. Genet.* 17, 643–656. doi:10.1038/nrg.2016.97.

Verstrepen, K. J., Jansen, A., Lewitter, F., and Fink, G. R. (2005). Intragenic tandem repeats generate functional variability. *Nat. Genet.* 37, 986–90. doi:10.1038/ng1618.

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K. J. (2009). Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability. *Science (80-.).* 324, 1213–1216. doi:10.1126/science.1170995.

Voineagu, I., Surka, C. F., Shishkin, A. A., Krasilnikova, M. M., and Mirkin, S. M. (2009). Replisome stalling and stabilization at CGG repeats, which are responsible for chromosomal fragility. *Nat. Struct. Mol. Biol.* 16, 226–228. doi:10.1038/nsmb.1527.

Wang, J. W., Wang, A., Li, K., Wang, B., Jin, S., Reiser, M., et al. (2015). CRISPR/Cas9 nuclease cleavage combined with Gibson assembly for seamless cloning. *Biotechniques* 58, 161–170. doi:10.2144/000114261.

Wang, Y.-H., and Griffith, J. (1995). Expanded CTG Triplet Blocks from the Myotonic Dystrophy Gene Create the Strongest Known Natural Nucleosome Positioning Elements. *Genomics* 25, 570–573. Available at: https://ac-els-cdn-com.kuleuven.ezproxy.kuleuven.be/088875439580061P/1-s2.0-088875439580061P-main.pdf?_tid=9b94d120-9306-485a-ad2d-b28eda8508be&acdnat=1544882237_a645030e332c4803eed7f7c71b516165 [Accessed December 15, 2018].

Weiser, J. N., Love, J. M., and Moxon, E. R. (1989). The molecular mechanism of phase variation of H. influenzae lipopolysaccharide. *Cell* 59, 657–665. doi:10.1016/0092-8674(89)90011-1.

Wohrle, D., Hennig, I., Vogel, W., and Steinbach, P. (1993). Mitotic stability of fragile X mutations in differentiated cells indicates early post-conceptual trinucleotide repeat expansion. *Nat. Genet.* 3, 73–96.

Wöhrle, D., Hirst, M. C., Kennerknecht, I., Davies, K. E., and Steinbach, P. (1992). Genotype mosaicism in fragile X fetal tissues. *Hum. Genet.* 89, 114–116. doi:10.1007/BF00207057.

Yanovsky-Dagan, S., Avitzour, M., Altarescu, G., Renbaum, P., Eldar-Geva, T., Schonberger, O., et al. (2015). Uncovering the Role of Hypermethylation by CTG Expansion in Myotonic Dystrophy Type 1 Using Mutant Human Embryonic Stem Cells. *Stem Cell Reports* 5, 221–231. doi:10.1016/j.stemcr.2015.06.003.

Yao, K., Muruvanda, T., Roberts, R. J., Payne, J., Allard, M. W., and Hoffmann, M. (2016). Complete Genome and Methylome Sequences of Salmonella enterica subsp. enterica Serovar Panama (ATCC 7378) and Salmonella enterica subsp. enterica Serovar Sloterdijk (ATCC 15791): TABLE 1. *Genome Announc.* 4, e00133-16. doi:10.1128/genomeA.00133-16.

Yrigollen, C. M., Durbin-Johnson, B., Gane, L., Nelson, D. L., Hagerman, R., Hagerman, P. J., et al. (2012). AGG interruptions within the maternal FMR1 gene reduce the risk of offspring with fragile X syndrome. *Genet. Med.* 29, 729–736. doi:10.1038/gim.2012.34.

Yrigollen, C. M., Martorell, L., Durbin-Johnson, B., Naudo, M., Genoves, J., Murgia, A., et al. (2014a). AGG interruptions and maternal age affect FMR1 CGG repeat allele stability during transmission. *J. Neurodev. Disord.* 6, 24. doi:10.1186/1866-1955-6-24.

Yrigollen, C. M., Sweha, S., Durbin-Johnson, B., Zhou, L., Berry-Kravis, E., Fernandez-Carvajal, I., et al. (2014b). Distribution of AGG interruption patterns within nine world populations. *Intractable Rare Dis. Res.* 3, 153–161. doi:10.5582/irdr.2014.01028.

Yu, Z., Zhu, Y., Chen-Plotkin, A. S., Clay-Falcone, D., McCluskey, L., Elman, L., et al. (2011). PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. *PLoS One* 6, 14–19. doi:10.1371/journal.pone.0017951.

