Detecting AGG interruptions in male and female *FMR1* premutation carriers by single-molecule sequencing

Simon Ardui, Valerie Race, Alena Zablotskaya, Matthew S. Hestand, Hilde Van Esch, Koenraad Devriendt, Gert Matthijs, Joris R. Vermeesch*

Department of Human Genetics, KU Leuven, Leuven, Belgium

*Corresponding author. Correspondence to:

Joris R. Vermeesch,

Department of Human Genetics, KU Leuven

O & N I Herestraat 49, box 602

3000 Leuven, Belgium

Email: joris.vermeesch@uzleuven.be

Abstract

The *FMR1* gene contains an unstable CGG repeat in its 5' untranslated region. Premutation alleles range between 55 and 200 repeat units and confer a risk for developing fragile X-associated tremor/ataxia syndrome or fragile X-associated primary ovarian insufficiency. Furthermore, the premutation allele often expands to a full mutation during female germline transmission giving rise to the fragile X syndrome. The risk for a premutation to expand depends mainly on the number of CGG units and the presence of AGG interruptions in the CGG repeat. Unfortunately, the detection of AGG interruptions is hampered by technical difficulties. Here, we demonstrate that single-molecule sequencing enables the determination of not only the repeat size, but also the complete repeat sequence including AGG interruptions in male and female alleles with repeats ranging from 45 to 100 CGG units. We envision this method will facilitate research and diagnostic analysis of the *FMR1* repeat expansion.

Keywords

Introduction

The fragile X mental retardation gene (*FMR1*; MIM *309550) is located on band q27.3 of the X chromosome and contains a CGG tandem repeat in its 5' untranslated region (UTR)(Harrison et al. 1983; Verkerk et al. 1991). The size of the CGG repeat is variable and this characteristic is used to classify the repeats into 4 categories: normal (<45 units), gray-zone (45-54 units), premutation (55-200 units) and full mutations (>200 units)(Biancalana et al. 2015). The repeat size influences the instability of the repeat and also the phenotype of the individual will be different. Carriers of a gray-zone allele have a normal phenotype. The small repeats of this category might be unstable upon transmission, but they rarely expand into a premutation allele (Biancalana et al. 2015). In contrast, individuals carrying a premutation allele are at risk for developing the late-onset neurodegenerative disorder fragile X-associated tremor/ataxia syndrome (FXTAS; MIM #300623) or fragile X-associated primary ovarian insufficiency (FXPOI; MIM #311360) (Sullivan et al. 2005; Hagerman and Hagerman 2013). Moreover, female premutation carriers are at risk of transmitting a full mutation (>200 CGG units) to their offspring (Oberlé et al. 1991; Yu et al. 1991; Penagarikano et al. 2007). This large expansion induces abnormal methylation of the promoter of the *FMR1* gene which disrupts the production of the Fragile X Mental Retardation Protein (FMRP)(Pieretti et al. 1991). The absence of FMRP causes the fragile X syndrome (FXS; MIM #300624) which is the most frequent monogenic cause of X-linked Intellectual Disability (XLID) and autism. Other phenotypes often observed in patients with FXS are hypersensitivity, hyperactivity, attention deficit and mild dysmorphic features (Penagarikano et al. 2007; Usdin et al. 2014).

Due to the severity of the FXS, an accurate estimate of the risk for a woman with a premutation allele to transmit a full mutation to her offspring is crucial in genetic counseling. For example, a woman with a high risk can decide to choose for preimplantation genetic

diagnosis (PGD) where one could select for unaffected males or non-carrier female embryos (Sermon et al. 1999; Burlet et al. 2006). On the other hand, women with only a minor risk would choose for normal conception, optionally in combination with prenatal diagnosis to monitor the fragile X status of their fetus (Biancalana et al. 2015). With an incidence of about 1 in 200 females carrying a premutation allele, this is a pertinent question in genetic counseling (Tassone et al. 2012).

The risk of a premutation to expand to a full mutation depends mainly on the size of the premutation allele, whereby larger alleles expand faster into full mutations. For example, alleles larger than 100 CGG repeat units have a risk of almost 100% to expand into a full mutation. In contrast, small premutation alleles (55 to 59 CGG units) have only 3% expansion risk to a full mutation and alleles ranging between 60 and 100 CGG tandem repeats have intermediate risks (Yrigollen et al. 2012). The *FMR1* CGG repeat is often interrupted by 1-4 AGGs clustered towards the 5' end. It has been shown that AGG interruptions will increase the stability and reduce the risk for expansions. (Eichler et al. 1994; Nolin et al. 2014). Large population studies indicate that the risk reduction of AGGs is the biggest for repeat sizes between 55 and 85 CGG units (Nolin et al. 2014; Yrigollen et al. 2014). For example, for a woman with 75 repeats interspersed with 2 AGG units the risk to transmit a full mutation to her offspring is 12% but increases to 77% if AGG interruptions are present(Yrigollen et al. 2012). For alleles smaller than 55 there is no risk difference with or without AGG interruptions. The differential risk reduces to less than 38% for alleles larger than 85 CGG units. Since for those large alleles the overall risk for an expansion is larger than 50% irrespective of the number of AGG interruptions, the clinical importance of mapping the interruptions in larger expansions drops. Hence, mapping the location and number of AGGs within the CGG repeat is essential for an accurate risk prediction and genetic counseling, especially for alleles with 55-85 CGG units. However, AGG

