


Single-locus enrichment without amplification for sequencing and direct detection of epigenetic modifications

Thang T. Pham¹ · Jun Yin^{2,3} · John S. Eid^{1,4} · Evan Adams² · Regina Lam¹ · Stephen W. Turner¹ · Erick W. Loomis^{2,5} · Jun Yi Wang² · Paul J. Hagerman² · Jeremiah W. Hanes¹ 

Received: 18 June 2015 / Accepted: 6 January 2016 / Published online: 29 January 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract A gene-level targeted enrichment method for direct detection of epigenetic modifications is described. The approach is demonstrated on the CGG-repeat region of the *FMR1* gene, for which large repeat expansions, hitherto refractory to sequencing, are known to cause fragile X syndrome. In addition to achieving a single-locus enrichment of nearly 700,000-fold, the elimination of all amplification steps removes PCR-induced bias in the repeat count and preserves the native epigenetic modifications of the DNA. In conjunction with the single-molecule real-time sequencing approach, this enrichment method enables direct readout of the methylation status and the CGG repeat number of the *FMR1* allele(s) for a clonally derived

cell line. The current method avoids potential biases introduced through chemical modification and/or amplification methods for indirect detection of CpG methylation events.

Keywords Targeted enrichment · Single molecule sequencing · *FMR1* · Fragile X syndrome · Epigenetic modification · Tandem repeats

Introduction

The inability to sequence microsatellite DNA expansions associated with a broad range of clinical disorders impedes the characterization of these loci and hampers epigenetic mapping within many of these regions (Kieleczawa 2006; Mirkin 2007; Walker 2007; Deaton and Bird 2011; Marmolino 2011; Udd and Krahe 2012; Evans-Galea et al. 2013; Nelson et al. 2013). Therefore, development of a targeted enrichment methodology is essential to the epigenetics study of these regions. At present, several different enrichment methods have been employed for such investigations; however, none of them can be used for direct genomic DNA-level epigenetic analysis. Polymerase chain reaction (PCR) can routinely target regions of the genome up to ~10 kb in length, but suffers the dual disadvantages of being an error-prone replication method, particularly for amplification of microsatellite sequence (Loomis et al. 2013) and regions of extreme GC content (Mutter and Boynton 1995; Kieleczawa 2006), and of destroying information about the methylation state of the sequence. Methylation information can now be read directly from genomic DNA through single molecule real-time (SMRT) DNA sequencing (Flusberg et al. 2010), however the SMRT methodology does not intrinsically focus sequencing from a sample embodying the whole genome onto a single locus.

Communicated by S. Hohmann.

T. T. Pham and J. Yin contributed equally to this manuscript.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-016-1167-2) contains supplementary material, which is available to authorized users.

✉ Jeremiah W. Hanes
jhanes@pacb.com

- ¹ Pacific Biosciences, Menlo Park, CA 94025, USA
- ² Department of Biochemistry and Molecular Medicine, University of California, Davis, School of Medicine, Davis, CA 95616, USA
- ³ Present Address: Dendrite Morphogenesis and Plasticity Unit, National Institute of Neurological Disorders and Stroke, Bethesda, MD 20892, USA
- ⁴ Present Address: Whole Biome, Inc., San Francisco, CA 94107, USA
- ⁵ Present Address: Faculty of Medicine, Department of Surgery & Cancer, Institute of Reproductive and Developmental Biology, Hammersmith Campus Imperial College London, London, UK

Hybridization capture methods (Mamanova et al. 2010; Teer et al. 2010) have been widely used in exome sequencing (Choi et al. 2009), resulting in extensive enrichment and focused sequencing of exomes. However, such methods do not presently yield fragments long enough to exploit the long-read technologies now available for detection of structural variations and phasing of mutations. Ligation-based target-enrichment methods have been applied to good effect on panels of genes, but because these methods rely on circularization of DNA (Dahl et al. 2005), limitations on the kinetics of ligation-based circle closure limit the applicability of these methods to fragments well below a kilobase in length. In addition, the available implementations of these methods still rely on PCR to produce an amount of targeted material suitable for sequencing, which destroys the epigenetic modifications.

Electrophoretic techniques, such as synchronous coefficient of drag alteration (SCODA) that are suitable for processing large amounts of input material, are being adapted to the task of target enrichment (So et al. 2010), but these methods are so far limited to short fragments and have the disadvantage of unlinking the sense and antisense strands of duplex DNA, confounding the analysis of hemi-methylation (Murray et al. 2012). To circumvent the limitations described above, we present a method for enriching a specific genomic locus that does not rely on amplification, thus preserving the methylation information contained in the genomic fragments, as well as inter-strand linkage information.

As a specific example of the applicability of our method, we have focused on the fragile X (*FMR1*) locus, where CGG-repeat expansions and epigenetic silencing give rise to fragile X syndrome (FXS), the leading heritable form of intellectual disability and leading single-gene form of autism (Hagerman et al. 2010), and the fragile X-associated disorders, including the neurodegenerative disorder, fragile X-associated tremor/ataxia syndrome (FXTAS) (Hagerman 2013). In the United States, the carrier frequency for expanded-repeat alleles is approximately 0.5 %, and a much larger fraction (~3 %) is indicated for testing based on increased risk (Hagerman and Hagerman 2013). There is a complex relationship between the size of the CGG-repeat and the nature of the clinical phenotype, with distinct CGG-repeat ranges corresponding to qualitatively distinct groups of patient outcomes (Gallagher and Hallahan 2012; Hagerman and Hagerman 2013). Further complicating the molecular analysis of the *FMR1* locus is the fact that its methylation state is an important modifier of the phenotypic impact of the repeat expansion on the affected individual.

To date, no method has been capable of direct analysis (i.e., avoiding bisulfite modification, cloning, and/or PCR amplification) of the patterns and extent of methylation

across the promoter region of the *FMR1* locus, particularly within the CGG-repeat. Recently, it was demonstrated that SMRT sequencing is capable of sequencing the CGG-repeat region, even for highly expanded CGG-repeat alleles in the full mutation range (>200 CGG repeats) (Loomis et al. 2013), despite its highly repetitive structure and 100 % GC content. However, the locus was isolated using either cloning or PCR to provide the necessary enrichment, resulting in loss of the methylation status of the gene. In the current work, we report that a combination of single-locus (*FMR1*) enrichment/capture, coupled with the unique capability of SMRT sequencing to follow the kinetics of nucleotide incorporation, facilitates the direct mapping of methylated cytosines at the level of genomic DNA.

Materials and methods

Restriction enzyme DNA fragmentation

Genomic DNA from the lymphoblastoid cell line, AG09391 (“AG”; NIA Cell Repository) from a normal female (16, 29 CGG-repeat alleles, by PCR-based sizing) (Tassone et al. 2000; Primerano et al. 2002; Arocena et al. 2005), was extracted and purified to remove traces of RNA, ssDNA, and contaminants that interfere with restriction enzyme (RE) and ligase activity. The gDNA was digested to completion using type IIS restriction enzymes, Bsm AI or Bco DI (isoschizomer of Bsm AI) (NEB), in the optimal buffer. For preparation of each single-locus capture library, 18–20 µg gDNA was digested at 55 °C for 16 h at a final concentration of 20 µg/mL. Five units of Bsm AI were used per microgram of gDNA. The efficiency of RE digestion was verified by PCR using primers across the RE sites. Bsm AI has 5-base recognition sequences, GTCTC and GAGAC, so the average Bsm AI-digested fragment is 512 bp ($=4^5/2$) assuming perfectly random sequence. For example, a 6.4 Gb genome of a typical female would yield $\sim 12.5 \times 10^6$ fragments [$=6.406 \times 10^9 / (4^5/2)$]. Each end of the Bsm AI-fragments has a 4-base overhang determined only by the local sequence at the site of cleavage resulting in 256 ($=4^4$) different 4-base combinations. Therefore, a specific Bsm AI-fragment with the same two ends would be found only once in every 65,536 fragments ($4^4 \times 4^4$).