Zamani Esteki, M., Dimitriadou, E., Mateiu, L., Melotte, C., Vander Aa, N., Kumar, P., et al. (2015). Concurrent Whole-Genome Haplotyping and Copy-Number Profiling of Single Cells. *Am. J. Hum. Genet.* 96, 894–912. doi:10.1016/j.ajhg.2015.04.011.

Zeidler, S., de Boer, H., Hukema, R. K., and Willemsen, R. (2017). Combination Therapy in Fragile X Syndrome; Possibilities and Pitfalls Illustrated by Targeting the mGluR5 and GABA Pathway Simultaneously. *Front. Mol. Neurosci.* 10, 368. doi:10.3389/fnmol.2017.00368.

Zeng, X., King, J. L., Stoljarova, M., Warshauer, D. H., Larue, B. L., Sajantila, A., et al. (2015). High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. *Forensic Sci. Int. Genet.* 16, 38–47. doi:10.1016/j.fsigen.2014.11.022.

Zhang, X., Zhuang, X., Gan, S., Wu, Z., Chen, W., Hu, Y., et al. (2012). Screening for FMR1 expanded alleles in patients with parkinsonism in mainland China. *Neurosci. Lett.* 514, 16–21. doi:10.1016/j.neulet.2012.02.036.

Zuñiga, A., Juan, J., Mila, M., and Guerrero, A. (2005). Expansion of an intermediate allele of the FMR1 gene in only two generations. *Clin. Genet.* 68, 471–473. doi:10.1111/j.1399-0004.2005.00514.x.

List of Abbreviations

6mA	N6-methyladenosine
AFF2	AF4/FMR2 family member 2
ALS	amyotrophic lateral sclerosis
AR	androgen receptor
ATN1	atrophin 1
ATXN	ataxin
ATXN8OS	ataxin 8 opposite strand
BER	base excision repair
CACNA1A	calcium voltage-gated channel subunit alpha1 A
CCS	circular consensus sequence
CLR	continuous long reads
CNBP	CCHC-type zinc finger nucleic acid binding protein
CRISPR/CAS	clustered regularly interspaced short palindromic repeats/CRISPR-associated systems
dCAS9	dead CAS9
DM	myotonic dystrophy
DMPK	dystrophic myotonic protein kinase
DRPLA	dentatorubral-pallidoluysian atrophy
EPM1	progressive myoclonic epilepsy 1
ESC	embryonic stem cell
FMR1	fragile-X mental retardation 1
FMRP	fragile X mental retardation protein
FRAX-E	fragile XE syndrome
FRDA	friedreich’s ataxia
FTD	frontotemporal dementia
FXN	frataxin
FXPOI	fragile X-associated primary ovarian insufficiency
FXS	fragile X syndrome
FXTAS	fragile X-associated tremor/ataxia syndrome
HD	huntington's disease
HDL2	Huntington disease-like 2
hPSC	human pluripotent stem cell
HTT	huntingtin
IPD	interpulse duration
iPSCs	induced pluripotent stem cells
JPH3	junctophilin 3
m4C	N4-methylcytosine
mGluR	metabotropic glutamate receptor
MMR	mismatch repair system
MPS	massively parallel sequencing
MSH	mutS homologue
NGS	next-generation sequencing

OGG1	7,8-dihydro-8-oxoguanine DNA glycosylase
ONT	oxford nanopore technologies
OPMD	oculopharyngeal muscular dystrophy
PacBio	pacific biosciences
PAPBN1	poly(A) binding protein nuclear 1
PolyA	polyalanine
PolyQ	polyglutamine
PPP2R2B	protein phosphatase 2 regulatory subunit B beta
RAN	repeat associated non-ATG translation
ROI	reads-of- insert
SCA	spincerebellar ataxia
sgRNA	short-guide RNA
SMBA	Spinal and bulbar muscular atrophy
SMRT	single molecule real-time
SNP	single nucleotide polymorphism
SP-PCR	small-pool PCR
STR	short tandem repeat
TBP	TATA-box binding protein
TCR	transcription-coupled repair
TP-PCR	triplet-primed PCR
TR	tandem repeat
UTR	untranslated region
ZMW	zero-mode waveguide