measurement is not yet a standard feature of *FRM1* diagnostic work-up in most laboratories worldwide (Jacquemont et al. 2011; Monaghan et al. 2013; Biancalana et al. 2015).

Measurement of AGG interruptions has been hampered by technical difficulties. Traditional Southern blotting cannot localize the AGG interruptions. If determined, AGG interruptions are detected by a Triplet-Primed PCR (TP-PCR)(Chen et al. 2010). This is an indirect method whereby the forward and reverse primer of a standard PCR are complemented with a third primer which will anneal right into the repeat. By adding the third primer, the PCR will produce a ladder of peaks which will be visible on an agarose gel 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 two X-chromosomes each containing a different CGG repeat with a specific set of AGG units. If the structure of those two 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 (Texas, USA) 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 sequencing technology is able to sequence through large and very GC-rich repeats, including CGG repeats (Loomis et al.

2013; Shin et al. 2013; Chaisson et al. 2015). Furthermore, Loomis et al. (2013) proved single molecule sequencing enabled the detection of AGG interruptions as they showed 1 AGG unit embedded in a CGG repeat cloned in a plasmid. Finally, the long reads 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 which are randomly distributed across the reads and generate very accurate reads-of-insert (Carneiro et al. 2012; Jiao et al. 2013; Hestand et al. 2016).

We demonstrate that single-molecule sequencing enables reading through repeat- and GC-rich regions with a very high accuracy which permits reconstruction of the complete repeat structure for gray zone and premutation alleles, not only for males, but also for females.

Materials and Methods

DNA samples. 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 34 females (5 females with a normal and a gray zone allele, 28 females with a normal and a premutation allele and 1 female with 2 premutation alleles). 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. After informed consent was obtained from the patients, DNA was isolated from peripheral white blood cells according to standard procedures. This study was approved by the local ethics committee.

Amplicon generation. First, the *FMR1* CGG repeat was amplified using the PCR$_X$ Enhancer System (Invitrogen, Carlsbad, CA) with 40ng/ul DNA input and previously published specific primers (Figure 1A)(Filipovic-Sadic et al. 2010). In order to integrate barcodes, an M13 tail was attached at both the forward (M13-Forward tail:

TGTAAAACGACGGCCAGT) and the reverse primer (M13-Reverse tail: CAGGAAACAGCTATGACC). Next, a reaction mixture was prepared by combining 4 µl of 10X PCR$_X$ Amplification Buffer, 1,2 µl 50 mM MgSO$_4$, 4 µl of 2 mM dNTPs (Invitrogen), 16 µl of 10X PCR$_X$ Enhancer Solution, 4 µl of a 2,5 uM 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 minutes, followed by 25 cycles of 95°C for 30 seconds, 64°C for 60 seconds and 68°C for 90 seconds, followed by a final extension step at 72°C for 5 minutes where after 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 ul of water. Next, barcoded primers from PacBio's 96-well barcoding kit were attached to the amplicons by their M13 tail, which allowed pooling 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 seconds followed by 5 cycles of 95°C for 30 seconds, 45°C for 60 seconds and 68°C for 90 seconds, followed by another 5 cycles of 95°C for 30 seconds, 65°C for 60 and 68°C for 90 seconds and a final elongation step at 72°C for 5 minutes. Afterwards, the amplicons were again purified by 1X washed Agencourt AMPure XP beads, visualized on the Fragment Analyser and pooled equimolar together. By using this strategy, 3 pools with a maximum of 24 samples were generated.

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 minutes movie (Figure 1B). All runs used PacBio's DNA Sequencing Reagent Kit 4.0 v2. The 3 pools were sequenced in three different sequencing runs. The first pool was sequenced on 2 SMRT cells, whereas pool 2 and 3 only used a single SMRT cell.

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 1C). 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 1D).

- 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 convert to BED format and select the on-target reads-of-inserts BEDtools v2.20.1 was used (Quinlan and Hall 2010).

- Seperation of the two 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 (Figure 1C), followed by separation of the normal from the premutation allele based on differences of the read size (Figure 1E). 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 1F) 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. 2014). To perform an assembly on the normal reads, MIRA was run with custom tuned parameters which can be found in the Supplemental 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 and reported according to nomenclature described by the Human Genome Variation Society (HGVS). To control the assembly process, the repeat structure extracted from the *de novo* assembly was compared to the repeat characteristics (repeat size, AGG interruptions) of the individual reads-of-insert. All variants are submitted to the *FMR1* locus-specific database which can be found at http://www.lovd.nl/FMR1 (Fokkema et al. 2011).