Adapter ligation

Based on the estimated 12.5×10^6 fragments created by Bsm AI, and an added specificity of 256-fold for each 4-base adapter-end, the ligation step using two sequence-specific adapters is expected to generate 764 fragments with adapters at both ends, ~200,000 fragments with only one adapter, and $>10^6$ fragments with no adapters (Table

S1). Among the 764 molecules that should have adapters at both ends to form cyclized SMRTbell templates, 573 (75 %) have at least one Bsm AI recognition site inside the fragment. Each fragment has either 0 (~25 %), 1 (~50 %), or 2 (~25 %) Bsm AI recognition sites, which can be recut by Bsm AI. Thus, Bsm AI was allowed to remain active during and after the ligation step to destroy these non-target molecules. For these reasons, the Bsm AI digestion was used directly for the adapter ligation reaction. Two specific hairpin adapters were designed with a 5'-CTGT overhang and a 5'-AATG overhang, respectively, such that the 5'-end of each adapter has a single-strand overhang that is complementary to the targeted 1.1 kb *FMRI* fragment. The sequences of adapter A and adapter B were 5'-pCTGTATCTCTCTCTTTTGCTCCTCCTCCTCCGTTGATTGTTGTTGGAGAGAGAT and 5'-pAATGATCTCTCTCTTTTGCTCCTCCTCCTCCGTTGATTGTTGTTGGAGAGAGAT, respectively.

A stoichiometric excess of the hairpin adapters is required to minimize self-ligation of the fragments. For high-fidelity sticky-end ligation, *E. coli* ligase (NEB) was found to be superior to T4 DNA ligase, as the latter was much more permissive of the ligation of non-complementary ends. Thus 200 nM of each adapter was incubated with 20 µg/ml of Bsm AI-digested DNA fragments (50 nM, based on an estimated average size of 512 bp for the DNA fragments) in 1× *E. coli* ligase buffer for 30 min at 37 °C. The ligation reaction was then started by adding *E. coli* ligase (0.15 U/µl *E. coli* ligase; ~10 U ligase per µg DNA fragments) followed by an additional incubation at either RT (~22 °C) or 37 °C (comparison in Table 1) for ~16 h.

DNA size selection

Following the ligation step, 0.35× and 0.65× volumes of AMPure beads were used for DNA clean-up and size selection. Since the targeted *FMRI* fragment is about 1.1 kb, DNA fragments between 0.5 and 3 kb were selected and purified from the ligation reaction. First, 0.35× volume

of washed AMPure beads (PacBio) was added into the ligation products to remove DNA fragments larger than 3 kb. After mixing at 500 rpm using a vortex mixer for 10 min, the beads were separated on the wall of the tube using a magnetic stand. The supernatant, with DNA fragments less than ~3 kb, was transferred to a new tube. An additional 0.30× volume of AMPure beads was added into the reserved supernatant such that the final bead volume is 0.65× of the original sample. After mixing at 500 rpm for 10 min, DNA fragments larger than 500 bp were attracted to the magnetic beads. The DNA fragments were cleaned further by two 75 % ethanol washes, and then eluted from beads using 10 mM Tris-HCl, pH 8.0 (or EB buffer from Qiagen).

Digestion of non-target DNA fragments

DNA fragments with 0 or 1 adapter (non-cyclized) are good substrates for exonuclease III (NEB) and Exonuclease VII (USB), whereas the fully cyclized fragments with two adapters should be resistant to the exonucleases. When all non-cyclized fragments have been eliminated, the quantity of double-stranded DNA (dsDNA) should be ~65,000-fold (accounting for the two-adapter fragments that can be recut by Bsm AI) lower than the starting material (e.g., ~300 pg DNA from 20 µg of starting gDNA). In practice, the Exo-treatment is stopped when the remaining quantity of dsDNA is ~50 ng, in order to maintain a sufficient amount of gDNA to carry out downstream steps.

In the exonuclease treatment reaction, 1.7 unit/µl Exo III and 0.1 unit/µl Exo VII were used for 100 ng/µl DNA in 1× NEBuffer 3. The reaction was incubated for 1–2 h at 37 °C, and a Qubit fluorometer was used to monitor the concentration of dsDNA. The reaction was stopped when total dsDNA was reduced to ~45 to 50 ng. Remaining DNA was purified by using 0.65× volume of AMPure beads.

T7 Exonuclease, Exonuclease I, and Rec Jf (NEB) were found to have lower endonuclease activity compared to Exo III and Exo VII. These exonucleases were used for

Table 1 Enrichment efficiency depends on ligation conditions

Sample	Total post-filtered reads	Mapped reads to human	Mapped reads to <i>FMRI</i>	<i>FMRI</i> specificity	Enrichment ^a
<i>E. coli</i> ligase @ 37 °C	4517	2968	325	0.1095	685,198
<i>E. coli</i> ligase @ 22 °C	4523	3350	278	0.0830	519,142
T4 ligase @ 37 °C	10,807	8694	222	0.0255	159,751
T4 ligase @ 22 °C	14,378	12,330	246	0.0200	124,812
T4 ligase @ 16 °C (no active Bsm AI during ligation)	46,675	42,707	46	0.001077	6738

Bsm AI-digested human female diploid (6.406×10^9 bp): 512 bp average fragment size

^a Fold of enrichment = fraction of *FMRI* read/fraction of *FMRI* fragment after Bsm AI digest = fraction of *FMRI* read/[2/($6.406 \times 10^9/512$)] using genomic DNA from the lymphoblastoid cell line, AG09391 from a normal female (16, 29 CGG-repeat alleles)

the male gDNA samples (samples with expanded CGG repeat allele). A combination of 10 units of T7 Exo, 10 units of Exo I, and 10 units of Rec Jf per μg DNA were used; each sample had a concentration of 100 ng/ μl DNA in $1\times$ NEBuffer 4. A carrier supercoiled plasmid (pUC18 or pBR322), 500 ng, was added to each DNA sample before the Exo-digestion to aid in the recovery of the enriched templates during AMPure purification. The reaction was incubated for 4 h at 37 °C. The extent of exonuclease treatment was evaluated by measuring the amount of remaining dsDNA by Qubit, where the reaction was considered complete as it approached the amount of plasmid carrier added (500 ng). The enriched templates and plasmid DNA were purified from the reaction by using $0.65\times$ volume of AMPure beads.

Annealing primer to the enriched template with ligated adapters

A sequencing primer that has reversed and complementary sequence to the loop region of the adapter are used for polymerase binding and DNA synthesis. The primer sequence is: 5'-CAACGGAGGAGGAGGAGC-3' (IDT, Iowa City, Iowa). The ratio of primer concentration to template concentration is approximately 10, such that all templates with hairpin adapters can have 2 primers per template. Hybridization of the primer to the template was carried out in $1\times$ Primer buffer (10 mM Tris-OAc, pH 8, 12 mM KOAc) using a thermocycler setting at 70 °C for 5 min, and temperature decreases by 0.1 °C per second until it reaches 22 °C. The primer-annealed templates can be stored at 4 °C.

Formation of polymerase-template complexes

To form the polymerase-template complex for primer extension and final sequencing, C2 or P5 polymerase (PacBio) was bound to the primer-annealed templates. In the reaction, 30 nM of polymerase was incubated with 0.5 ng/ μl primer-annealed templates (0.7–10 nM) in buffer containing 10 mM Tris-OAc, pH 8.0, 10 mM KOAc, 0.05 % Tween-20, 40 mM DTT, 0.4 mM Strontium acetate, 1 μM dNTP at 30 °C for at least 3 h.