Scientific Acknowledgement

First, we are grateful to the patients and their families for their participation in the projects of this study. We also acknowledge the clinicians of the Center for Human genetics (UZ Leuven, Leuven), especially prof. dr. Thomy de Ravel de l'Argentière, prof. dr. Koenraad Devriendt, prof. dr. Eric Legius and prof. dr. Hilde Van Esch for the consultation, recruitment and follow-up of the patients included in this study. We are grateful for the collaboration with the Eugin Clinic (Barcelona Spain) on the *FMR1* intermediate allele instability project and to the research groups of Karen Sermon (VUB, Brussel) and Thierry Vandendriessche (VUB Brussel) on the sequencing of the *DMPK* CTG repeat. We also would like to thank Greet Peeters, Wim Meert and Steve Smekens for their kindness and help with the laboratory work. Finally, we would also like to thank all contributing authors and prof. dr. Matthew Hestand for reading and correcting the manuscripts.

Funding was provided by the European Union's Research and Innovation funding program Horizon 2020 WIDENLIFE: 692065; the University of Leuven (KU Leuven) GOA (GOA/12/015 to J.R.V., K.D. and H.V.E.), SymBioSys (PFV/10/016), the Hercules foundation (ZW11-14) and the Belgian Science Policy Office Interuniversity Attraction Poles (BELSPO-IAP) program through the project IAP P7/43-BeMGI. SA received a Ph.D. grant (SB/131787) for strategic basic research of the Agency for Innovation by Science and Technology (IWT).

Personal Contribution

Chapter 3

In chapter 3 (Development of an amplification-free enrichment method targeting the FMR1 CGG repeat), all experiments, strategies and sgRNA's were designed by Simon Ardui (SA). The candidate also performed the experiments described in this chapter (including DNA extraction, PCR, CRISPR/CAS9 digestion, PacBio library preparation and restriction digestion). Sequencing of the generated libraries was performed at the Genomics Core (KU Leuven/UZ Leuven). After sequencing, SA generated the correct data formats and performed the bio-informatic analysis (including repeat and kinetic analysis).

Chapter 4

In Chapter 4 (Detecting AGG Interruptions in Male and Female *FMR1* Premutation Carriers by Long-Read Sequencing) all experiments were designed by SA. The amplicon generation, pooling and library preparation were exclusively done by SA whereafter the generated libraries were sequenced at the Genomics Core (KU Leuven/UZ Leuven). Subsequently, the candidate created the bioinformatic pipelines and performed analysis of the sequencing data described in chapter 4.

Chapter 5

Chapter 5 (Leveraging the power of SMRT sequencing to improve *DMPK* CTG repeat characterization) reports on the use of SMRT sequencing to study the *DMPK* CTG repeat. Here, the candidate defined and developed appropriate sequencing strategies to answer the research questions put forward by the collaborators in this chapter (Prof. Karen Sermon and prof. Thierry VandenDriessche from the Vrije Universiteit Brussel (VUB).

The experiments performed to determine the CRISPR/CAS9 gene editing efficiency (PCR and library preparation) were done by SA. To determine the variability of the DMPK CTG repeat, amplification was done by Silvie Franck (SF) from the research group of Karen Sermon. Library preparation was done by both SA and SF. All libraries were sequenced at the Genomics Core (KU Leuven/UZ Leuven). Finally, the creation of the bioinformatic pipelines and the analysis of the data was done by SA.

Conflict of Interest

The authors declare that there is no conflict of interest.

Curriculum Vitae

Simon Ardui

°10-05-1989, Lier (Belgium)

Address:
Rostal 4B, 2280 Grobbendonk
simonardui@gmail.com
0494325731

Experience

2014 - 2018	**PhD researcher** at the Laboratory for Cytogenetics and Genome Research, Centre of Human Genetics, KU Leuven.
2012 - 2013	**Research assistant** at the Laboratory for Cytogenetics and Genome Research, Centre of Human Genetics, KU Leuven.

Education

2016 - 2017	Academic Teacher Training (CVO De Oranjerie)
2015 - 2016	Postgraduate course in Human genetics (interuniversity course)
2010 - 2012	Master of Science in Bioscience Engineering (*Cum Laude*, KU Leuven) *Specialization: biomolecular engineering*
2007-2010	Bachelor of Science in Bioscience Engineering (*Cum Laude*, KU Leuven) *Erasmus at the university of Zaragoza (Spain)*

Additional Training

2015-2016	Introduction to Python Programming (VIB-BITS)
2015-2016	Introduction to BioPython (VIB-BITS)
2014-2015	Scientific Tools – Statistics (KU Leuven)
2014-2015	New Tools and Developments for Gene Editing and DNA Synthesis (VIB)
2014-2015	Presentation and Seminar Skills for Biomedical Researchers (KU Leuven)
2014-2015	Essential Tools for R - Introduction to R (VIB-BITS)
2013-2014	Practical Computing for Bioinformatics (KU Leuven)