Determination of the precision and robustness of AGG interruption detection. In order to describe the precision of the described AGG interruption detection method we use 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 one female.

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, Illininois, USA) carried out following the manufacturer's instructions. AGG information was extracted from the TP-PCR result if the electropherogram generated an interpretable result.

Results

FMR1 **CGG repeat structure determination**

To determine the *FMR1* repeat structure we performed single-molecule sequencing of PCR amplicons from 34 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 (Supp. Figure S1). The CGG repeat of the 7 males was supported by a mean coverage of 261 [84-614] reads (Supp. Figure S2A). In females, the premutation allele contains more CGG units than the normal allele and therefore amplifies worse during PCR. Consequently, the premutation will be covered by less reads. The female samples (1-34) had a mean coverage of 277 [83-458] and 158 [22-332] for the normal and premutation alleles respectively (Supp. Figure S2B-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 2A). Subsequently, the repeat structure was investigated (Table 1). For all male samples the number and position of AGG units observed by single molecule sequencing was 100% concordant with TP-PCR (Supp. Figure S3A).

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 2). All assembled normal and premutation repeat sizes fell within the error range of PCR (Figure 2B-C). The repeat structure of single-molecule sequencing was validated by TP-PCR for 6 females (Supp. Figure S3B).

For 4 females (female 31-34) 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 siz distribution of the reads-of-insert, it became difficult to separate the reads derived from the different alleles based on this characteristic (Figure 3A). Therefore the *de novo* assembly was performed on the mixture of reads. To make sure both alleles were generated by the assembler, also the distribution of the number of AGG interruptions as function of the repeat sizes of the individual reads-of-insert were mapped (Figure 3B). 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 two female alleles. In figure 3B also smaller circles are present flanking the most frequently occurring repeat sequences which represent stutter products inherent to PCR amplification of repeat rich regions. An overview of the females with small repeat size differences between their alleles can be found in Table 3. We also tested PacBio's Long Amplicon Analysis tool, but this performed overall worse than the MIRA assembly pipeline (see supplementary data).

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 4). Except for female 17, also the size of the repeats was 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 two samples (7 & 19 reads respectively, Supp. Figure S2C) 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 4; female 12). Thus, single-molecule sequencing generates results with a high precision and robustness.

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. 2014; Yrigollen et al. 2014). This is also increasingly recognized in international guidelines on *FMR1* genetic testing (Monaghan et al. 2013; Biancalana et al. 2015). 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

which will become even more strong thanks to PacBio's new Sequel system which is more high-throughput and 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). As this is an amplicon based method, also stutter products will be present next to products with the true repeat size. Luckily those stutter products are mainly ± 1 repeat unit and therefore their influence on the expansion risk is negligible. 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 two alleles of females starting from genomic DNA.

Knowledge of the risk for FMR1 CGG expansion to occur has a profound impact on reproductive choices. Couples at risk for offspring with FXS can choose for prenatal diagnosis with possible termination of an affected pregnancy (Burlet et al. 2006). 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. Therefore easy access to accurate AGG information will be extremely valuable in guiding and reassuring couples to make the right decisions.

In rare cases where women carry two expanded alleles, selection can target the allele with the highest risk. We already reported the CGG sequencing result obtained in this study to a female (female 34) 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. Since full mutations can also be amplified and single-molecule sequenced (Loomis et al. 2013), AGG interruptions can possibly also be detected in this repeat category. Furthermore, recently a method using a PCR-free approach to do targeted enrichment was published (Pham et al. 2016), and further improvements can be expected in the near future. Those strategies will enable to detect the original repeat size distribution and will in addition simultaneously reveal the epigenetic signature of the repeat.

In conclusion 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 two alleles of females. Single-molecule sequencing enables for the first time to separate unambiguously the two female repeats which enabled the generation of the exact repeat structure for both the normal and premutation allele. 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 (Kroutil et al. 1996; Matsuura et al. 2006; Gao et al. 2008; Musova et al. 2009; Braida et al. 2010;

Holloway et al. 2011; Yu et al. 2011; Menon et al. 2013). We foresee that this technology will replace current tests and has the potential to improve risk estimates allowing for improved genetic counseling.

Acknowledgements

References

Biancalana V, Glaeser D, McQuaid S, 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.

Braida C, Stefanatos RKA, Adam B, Mahajan N, Smeets HJM, Niel F, Goizet C, Arveiler B, Koenig M, Lagier-Tourenne C, Mandel JL, Faber CG, 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.