Capture-hook hybrid selection method

To further enrich the targeted region, a capture-hook hybridization selection method (developed at PacBio as the SMRThook method) was performed. $0.1\times$ volume of 5 mM magnesium acetate was added to the polymerase-template complex for 30 min incubation at RT, thereby extending the annealed primers by ~30–50 bases to form open single-stranded DNA sections in the stem of the

SMRTbell template. The extension reaction was stopped by adding $0.07\times$ volume of 30 mM EDTA to the mixture (final 2 mM EDTA). After incubation for 5 min, a $0.1\times$ volume of 50 mM strontium acetate (final concentration of 5 mM Sr^{2+}) was added to stabilize the “open complex”. The single-stranded DNA in the “open complex” is specific for each template. A capture-hook DNA oligonucleotide is designed to have an 18 nucleotide probe sequence specific to the targeted *FMRI* open complex and a 23 nucleotide oligo-dA tail, allowing hybridization with $(\text{dT})_{25}$ oligos derivatized on the surface of magnetic beads in the following procedure. The sequence of *FMRI* capture-hook oligo is: 5'-CTAGCGCCTATCGAAATGGT-CAAAAAAAAAAAAAAAAAAAAAAAAAA-3'.

A $0.1\times$ volume of 2 μM capture-hook oligo was added to the “opened complex” solution such that the capture-hook concentration is about 200 nM (>200-fold excess of the targets). Since the concentration of salt is low in the “opened complex” sample (<10 mM KOAc), a $0.1\times$ volume of bead wash buffer (BWB; 400 mM KOAc; PacBio) is added to the sample to allow efficient hybridization of the capture-hook oligo to the opened complexes and to the $(\text{dT})_{25}$ oligos on beads. The hybridization reaction is carried out at RT using a rotating platform for 2 h. Then the opened complexes with the annealed probe oligo were captured on magnetic beads through the interaction of $(\text{dA})_{23}$ on the probe oligo and the $(\text{dT})_{25}$ oligos, which are covalently coupled to the beads. For each sample, 50 μl of magnetic beads- $(\text{dT})_{25}$ oligos was washed with 50 μl aliquots of BWB and then bead binding buffer (BBB) from PacBio. Before binding to the complexes, the BBB was removed from solid beads using the magnetic stand. The sample of open complexes hybridized to the probe oligo was applied to the solid beads, which were mixed well by gently pipetting up and down. The hybridization reaction is carried out at RT for 1 h using a rotating platform for efficient annealing of the capture-hook to the $(\text{dT})_{25}$ beads. Complexes that do not have the annealed capture-hook oligo were washed away from magnetic beads using the reagents and protocol described in the Bead-binding kit (PacBio).

The retained opened complexes on the magnetic beads, with the highly enriched targeted templates, were used for loading the active Pol-template complexes into ZMWs on a SMRT Cell for sequencing on the PacBio RS II system.

Sequencing and analysis on the PacBio RS II system

SMRT sequencing was carried out on the PacBio RS II (Pacific Biosciences, Menlo Park, CA, USA) using standard C2-C2 chemistry for bead-loading of SMRTbell libraries. Sequencing reads were processed and mapped to the respective reference sequences using the BLASR mapper

Fig. 1 A schematic of the amplification-free enrichment approach. Purified, unsheread genomic DNA is digested with specific type-IIS restriction enzymes (RE) selected to produce cuts on both ends of the desired target region. Ligation to hairpin adapters with complementary overhangs yields closed circular (SMRTbell) DNA, which is refractory to subsequent digestion with exonuclease types III and VII. Fully formed off-target SMRTbell templates can be cut through the use of additional REs (chosen to not cut within the desired target sequence) or the same RE (since many off target molecules will still maintain the recognition site within the SMRTbell template). The enriched region of interest is primer-annealed, and polymerase is added and allowed to extend by ~40 nucleotide into the locus-specific DNA region, thus allowing for further selectivity based on annealing of a locus-specific SMRThook oligonucleotide.

Fragmentation

- Type IIS Restriction Enzyme (RE)

Target Protection

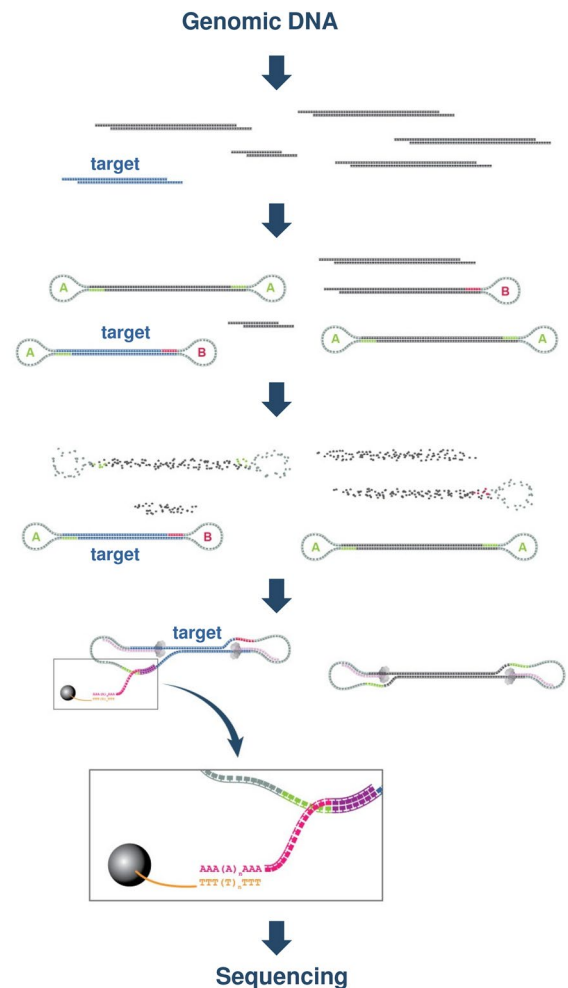
- Ligate sequence specific adaptors **A** and **B**

Complexity reduction

- Add off-target cutting REs
- Add Exonucleases III & VII

Hybridization selection

- Purify; add primer & polymerase
- Extend primer
- Capture exposed ssDNA section using magnetic beads with capture-probe oligonucleotide



the type IIS enzymes cut at a specific distance outside of their recognition sequence, cleavage yields DNA fragments with a single-stranded overhang sequence determined only by the local context at the site of cleavage. In the case of targeted enrichment, the overhang sequences are specified by the sequence at the locus of interest, thus allowing for the design of two independent hairpin adapters (“A” and “B” in Fig. 1) to specifically ligate on the ends to form a circular SMRTbell template that is resistant to exonuclease digestion. For four nucleotide overhangs, the specific overhang sequences at each end are expected to provide an additional 256-fold ($=4^4$) specificity or 65,536-fold ($=4^4 \times 4^4$) specificity for a retained DNA fragment. Ligation of the locus-specific adaptors to the digested genomic DNA, followed by exonuclease digestion of the unligated material, enriches the SMRTbell fraction from genomic DNA with zero or only one ligated adaptor (Fig. 1). Theoretical numbers for the target and non-target fragments are presented in Table S1 and are based on ideal conditions for the following steps: gDNA digestion by RE (Bsm AI), ligation of adaptors, and exonuclease digestion. The true number is

expected to be lower due to a number of potential factors: reduced RE efficiency due to methylation; contaminants in the DNA sample (e.g. ssDNA, RNA); mutation of the DNA, especially at the ends of the targeted fragment; and the specificity and extent of optimization of the enzymes (RE, ligase, exonuclease) themselves.

To further enrich the locus of interest, we included a sequence-specific “capture-hook” method in which the annealed primers are extended to form open single-stranded DNA sections (Fig. 1) in the stem of the SMRTbell template. This exposed single-stranded portion allows for targeted capture using an oligo containing 15–25 bases of locus-specific complementary sequence as well as a (dA)₂₃ tail to link the complex to (dT)₂₅-magnetic beads. Once captured on magnetic beads, the sample can be loaded directly onto the SMRT Cell for sequencing.