Languages

Dutch	mother tongue
English & French	fluent
Spanish & German	basic

Publications

- **Ardui, S.**, Race V, De Ravel, T., Van Esch H, Devriendt K, Matthijs G, Vermeesch J (2018). Detecting AGG Interruptions in Females With a FMR1 Premutation by Long-Read Single-Molecule Sequencing: A 1 Year Clinical Experience. Frontiers in Genetics 9, 150.
- **Ardui, S**., Ameur, A., Vermeesch, J., Hestand, M. (2018). Single Molecule Real-Time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 46, 2159–2168.
- Dastidar, S., **Ardui, S**., Singh, K., Majumdar, D., Nair, N., Fu, Y., Reyon, D., Samara, E., Gerli, M.F.M., Klein, A.F., et al. (2018). Efficient CRISPR / Cas9-mediated editing of trinucleotide repeat expansion in myotonic dystrophy patient-derived iPS and myogenic cells. Nucleic Acids Res. 1, 1–24.
- **Ardui, S**., Race V, Zablotskaya A, Hestand M, Van Esch H, Devriendt K, Matthijs G, Vermeesch J. (2017). Detecting AGG Interruptions in Male and Female FMR1 Premutation Carriers by Single-Molecule Sequencing. Hum Mutat 38:324–331
- Bayandir, B., Dehaspe, L, Brison, N., Brady, P., **Ardui, S**., Kamoun, M., Van der Veken, L., et al. (2015). Non-Invasive Prenatal Testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. European Journal of Human Genetics 23:1286-1293.
- Brady, P., **Ardui, S**., Vermeesch, J.R. (2013). The Future of Prenatal Cytogenetics: From Copy Number Variations to Non invasive Prenatal Testing. Current Genetic Medicine Reports 1: 91-98.

Meetings

- 3rd International conference of FMR1 premutation (2017, Jeruzalem). Oral presentation: Detecting AGG interruptions in male and female FMR1 premutation carriers by single-molecule sequencing
- European Society Meeting for Human Genetics (2016, Barcelona). Oral presentation: Detection of AGG interruptions in FMR1 premutation females by single-molecule sequencing.
- Genetics & Genomics symposium (2016, Leuven). Oral presentation: Single Molecule Real-Time Sequencing of the CGG repeat in the FMR1 gene. Awarded as best oral presentation.
- Joint Meeting BESHG/NVHG (2016, Leuven). Poster Presentation: Detection of AGG interruptions in FMR1 premutation females by Single-molecule sequencing.
- Genetics Retreat Meeting (2015, Kerkrade (NL)). Oral presentation. Long Read Sequencing of the FMR1 CGG Repeat: a Break Free of the Limited GC-Content Sequencability of Massively Parallel Sequencing Technologies.
- Belgian Society of Human Genetics meeting (Brussels, 2013). Oral presentation: Validation of a 60 k Cytosure ISCA + SNP array for prenatal diagnosis. Awarded as best young investigator.

Personal acknowledgements

I would like to thank my promotor Prof. Joris Vermeesch to have given me the opportunity to do this PhD. Joris, you often refueled my energy with your continuous enthusiasm and by emphasizing the value of the project and the generated results. I also appreciated the help of my co-promotor Gert Matthijs. Thanks for the input in the project, proofreading papers and giving me new insights. I also acknowledge the clinicians of the Center for Human genetics (UZ Leuven, Leuven), especially prof. dr. Thomy de Ravel de l'Argentière, prof. dr. Koenraad Devriendt, prof. dr. Eric Legius and prof. dr. Hilde Van Esch for their enthusiasm for my project and help with cell lines, patients or correcting manuscripts.

I also want to express my gratitude towards the members of the jury, Prof. dr. Adam Ameur, Prof. dr. Diether Lambrechts, Prof. dr. Karen Sermon and Prof. dr. Kevin J. Verstrepen for taking your time to carefully evaluate this PhD project. Your input definitely improved this manuscript. Also, many thanks to professor Anton Roebroek for accepting the role of chairman.