Burlet P, Frydman N, Gigarel N, Kerbrat V, Tachdjian G, Feyereisen E, Bonnefont JP, Frydman R, Munnich A, Steffann J. 2006. Multiple displacement amplification improves PGD for fragile X syndrome. Mol Hum Reprod 12:647–652.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, Depristo MA. 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13:375.

Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. Nature 517:608–611.

Chen L, Hadd A, Sah S, Filipovic-Sadic S, Krosting J, Sekinger E, Pan R, Hagerman PJ, Stenzel TT, Tassone F, Latham GJ. 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.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159.

Eichler EE, Holden JJ, Popovich BW, Reiss AL, Snow K, Thibodeau SN, Richards CS, Ward PA, Nelson DL. 1994. Length of uninterrupted CGG repeats determines instability in the FMR1 gene. Nat Genet 8:88–94.

Filipovic-Sadic S, Sah S, Chen L, Krosting J, Sekinger E, Zhang W, Hagerman PJ, Stenzel TT, Hadd AG, Latham GJ, Tassone F. 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.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. 2011. LOVD v.2.0: The next generation in gene variant databases. Hum Mutat 32:557–563.

Gao R, Matsuura T, Coolbaugh M, Zühlke C, Nakamura K, Rasmussen A, Siciliano MJ, Ashizawa T, Lin X. 2008. Instability of expanded CAG/CAA repeats in spinocerebellar ataxia type 17. Eur J Hum Genet 16:215–222.

Guo X, Zheng S, Dang H, Pace RG, Stonebraker JR, Jones CD, Boellmann F, Yuan G, Haridass P, Fedrigo O, Corcoran DL, Seibold MA, et al. 2014. Genome Reference and Sequence Variation in the Large Repetitive Central Exon of Human MUC5AC. Am J Respir Cell Mol Biol 50:223–232.

Hagerman R, Hagerman P. 2013. Advances in clinical and molecular understanding of the FMR1 premutation and fragile X-associated tremor/ataxia syndrome. Lancet Neurol 12:786–798.

Harrison CJ, Jack EM, Allen TD, Harris R. 1983. The fragile X: a scanning electron microscope study. J Med Genet 20:280–285.

Hestand MS, Houdt J Van, Cristofoli F, Vermeesch JR. 2016. Polymerase Specific Error Rates and Profiles Identified by Single Molecule Sequencing. Mutat Res 784–785:39–45.

Holloway TP, Rowley SM, Delatycki MB, Sarsero JP. 2011. Detection of interruptions in the GAA trinucleotide repeat expansion in the FXN gene of Friedreich ataxia. Biotechniques 50:182–186.

Jacquemont S, Birnbaum S, Redler S, Steinbach P, 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.

Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, Sun Q, Sherman T, Hu X, Jones K, Raley C, Tran B, Munroe DJ, 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.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32:493–496.

Kroutil LC, Register K, Bebenek K, Kunkel TA. 1996. Exonucleolytic proofreading during replication of repetitive DNA. Biochemistry 35:1046–1053.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, et al. 2011. The European nucleotide archive. Nucleic Acids Res 39:D28–D31.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, Hagerman PJ. 2013. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res 23:121–128.

Matsuura T, Fang P, Pearson CE, Jayakar P, Ashizawa T, Roa BB, Nelson DL. 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.

Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Müller CR, Pratt V, Wallace A. 2010. A standardized framework for the validation and verification of

clinical molecular genetic tests. Eur J Hum Genet 18:1276–1288.

Menon RP, Nethisinghe S, Faggiano S, Vannocci T, Rezaei H, Pemble S, Sweeney MG, Wood NW, Davis MB, Pastore A, Giunti P. 2013. The Role of Interruptions in polyQ in the Pathology of SCA1. PLoS Genet 9:e1003648. doi:10.1371/journal.pgen.1003648.

Monaghan KG, Lyon E, Spector EB. 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.

Mononen T, Von Koskull H, Airaksinen RL, 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.

Musova Z, Mazanec R, Krepelova A, Ehler E, Vales J, Jaklova R, Prochazka T, Koukal P, Marikova T, Kraus J, Havlovicova M, Sedlacek Z. 2009. Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. Am J Med Genet Part A 149A:1365–1369.

Natesan S a, Bladon AJ, Coskun S, Qubbaj W, Prates R, Munne S, Coonen E, Dreesen JCFM, Stevens SJC, Paulussen ADC, Stock-Myer SE, Wilton LJ, et al. 2014. Genome-wide karyomapping accurately identifies the inheritance of single-gene defects in human preimplantation embryos in vitro. Genet Med 16:838–45.

Nolin SL, Glicksman A, Ersalesi N, Dobkin C, Brown WT, Cao R, Blatt E, Sah S, Latham GJ, Hadd AG. 2014. Fragile X full mutation expansions are inhibited by one or more AGG interruptions in premutation carriers. Genet Med 17:358–364.

Nolin SL, Sah S, Glicksman A, Sherman SL, Allen E, Berry-Kravis E, Tassone F, Yrigollen

C, Cronister A, Jodah M, Ersalesi N, Dobkin C, 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.