FMRI enrichment example

For the current application, genomic DNA was isolated and purified from an Epstein-Barr virus (EBV)-transformed

lymphoblastoid line (designated AG) derived from a normal female (~16 and ~29 CGG repeats were estimated previously by polyacrylamide gel electrophoresis) (Primerano et al. 2002). For preparation of each single locus capture library, 18.3 µg of genomic DNA (corresponding to 6.75 pg of the ~1.1 kbp target locus) was digested using the type IIS enzyme, Bsm AI (GTCTCN[^]NNNN), which leaves a 4-base overhang specified by the sequence context of the cut site. The 1.1 kb fragment of interest contains the CGG repeat site with an upstream 5'-ACAG overhang and a downstream 5'-CATT overhang. The resulting fragment pool is then circularized by ligation to specific adapters with overhang sequences that are complementary to the overhangs created by Bsm AI, thus yielding increased *FMRI* specificity (5'-CTGT on the upstream "A" adapter and 5'-AATG on the downstream "B" adapter). Note that, except for the 4-base overhangs, these sequences are similar to the standard SMRTbell preparation adapters (Travers et al. 2010). High-fidelity *E. coli* ligase (NEB) was used under specific conditions to reduce the fraction of off-target ligation ("Materials and methods").

When, as in this case, the target fragment does not embody the recognition sequence, the same restriction enzyme can be allowed to remain active during and after the ligation, so that the non-target fragments that have ligated adapters at both ends but do contain a Bsm AI recognition motif will be cut open once again. Fragments with at least one open end were then digested using exonucleases III and VII. Circular fragments closed at both ends are resistant to exonuclease activity. To render the locus-specific probe available for hybridization, the SMRTbell templates are primed in the hairpin region and bound with sequencing polymerase (Pacific Biosciences, C2 chemistry), and then extended in a solution that contains dNTPs and Mg²⁺, as well as Sr²⁺ to slow the reaction. The priming reaction is quenched with EDTA, and additional Sr²⁺ is added after exposure of ~40 bases of single-stranded insert DNA at one end of the *FMRI* fragment. After quenching, the resulting open-complex comprising the partially strand-displaced fragment, extended primer, and polymerase is annealed with the specific bridging oligo at 30 °C, and this target specific complex is captured by oligo-dT-derivatized magnetic beads ("Magbeads") as in the standard Magbead loading protocol. The retained complexes and the Magbeads were then applied to a SMRT Cell and sequenced. This process was repeated with appropriate modifications for a number of control samples as well.

Methylation-positive controls were prepared by performing an in vitro methylation of a synthetic 20-CGG-repeat containing molecule, and plasmid-derived 30-CGG-repeat containing species using SssI methyltransferase. The level of methylation was confirmed using bisulfite sequencing (Table S2 and Fig. S1).

Sequencing and analysis

The native targeted DNA sequencing run, from a sample using *E. coli* ligase at 37 °C (see Table 1), yielded 2968 reads that map to the human genome with non-mapping reads comprising mitochondrial sequences, adapter dimers, and other contaminating DNA. Of the reads that map to human (human_g1k_v37 reference), 325 of them (11 %) were specific to the 1.1 kbp *FMRI* fragment region, representing ~692-fold coverage (average of 2.1 sub-reads per molecule). A coverage map of the X-chromosome reveals a clear peak at the *FMRI* locus, and a read-map of this region indicates that the vast majority of reads begin and end where expected (Fig. 2). Without enrichment, one read in roughly 6.26×10^6 $[=(6.41 \times 10^9/512)/2]$ would be expected to map to this locus; therefore, an on-target rate of 11.0 % ($=325/2968$) corresponds to an estimated enrichment factor of ~688,600 ($=\text{fraction of } FMRI \text{ reads}/\text{fraction of } FMRI \text{ fragment in the RE digest} = 0.11 \times 6.26 \times 10^6$). The procedures were performed in duplicate with the exception of using different ligation temperatures, 37 and 22 °C; the number of targeted reads and specificity of enrichment were within approximately 20 % due to better ligation fidelity at higher temperature for sticky end ligation.

Sequences that mapped to the *FMRI* region and also possessed at least three subreads were selected for further analysis, as in Loomis et al. (2013). The most likely consensus sequence is arrived at through an algorithm that combines the subreads by taking into account the expected error profile (Chin et al. 2013). To further minimize homopolymer slip, only the CG or GC transitions are counted in estimating the repeat length. The results are shown in Fig. 3. There were two distinct populations of repeat lengths with modes at 20 and 30 repeats. This result is consistent with the earlier PCR-electrophoresis result (~16 and 29 CGG repeats), given the approximate nature of the PCR amplification process, and the current assumptions involved with assembly within the CGG repeat; therefore, we have used the values of 20 and 30 CGG repeats for the remainder of the current study. The standard deviations of the two clusters were 0.90 and 1.0 for the clusters at 20 and 30 repeats, respectively. The numbers of reads observed in each cluster (42 and 46, respectively) are consistent with a heterozygous female.

Kinetic analysis for methylation

The information provided by SMRT sequencing inherently includes the kinetics of each nucleotide incorporation, which has been shown to inform on the methylation status of DNA. Because the data are from unamplified gDNA, cytosine methylation remains intact and can thus be directly queried without bisulfite conversion or similar approach. The inter-pulse duration (the interval between the end of a

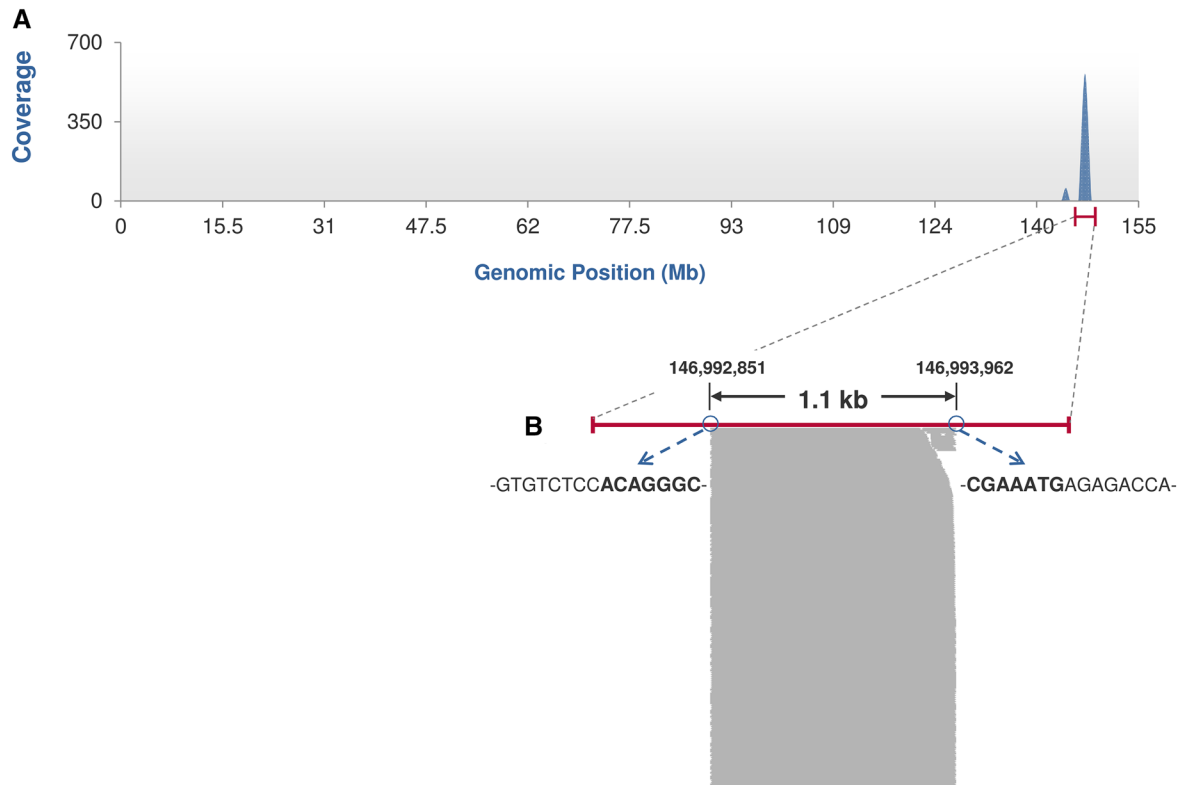


Fig. 2 **a** Coverage map of the entire X-chromosome showing the main localization at the *FMR1* region with 692 \times coverage (red bar) with one minor off-target site that contains both Bsm AI cleavage

sites and a poly(A) tract that is non-specifically bound to the beads. **b** Zoom in on the area immediately surrounding the *FMR1* region, demonstrating the precise restriction site ends of the reads