Verschillende mensen op de gang van het zesde verdiep waren erg belangrijk tijdens mijn doctoraat. Ik denk hierbij in de eerste plaats aan de mensen van de Genomics Core waarbij ik te pas en te onpas over de vloer kwam voor de reservatie van een PCR machine of een probleempje met de fragment analyzer. In het bijzonder wil ik graag Wim Meert en Stephanie Deman bedanken voor hun toegewijdheid en motivatie om steeds het beste uit mijn Pacbio libraries te halen. Bedankt! Eén deurtje verder moet ik ook Erik Reggers bedanken voor de eindeloze hoeveelheid toegangsbadgen die ik bij hem mocht lenen. Tot vandaag heb ik nog een uitleenbadge in mijn portefeuille zitten. Daarnaast wil ik ook Veerle Mattheus bedanken om alle (administratieve) hulp met de glimlach te verlenen. Nog enkele deuren verder is de bureau van Valerie Race en Steve Smekens. Bedankt voor de input, het bespreken van experimenten en ideeën, de vele 'gratis' reagentia en het verzamelen van DNA stalen.

Mijn habitat was natuurlijk de bureau op het vijfde verdiep. Met mensen die constant in de bureau binnen en buiten fladderen deed hij me vaak denken aan een duiventil. Sinds ik destijds in de bureau aankwam is er nu nog enkel Greet, wat direct alles zegt over haar waarde voor de bureau en de groep van Joris. Danku Greet voor alle hulp de voorbije jaren, maar vooral voor de leuke babbels! Door af en toe fruit of Bicky Burgers te voorzien zorgde je voor een stukje huiselijkheid en gezelligheid op het werk, wat ik erg leuk vond! De toekomst van de groep is verzekerd met onder andere Heleen, Margot en Lisanne. Bedankt voor de leuke babbels en heel veel succes met jullie toekomst (doctoraten, kinderen, volleybal,…). It was really fun and interesting working in such an international group. Aimé, Alena, Molka, Francesca, Aspasia, Olga, Eftychia, Romain, Berardo, Darine and Matthew: thanks for the interesting conversations, movie- and superbowl evenings. Also, I really enjoyed all the sweets you brought from your home countries! Daarnaast apprecieerde ik ook erg de gesprekjes aan de printer, een momentje aan het koffiemachine of een goede grap vanachter de computer met Nele, Laura, Eline, Yoeri, Carlijn, Lise, Elfi en Lieselot. Ik had het geluk om mijn doctoraat samen te mogen starten met twee erg leuke kerels: Koen en Wolf. Jullie zorgden voor de nodige sfeer de voorbije 5 jaar! Koen, met jou had ik erg leuke babbels, vooral over muziek en festivals. Ik hoop dat we ooit toch samen eens een optreden kunnen doen. Wolf, de kassei die we destijds gingen eten na een les humane genetica zal me altijd bijblijven. Voor mij het hoogtepunt van mijn doctoraat 😉. Ik wens jullie veel succes met jullie doctoraten, huizen, jobs, pedalen,…

Een ander voordeel van het uitoefenen van een doctoraat is dat je de mogelijkheid hebt om in het Begijnhof te wonen. Een prachtige omgeving waar Tatyana en ik een fantastische tijd hadden. De vriendschap die we in die tijd met onze buren Andreas en Sara konden opbouwen zal ik blijven koesteren. Tot slot vonden we in Leuven een erg leuke en warme vriendengroep die ook wel eens graag een lekkere pintje lust. We missen jullie, maar met het kerstfeestje als ankerpunt blijven onze vriendschapsbanden zeker intact.

Natuurlijk wil ik ook heel graag mijn dichte familie bedanken voor alle steun. Anton, Marie & Librecht, mama & papa: misschien hielpen jullie niet direct aan het doctoraat, maar jullie stonden wel steeds klaar om me te helpen met alles daarbuiten! Jullie vroegen je de voorbije jaren vaak af waar ik nu eigenlijk mee bezig was, maar moesten het steeds stellen met enkele vage antwoorden. Ik hoop dat dit vandaag wat duidelijker geworden is 😊. Lang leve onze gezellige eetentjes op zondagavond!

Graag wil ik eindigen met mijn Chouke in de bloemetjes te zetten. Jij hebt de voorbije jaren zo hard meegeleefd met mij en me altijd zo gesteund! Zelfs als ik slechtgezind thuiskwam na een mislukt experiment raakte je tot mijn grote verbazing niet beu. In tegendeel, je slaagde er steeds in om me terug op te beuren en aan te moedigen om verder te gaan. Ik heb erg veel zin om samen met jou aan nieuwe avonturen te beginnen ♥