Oberlé I, Rousseau F, Heitz D, Kretz C, Devys D, Hanauer A, Boué J, Bertheas MF, Mandel JL. 1991. Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science 252:1097–1102.

Penagarikano O, Mulle JG, Warren ST. 2007. The pathophysiology of fragile x syndrome. Annu Rev Genomics Hum Genet 8:109–129.

Pham TT, Yin J, Eid JS, Adams E, Lam R, Turner SW, Loomis EW, Wang JY, Hagerman PJ, Hanes JW. 2016. Single-locus enrichment without amplification for sequencing and direct detection of epigenetic modifications. Mol Genet Genomics 291:1–14.

Pieretti M, Zhang F, Fu Y-H, Warren ST, Oostra BA, Caskey CT, Nelson DL. 1991. Absence of expression of the FMR-1 gene in fragile X syndrome. Cell 66:817–822.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Radvansky J, Ficek A, Minarik G, Palffy R, 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.

Sermon K, Seneca S, Vanderfaeillie A, Lissens W, Joris H, Vandervorst M, Van Steirteghem A, Liebaers I. 1999. Preimplantation diagnosis for fragile X syndrome based on the detection of the non-expanded paternal and maternal CGG. Prenat Diagn 19:1223–1230.

Shin SC, Ahn DH, Kim SJ, Lee H, Oh T-J, Lee JE, Park H. 2013. Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PLoS One 8:e68824.

doi:10.1371/journal.pone.0068824.

Sullivan AK, Marcus M, Epstein MP, Allen EG, Anido AE, Paquin JJ, Sherman SL. 2005. Association of FMR1 repeat size with ovarian dysfunction. 20:402–412.

Tassone F, Iong KP, Tong T-H, Lo J, Gane LW, Berry-Kravis E, Nguyen D, Mu LY, Laffin J, Bailey DB, Hagerman RJ. 2012. FMR1 CGG allele size and prevalence ascertained through newborn screening in the United States. Genome Med 4:100.

Usdin K, Hayward BE, Kumari D, Lokanga RA, Sciascia N, Zhao XN. 2014. Repeat-mediated genetic and epigenetic changes at the FMR1 locus in the Fragile X-related disorders. Front Genet 5:226.

Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, Reiner O, Richards S, Victoria MF, Zhang FP, Eussen BE, van Ommen GB, 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.

Yrigollen CM, Durbin-Johnson B, Gane L, Nelson DL, Hagerman R, Hagerman PJ, Tassone F. 2012. AGG interruptions within the maternal FMR1 gene reduce the risk of offspring with fragile X syndrome. Genet Med 29:997–1003.

Yrigollen CM, Martorell L, Durbin-Johnson B, Naudo M, Genoves J, Murgia A, Polli R, Zhou L, Barbouth D, Rupchock A, Finucane B, Latham GJ, et al. 2014. AGG interruptions and maternal age affect FMR1 CGG repeat allele stability during transmission. J Neurodev Disord 6:24.

Yu S, Pritchard M, Kremer E, Lynch M, Nancarrow J, Baker E, Holman K, Mulley JC, Warren ST, Schlessinger D, Sutherland GR, Richards RI. 1991. Fragile X genotype characterized by an unstable region of DNA. Science 252:1179–1181.

Yu Z, Zhu Y, Chen-Plotkin AS, Clay-Falcone D, McCluskey L, Elman L, Kalb RG, Trojanowski JQ, Lee VMY, van Deerlin VM, Gitler AD, Bonini NM. 2011. PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. PLoS One 6:14–19.

Zamani Esteki M, Dimitriadou E, Mateiu L, Melotte C, Vander Aa N, Kumar P, Das R, Theunis K, Cheng J, Legius E, Moreau Y, Debrock S, et al. 2015. Concurrent Whole-Genome Haplotyping and Copy-Number Profiling of Single Cells. Am J Hum Genet 96:894–912.