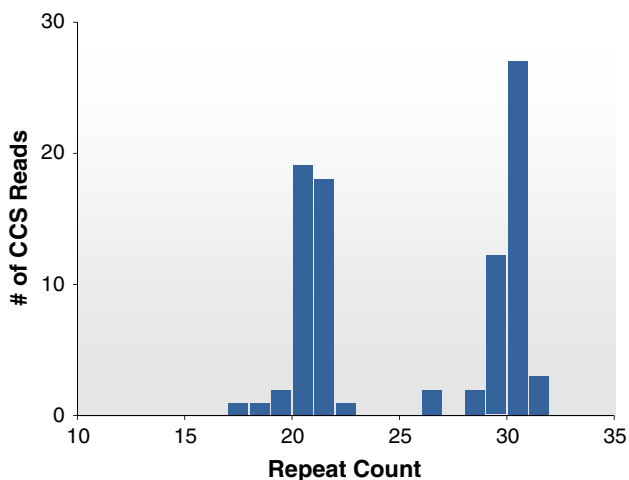


Fig. 3 Histogram of the CGG-repeat length from circular consensus sequence (CCS) reads, which shows that there are two populations that represent the two alleles present in this clonal female lymphoblast cell line

sequencing pulse and beginning of the subsequent pulse-IPD) is perturbed by the presence of many chemical modifications of the template, and is also sensitive to the local sequence context (Flusberg et al. 2010). Thus, a kinetic

reference representing the expected kinetics for unmodified DNA is needed to distinguish sequence context effects from actual modifications. The usual in silico reference approach relies on training data which, at present, does not contain a sufficient sampling of this rare repeat motif. Accordingly, for the analysis of the current data, an unmethylated sample with identical sequence was created using multiple displacement amplification (Hutchison et al. 2005) and sequenced under the same conditions. The upper row plots in Fig. 4 depict a comparison of the observed mean IPD values with associated standard errors of nucleotide incorporation for the forward strand of the two alleles between the native genomic DNA (red bars) and the amplified (and thus unmethylated) reference sample (blue bars). The pattern of sequence-context dependent kinetic variation that is common between the two samples is clearly visible and, therefore, it is convenient to plot the ratio of the native to unmethylated IPDs (bottom row of Fig. 4) which reveals methylation of the forward strand in the 20 CGG allele but not in the 30 CGG allele. All 4 enriched samples from the same female gDNA in Table 1 showed qualitatively similar patterns as the data shown in Fig. 4.

To confirm this finding, synthetic oligonucleotides reflecting the CGG repeat and flanking regions were

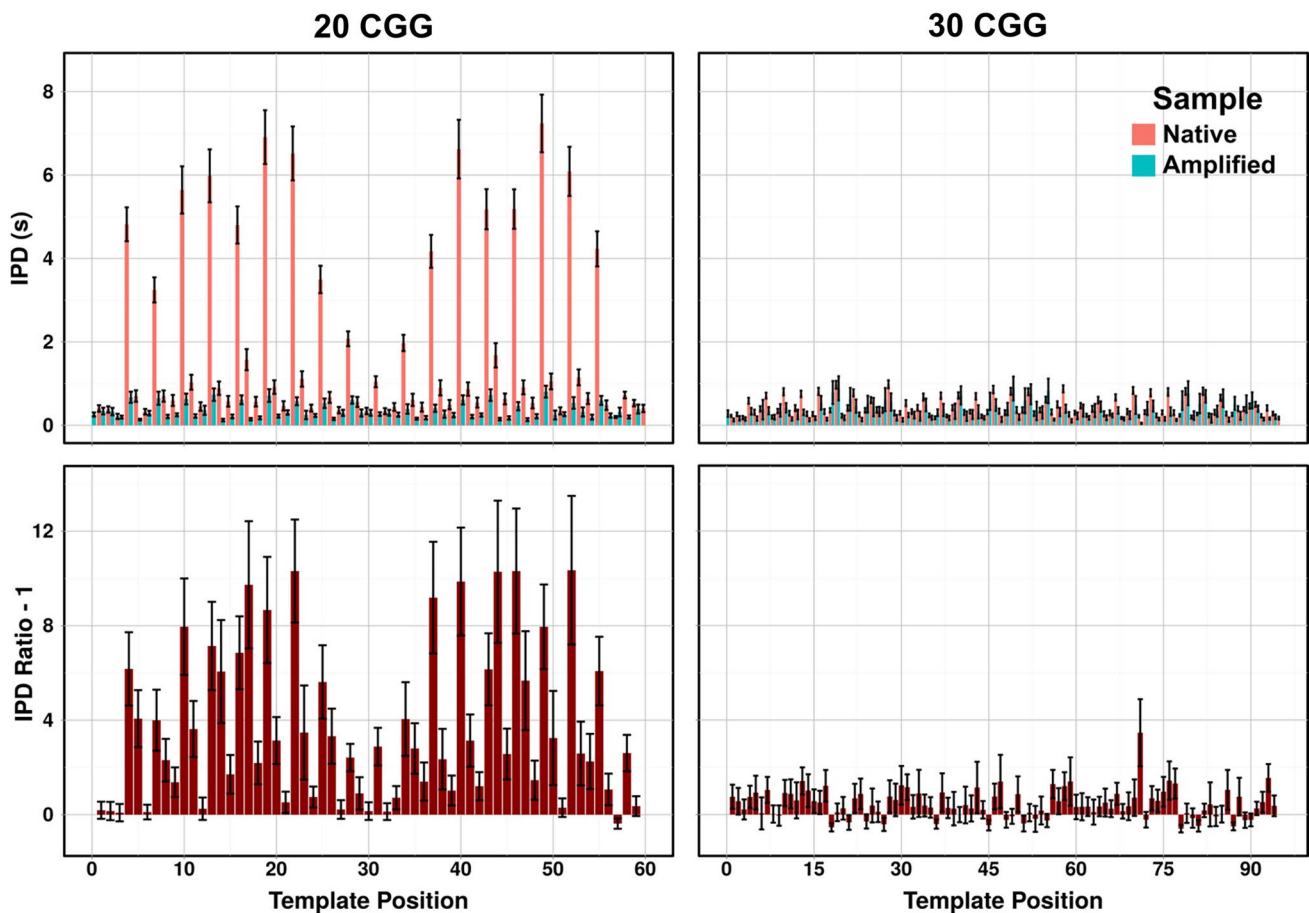


Fig. 4 Direct observation of X-inactivation in which the 20 CGG allele is highly methylated whereas the 30 CGG allele shows no evidence of methylation. The *top row* contains the raw mean IPD values of both the native (*red bars*) and amplified (*blue bars*) samples with standard error bars. The *bottom row* is the ratio between the native

and amplified samples at each template position minus one to highlight kinetic differences from an unmodified position. The standard error of the mean IPD values were propagated in the calculation of the ratio and shown as *error bars* in the plot

prepared and used either unmodified or in vitro-methylated (using Sss1 methyltransferase), as negative and positive controls, respectively. Figure 5 shows the IPD ratio plots for the positive control (top row), native DNA (middle row) and negative control (bottom row) for both forward and reverse strands and confirms that the forward strand of the 20 CGG allele is methylated and that the forward strand of the 30 CGG allele is not. The degree to which the forward strand of the 20 CGG repeat is methylated can be inferred from a direct comparison of the positive control to the native DNA (Figure S2). The values of the IPD ratios are largely within the standard error of the measurement in the CGG repeat sequence suggesting that the native DNA is very close to 100 % methylated. The same cannot be said regarding the reverse strand because the magnitude of the kinetic signal due to methylation is much smaller relative to the forward strand (see Fig. 5—positive control for reverse strand). Therefore, from these data alone, it is not possible to determine if the reverse strand of either the native 20 CGG repeat

or the native 30 CGG repeat is methylated even though it is likely that the reverse strand of the 20 CGG repeat is, in fact, methylated. However, because the standard error of the mean IPD values at each position is expected to decrease as a function of coverage, perhaps it would be possible to make a high certainty call with greater sequencing coverage than was obtained in this study. The difference in signal between the forward and reverse strands is likely due to the pronounced differences in kinetic response as a function of sequence context (Flusberg et al. 2010).