Figure Legends

Figure 1: Overview of the workflow. First, PCR amplicons are generated whereby a different barcode is introduced for each amplicon (A). Next, different amplicons are pooled together and sequenced by PacBio single-molecule sequencing (B). The long reads generated by single-molecule sequencing allow to cross a circulated double-stranded template molecule multiple times (C) from which reads-of-insert with a high quality are generated (C). After selecting the on-target reads, the distribution of the read sizes of those reads-of insert was plotted (D), 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). Note that the CGG repeat is described as a GGC repeat according to the nomenclature of the Human Genome Variation Society (HGVS).
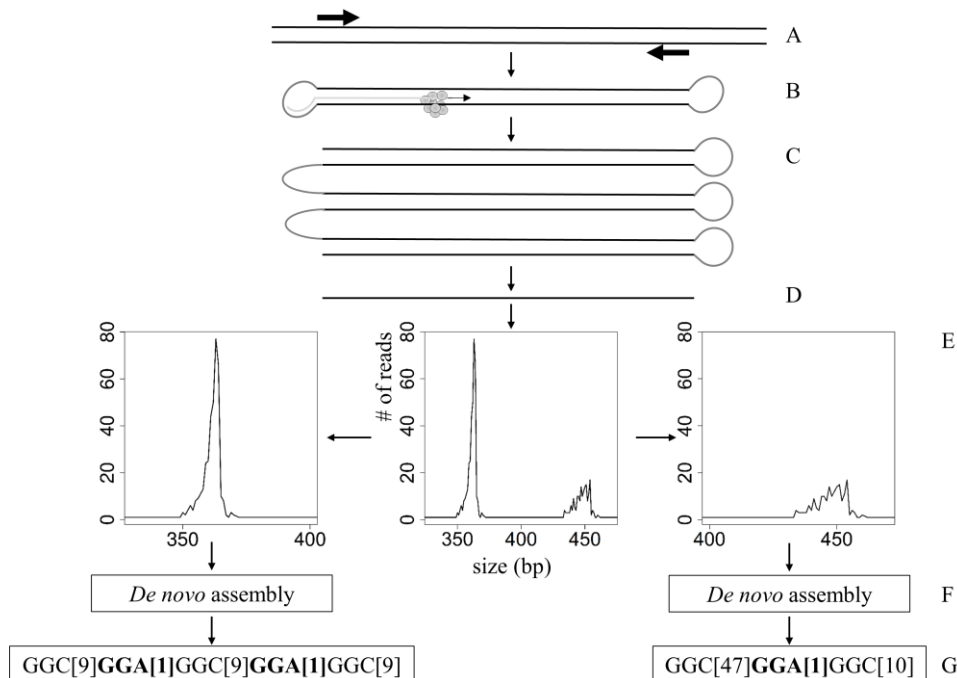
Figure 2: Correlation of the repeat size between single-molecule sequencing (Y axis) and PCR (X-axis) for: 7 male samples (A). The normal allele of all female samples (B). The grey-zone/premutation allele of all female samples (C). Samples are indicated by a dot, the grey lines show the error range of PCR which is ± 1 repeat unit for repeats smaller than 54 repeat units, ± 3 repeat units up until 80 CGG units and ±10% of the repeat size for repeats larger than 80 repeats.
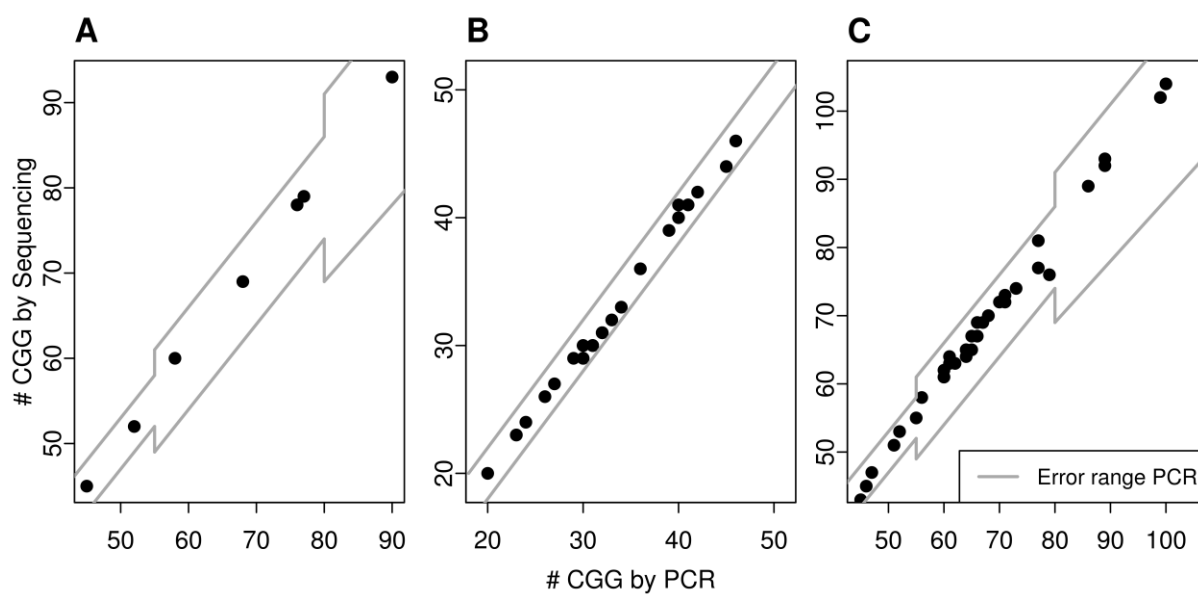
Figure 3: (A) Distribution of the read sizes of the reads-of insert for female 32 (A). The two alleles only differ by 5 repeat units and hence are difficult to separate based on differences in the size of the reads-of-insert. (B) Relation between the number of CGG units (X axis) and AGG interruptions (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.
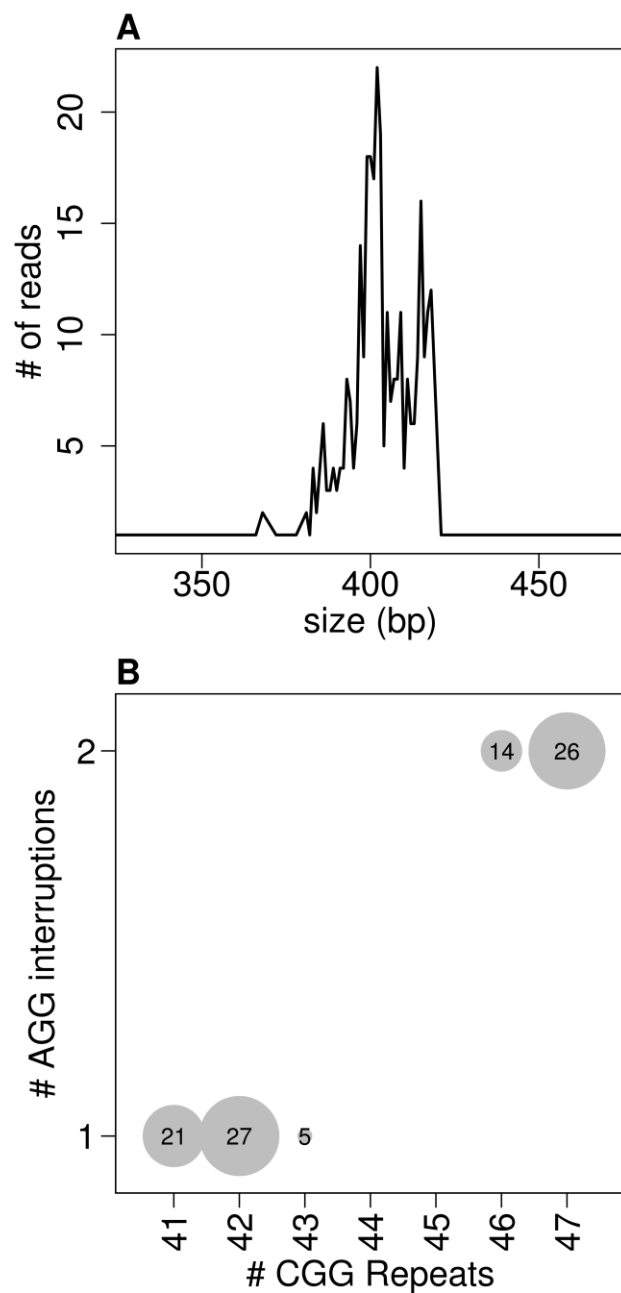
Table 1: Repeat characteristics for all male individuals.

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

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

Table 2: 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	# GGA	Repeat structure*	# Units	# GGA	Repeat structure*
1	29	56	29	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]	58	1	GGC[47]GGA[1]GGC[10]
2	33	51	32	2	GGC[9]GGA[1]GGC[12]GGA[1]GGC[9]	51	4	GGC[10]GGA[1]GGC[9]GGA[1]GGC[9] GGA[1]GGC[9]GGA[1]GGC[10]
3	40	68	40	2	GGC[20]GGA[1]GGC[9]GGA[1]GGC[9]	70	1	GGC[60]GGA[1]GGC[9]
4	30	65	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	67	0	GGC[67]
5	30	71	29	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]	73	2	GGC[53]GGA[1]GGC[9]GGA[1]GGC[9]
6	32	68	31	2	GGC[10]GGA[1]GGC[9]GGA[1]GGC[10]	70	2	GGC[50]GGA[1]GGC[9]GGA[1]GGC[9]
7	31	71	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	72	2	GGC[52]GGA[1]GGC[9]GGA[1]GGC[9]
8	26	89	26	1	GGC[16]GGA[1]GGC[9]	92	2	GGC[72]GGA[1]GGC[9]GGA[1]GGC[9]
9	32	60	31	2	GGC[10]GGA[1]GGC[9]GGA[1]GGC[10]	61	1	GGC[49]GGA[1]GGC[11]
10	31	61	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	64	1	GGC[54]GGA[1]GGC[9]
11	30	86	29	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]	89	1	GGC[79]GGA[1]GGC[9]
12	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
13	31	79	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	76	0	GGC[76]
14	30	89	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	93	2	GGC[73]GGA[1]GGC[9]GGA[1]GGC[9]
15	31	70	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	72	1	GGC[62]GGA[1]GGC[9]
16	23	67	23	1	GGC[9]GGA[1]GGC[13]	69	1	GGC[59]GGA[1]GGC[9]
17	36	99	36	1	GGC[25]GGA[1]GGC[10]	102	1	GGC[92]GGA[1]GGC[9]
18	31	66	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	69	0	GGC[69]
19	24	61	24	1	GGC[9]GGA[1]GGC[14]	63	1	GGC[53]GGA[1]GGC[9]
20	32	77	31	2	GGC[9]GGA[1]GGC[11]GGA[1]GGC[9]	77	1	GGC[67]GGA[1]GGC[9]
21	32	100	31	2	GGC[10]GGA[1]GGC[9]GGA[1]GGC[10]	104	2	GGC[86]GGA[1]GGC[7]GGA[1]GGC[9]
22	31	64	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	65	2	GGC[45]GGA[1]GGC[9]GGA[1]GGC[9]
23	27	52	27	1	GGC[17]GGA[1]GGC[9]	53	2	GGC[33]GGA[1]GGC[9]GGA[1]GGC[9]
24	31	66	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	67	2	GGC[47]GGA[1]GGC[9]GGA[1]GGC[9]
25	30	62	29	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]	63	2	GGC[43]GGA[1]GGC[9]GGA[1]GGC[9]