It should be noted that despite a dense cluster of high IPD ratios confined within the repeat region (Figs. 4, 5), these observations are consistent with pervasive methylation across this portion of the forward strand of the native 20 CGG allele. The appearance of the cluster is likely due to a synergistic interaction between adjacent methylcytosines when they reside within the 10–12 base template-footprint of the sequencing polymerase, given the high density of CpGs within the *FMRI* repeat region. Figure 6

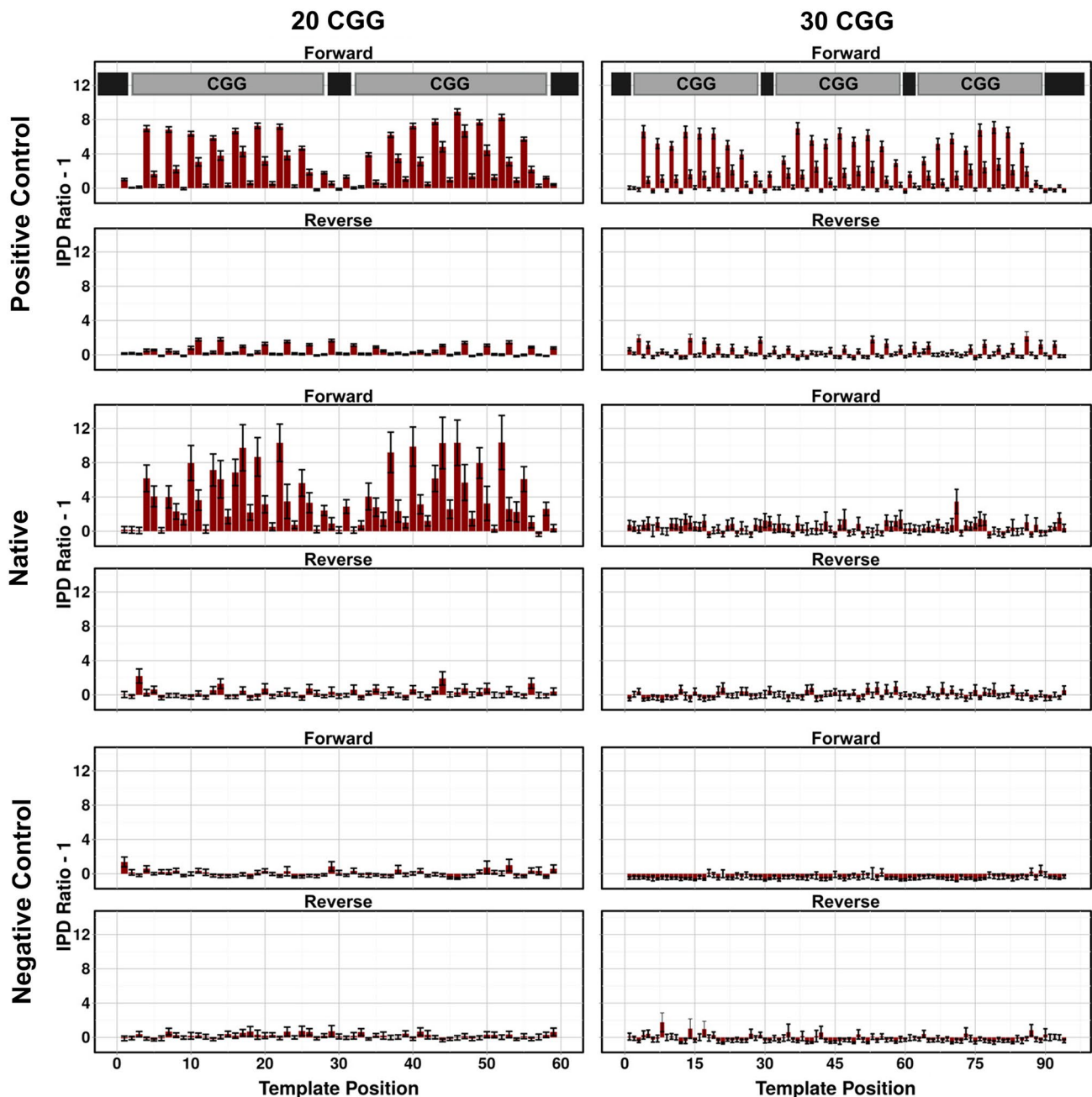


Fig. 5 Comparison of native samples (*middle row*) to negative (*bottom row*) and positive (*top row*) controls for 20 and 30 CGG repeat lengths. The forward strand of the 20 CGG allele mirrors the positive

shows the IPD ratio profile of the entire 1.1 kbp region for both alleles of the on-target fragment.

Premutation repeat alleles

Several gDNA preparations from male cell lines with normal and expanded CGG repeats were enriched in order to evaluate this technique on samples that more closely reflect premutation alleles. Some changes were made for these experiments

control, while the 30 CGG allele does not. The standard error of the mean IPD values were propagated in the calculation of the ratio and shown as *error bars* in the plot

due to the discovery of a slow, but significant, endonuclease activity present in Exo III and Exo VII. Instead, T7 Exo, Exo I, and RecJf were used, as this combination exhibited a greatly reduced rate of endonucleolytic cleavage (data not shown). At present, it is not clear if the endonuclease activity seen with Exo III/Exo VII is due to a contaminant or an inherent property of these enzymes. With this improvement, the sequence-specific “capture-hook” step was not included for these samples, as the required level of enrichment was