N°	# Units	# Units	# Units	# GGA	Repeat structure*	# Units	# GGA	Repeat structure*
26	23	60	23	1	GGC[10]GGA[1]GGC[12]	62	2	GGC[42]GGA[1]GGC[9]GGA[1]GGC[9]
27	34	77	33	2	GGC[13]GGA[1]GGC[9]GGA[1]GGC[9]	81	1	GGC[71]GGA[1]GGC[9]
28	39	64	39	0	GGC[39]	65	2	GGC[45]GGA[1]GGC[9]GGA[1]GGC[9]
29	30	60	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	61	1	GGC[51]GGA[1]GGC[9]
30	20	64	20	1	GGC[9]GGA[1]GGC[10]	64	2	GGC[44]GGA[1]GGC[9]GGA[1]GGC[9]

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

Table 3: Repeat characteristics for four females with similarly sized repeats.

| | (TP)-PCR | | SINGLE-MOLECULE SEQUENCING | | | | | |
| | ALLELE 1 | ALLELE 2 | ALLELE 1 | | | ALLELE 2 | | |
N°	# Units	# Units	# Units	# GGA	Repeat structure*	# Units	# GGA	Repeat structure*
31	41	46	41	2	GGC[20]GGA[1]GGC[9]GGA[1]GGC[10]	45	2	GGC[25]GGA[1]GGC[9]GGA[1]GGC[9]
32	42	47	42	1	GGC[21]GGA[1]GGC[20]	47	2	GGC[27]GGA[1]GGC[9]GGA[1]GGC[9]
33	40	45	41	2	GGC[21]GGA[1]GGC[9]GGA[1]GGC[9]	43	2	GGC[23]GGA[1]GGC[9]GGA[1]GGC[9]
34	65	73	65	2	GGC[45]GGA[1]GGC[9]GGA[1]GGC[9]	74	2	GGC[56]GGA[1]GGC[7]GGA[1]GGC[9]

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

Table 4: 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	# GGA	Repeat structure*	# Units	# GGA	Repeat structure*
2	1	40	33	51	32	2	GGC[9]GGA[1]GGC[12]GGA[1]GGC[9]	51	4	GGC[10]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]
	2	40	33	51	32	2	GGC[9]GGA[1]GGC[12]GGA[1]GGC[9]	51	4	GGC[10]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]
	2	40	33	51	32	2	GGC[9]GGA[1]GGC[12]GGA[1]GGC[9]	51	4	GGC[10]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]
	2	40	33	51	32	2	GGC[9]GGA[1]GGC[12]GGA[1]GGC[9]	51	4	GGC[10]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]
	3	40	33	51	32	2	GGC[9]GGA[1]GGC[12]GGA[1]GGC[9]	51	4	GGC[10]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]
12	1	40	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
	2	40	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
	2	40	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
	2	40	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
	3	40	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
	3	5	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
	3	100	31	55	30	2	GGC[9]GGA[1]GGC[9]GGA[1]GGC[10]	55	2	GGC[37]GGA[1]GGC[7]GGA[1]GGC[9]
17	1	40	36	99	36	1	GGC[25]GGA[1]GGC[10]	104	1	GGC[94]GGA[1]GGC[9]
	2	40	36	99	36	1	GGC[25]GGA[1]GGC[10]	102	1	GGC[92]GGA[1]GGC[9]
	2	40	36	99	36	1	GGC[25]GGA[1]GGC[10]	103	1	GGC[93]GGA[1]GGC[9]
	2	40	36	99	36	1	GGC[25]GGA[1]GGC[10]	104	1	GGC[94]GGA[1]GGC[9]
	3	40	36	99	36	1	GGC[25]GGA[1]GGC[10]	104	1	GGC[94]GGA[1]GGC[9]

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