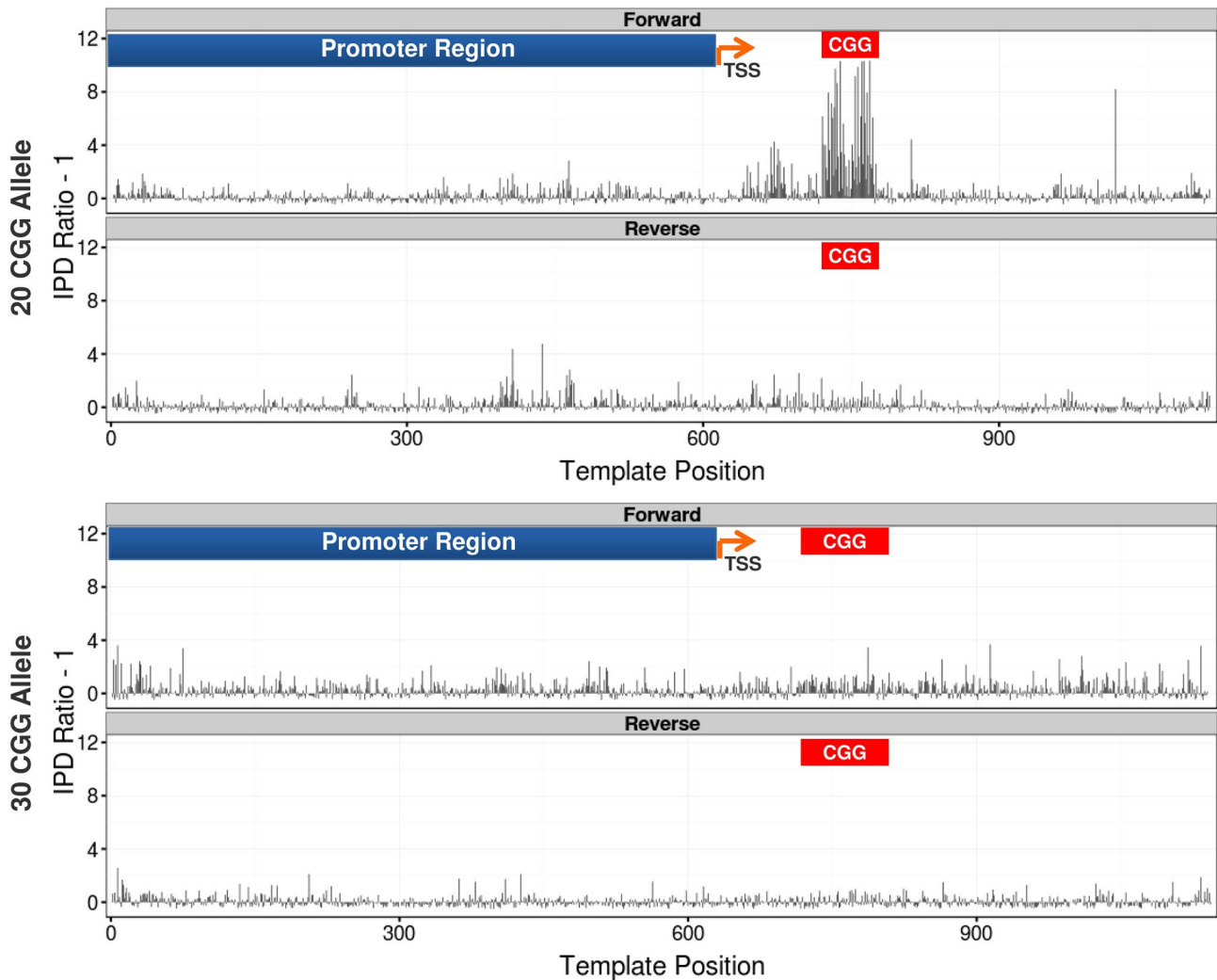


Fig. 6 View of the IPD ratio parameter over the 1.1 kb *FMR1* gene region that was enriched showing that areas outside the CGG repeat section (below the red ‘CGG’ boxes), of the 20 CGG allele, also

appear to be modified on both strands. The promoter region is indicated by the blue boxes and the TSS arrows delineate the transcription start site locations

Table 2 Enrichment of normal and premutation alleles from male gDNA

Sample	Total post-filtered reads	Mapped reads to human	Mapped reads to <i>FMR1</i>	<i>FMR1</i> specificity	Enrichment ^a	Repeat length
Library 1019-09/26 ^b	32,909	26,916	43	0.00160	20,253	28.5 ± 0.7
Library TS-107-12/71	32,101	25,603	53	0.00207	26,203	69.3 ± 2.4
Library 1066-09-RW/97	26,723	22,199	58	0.00261	33,038	94.9 ± 7.8
Library TS-109-12/128	20,853	16,037	22	0.00137	17,342	118.5 ± 7.6

^a For human male diploid, the fraction of Bsm AI-fragments with an *FMR1* locus is $\sim 7.9 \times 10^{-8}$

^b Line designation/CGG-repeat size

reached without additional purification. The enrichment factor was between roughly 17,000 and 33,000 and sufficient to estimate the length of the expansion (Table 2). The estimated CGG-repeat lengths were found to be in agreement with the

known lengths for these gDNA samples determined by PCR-electrophoresis. As expected, IPD ratio analyses were consistent with the interpretation that the CGG repeats in these 4 samples are unmodified (Fig. 7).

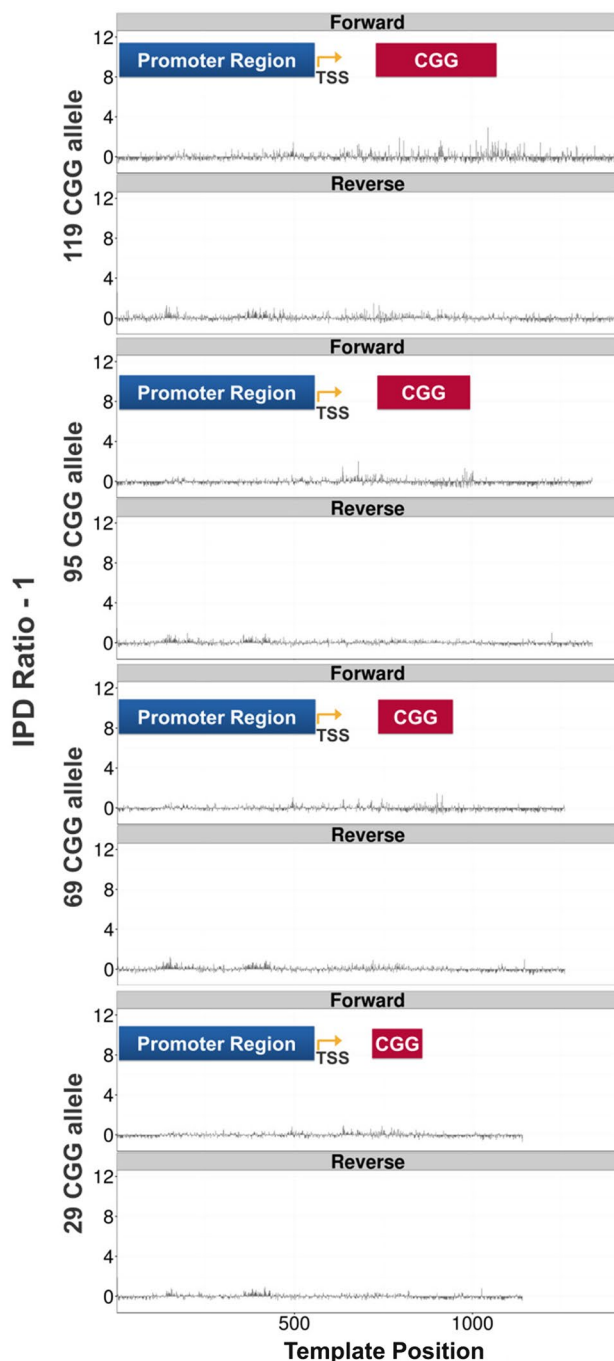


Fig. 7 IPD ratio analyses for the CGG repeat regions in 4 male samples. Comparison of IPD ratio over the same enriched *FMR1* gene region from 4 male gDNA samples indicates that the CGG repeats in these 4 samples are unmodified

Discussion

Expansions of tandem-repeat DNA are associated with a broad range of clinical disorders, heavily weighted to neurodevelopmental and neurodegenerative syndromes [e.g.,

fragile X syndrome (CGG) (Verkerk et al. 1991); Huntington disease and the spinocerebellar ataxias 1, 2, 3, 6, 7, 12 (CAG) (Walker 2007); myotonic dystrophy (CTG) (Udd and Krahe 2012); Friedreich's ataxia (GAA) (Marmolino 2011)]. However, most of the current sequencing technologies are not capable of sequencing long runs of tandem repeats, due to the absence of unique-sequence “landmarks” that would otherwise permit sequence tiling.

Given the high prevalence of *FMR1* expanded alleles in the general population (approximately 0.5 % carrier frequency in the United States), and the availability of promising new targeted treatments, there is an urgent need for rapid and cost-effective detection of CGG-expanded alleles in early childhood. As SMRT sequencing provides the high throughput capability needed to sequence hundreds of clinical samples in tandem, this single-locus sequencing technology could lead to more accurate, less expensive, and higher-throughput means for screening expanded alleles.

The single-locus capture method presented here is, in theory, applicable to a broad range of repeat-expansion disorders, as well as to the study of many other forms of tandem-repeat DNA where the distinguishing feature is the lack of the complex/unique sequence milestones. Moreover, our single-locus capture methodology should permit enrichment of any locus in the genome; thus, it is broadly applicable to sequencing of any locus, especially those that are refractory to accurate PCR or sequencing due to either size or GC content.

A rapidly increasing number of epigenetic modifications have been found to play important roles during development and disease involved pathogenesis, including mCG, mCH (non CpG), hmC, fmC, caC and 8-oxo-G (Taddei et al. 1997; Maga et al. 2007; Fu and He 2012; Lister et al. 2013; Shen et al. 2013; Shen and Zhang 2013; Song and He 2013; Song et al. 2013). Epigenetic studies of these modifications still mostly rely on chemical treatment of genomic DNA followed by PCR-based amplification, which often yields biased results, due to preferential utilization of primers targeting bisulfite-converted (or unconverted) DNA sequence and/or to selective reamplification of specific sequences formed during the first few rounds of PCR. Thus, a second specific advantage of our approach is that it enables one to study directly the patterns of modifications of genomic DNA, without having to chemically modify and amplify the DNA; an approach that has broad applicability not only to microsatellite sequencing, but also for direct characterization of genome-level base modifications through the kinetic sequencing capability of SMRT sequencing. Finally, as an intrinsically single-molecule approach SMRT sequencing should provide the means for examining mosaicism of allele size and modification, which are not readily accessible by other methods.

Acknowledgments The authors wish to thank the entire staff at Pacific Biosciences, in particular Leewin Chern for PCR experiments, Karl Voss for helpful discussions, and the families that have contributed to our fragile X research.

Compliance with ethical standards

Conflict of interest Thang T. Pham, John S. Eid, Regina Lam, Stephen W. Turner and Jeremiah W. Hanes were employed at Pacific Biosciences (manufacturer of the PacBio RS II DNA sequencing instrument used in this study) throughout the course of this study. Paul J. Hagerman is a nonremunerative collaborator with Pacific Biosciences and with Roche Diagnostics; he also holds a patent for PCR-based methods for sizing CGG repeats. All other authors declare no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Funding This work was supported by the National Institutes of Health (R01HD040661 to P.J.H.).

References

- Arocena DG, Iwahashi CK, Won N, Beilina A, Ludwig AL, Tassone F, Schwartz PH, Hagerman PJ (2005) Induction of inclusion formation and disruption of lamin A/C structure by pre-mutation CGG-repeat RNA in human cultured neural cells. *Hum Mol Genet* 14(23):3661–3671
- Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10(6):563–569
- Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 106(45):19096–19101
- Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res* 33(8):e71
- Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev* 25(10):1010–1022
- Evans-Galea MV, Hannan AJ, Carroddus N, Delatycki MB, Saffery R (2013) Epigenetic modifications in trinucleotide repeat diseases. *Trends Mol Med* 19(11):655–663
- Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* 7(6):461–465
- Fu Y, He C (2012) Nucleic acid modifications with epigenetic significance. *Curr Opin Chem Biol* 16(5–6):516–524
- Gallagher A, Hallahan B (2012) Fragile X-associated disorders: a clinical overview. *J Neurol* 259(3):401–413
- Hagerman P (2013) Fragile X-associated tremor/ataxia syndrome (FXTAS): pathology and mechanisms. *Acta Neuropathol* 126(1):1–19
- Hagerman R, Hagerman P (2013) Advances in clinical and molecular understanding of the FMR1 pre-mutation and fragile X-associated tremor/ataxia syndrome. *Lancet Neurol* 12(8):786–798
- Hagerman R, Hoem G, Hagerman P (2010) Fragile X and autism: intertwined at the molecular level leading to targeted treatments. *Mol Autism* 1(1):12
- Hutchison CA 3rd, Smith HO, Pfannkoch C, Venter JC (2005) Cell-free cloning using phi29 DNA polymerase. *Proc Natl Acad Sci U S A* 102(48):17332–17336
- Kieleczawa J (2006) Fundamentals of sequencing of difficult templates—an overview. *J Biomol Tech* 17(3):207–217
- Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Tonti-Filippini J, Heyn H, Hu S, Wu JC, Rao A, Esteller M, He C, Haghghi FG, Sejnowski TJ, Behrens MM, Ecker JR (2013) Global epigenomic reconfiguration during mammalian brain development. *Science* 341(6146):1237905
- 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(1):121–128
- Maga G, Villani G, Crespan E, Wimmer U, Ferrari E, Bertocci B, Hubscher U (2007) 8-oxo-guanine bypass by human DNA polymerases in the presence of auxiliary proteins. *Nature* 447(7144):606–608
- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7(2):111–118
- Marmolino D (2011) Friedreich’s ataxia: past, present and future. *Brain Res Rev* 67(1–2):311–330
- Mirkin SM (2007) Expandable DNA repeats and human disease. *Nature* 447(7147):932–940
- Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, Korlach J, Roberts RJ (2012) The methylomes of six bacteria. *Nucleic Acids Res* 40(22):11450–11462
- Mutter GL, Boynton KA (1995) PCR bias in amplification of androgen receptor alleles, a trinucleotide repeat marker used in clonality studies. *Nucleic Acids Res* 23(8):1411–1418
- Nelson DL, Orr HT, Warren ST (2013) The unstable repeats—three evolving faces of neurological disease. *Neuron* 77(5):825–843
- Primerano B, Tassone F, Hagerman RJ, Hagerman P, Amaldi F, Bagni C (2002) Reduced FMR1 mRNA translation efficiency in fragile X patients with premutations. *RNA* 8(12):1482–1488
- Shen L, Zhang Y (2013) 5-Hydroxymethylcytosine: generation, fate, and genomic distribution. *Curr Opin Cell Biol* 25(3):289–296
- Shen L, Wu H, Diep D, Yamaguchi S, D’Alessio AC, Fung HL, Zhang K, Zhang Y (2013) Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* 153(3):692–706
- So A, Pel J, Rajan S, Marziali A (2010) Efficient genomic DNA extraction from low target concentration bacterial cultures using SCODA DNA extraction technology. *Cold Spring Harb Protoc* 2010(10):pdb prot5506
- Song CX, He C (2013) Potential functional roles of DNA demethylation intermediates. *Trends Biochem Sci* 38(10):480–484
- Song CX, Szulwach KE, Dai Q, Fu Y, Mao SQ, Lin L, Street C, Li Y, Poidevin M, Wu H, Gao J, Liu P, Li L, Xu GL, Jin P, He C (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* 153(3):678–691
- Taddei F, Hayakawa H, Bouton M, Cirinesi A, Matic I, Sekiguchi M, Radman M (1997) Counteraction by MutT protein of transcriptional errors caused by oxidative damage. *Science* 278(5335):128–130
- Tassone F, Hagerman RJ, Taylor AK, Gane LW, Godfrey TE, Hagerman PJ (2000) Elevated levels of FMR1 mRNA in carrier males: a new mechanism of involvement in the fragile-X syndrome. *Am J Hum Genet* 66(1):6–15
- Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ, Program NCS, Margulies EH, Green ED, Collins FS, Mullikin JC, Biesecker LG (2010) Systematic

- comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* 20(10):1420–1431
- Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* 38(15):e159
- Udd B, Krahe R (2012) The myotonic dystrophies: molecular, clinical, and therapeutic challenges. *Lancet Neurol* 11(10):891–905
- Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, Reiner O, Richards S, Victoria MF, Zhang FP 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(5):905–914
- Walker FO (2007) Huntington's disease. *Lancet* 369(9557):218–228