Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions

J. Ekholm[1)], Y.C. Tsai[1)], T. Hon[1)], B. Bowman[1)], J. Ziegle[1)], B. Schule[2)], T. Ashizawa[3)], K.N. McFarland[4)], T.A. Clark[1]
[1)] PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025, [2)] Parkinson's Institute and Clinical Center, Sunnyvale, CA, USA, [3)] Houston Methodist Research Institute, Houston, TX, USA, [4)] Center for Translational Research in Neurodegenerative Disease and The McKnight Brain Institute, University of Florida, FL, USA

Abstract

Targeted sequencing has proven to be economical for obtaining sequence information for defined regions of the genome. However, most target enrichment methods are reliant upon some form of amplification which can negatively impact downstream analysis. For example, amplification removes epigenetic marks present in native DNA, including nucleotide methylation, which are hypothesized to contribute to disease mechanisms in some disorders. In addition, some genomic regions known to be causative of many genetic disorders have extreme GC content and/or repetitive sequences that tend to be recalcitrant to faithful amplification.

We have developed a novel, amplification-free enrichment technique that employs the CRISPR/Cas9 system to target individual genes. This method, in conjunction with the long reads, high consensus accuracy, and uniform coverage of SMRT Sequencing, allows accurate sequence analysis of complex genomic regions that cannot be investigated with other technologies. Using this strategy, we have successfully targeted a number of repeat expansion disorder loci (*HTT*[1,2], *FMR1*[1,2], *ATXN10*[1,3], *C9orf72*[1,4]).

With this data, we demonstrate the ability to isolate thousands of individual on-target molecules and, using the Sequel System, accurately sequence through long repeats regardless of the extreme GC-content. The method is compatible with multiplexing of multiple target loci and multiple samples in a single reaction. Furthermore, because there is no amplification step, this technique also preserves native DNA molecules for sequencing, allowing for the direct detection and characterization of epigenetic signatures. To this end, we demonstrate the detection of 5-mC in the CGG repeat of the *FMR1* gene that is responsible for Fragile X syndrome.

Method Overview

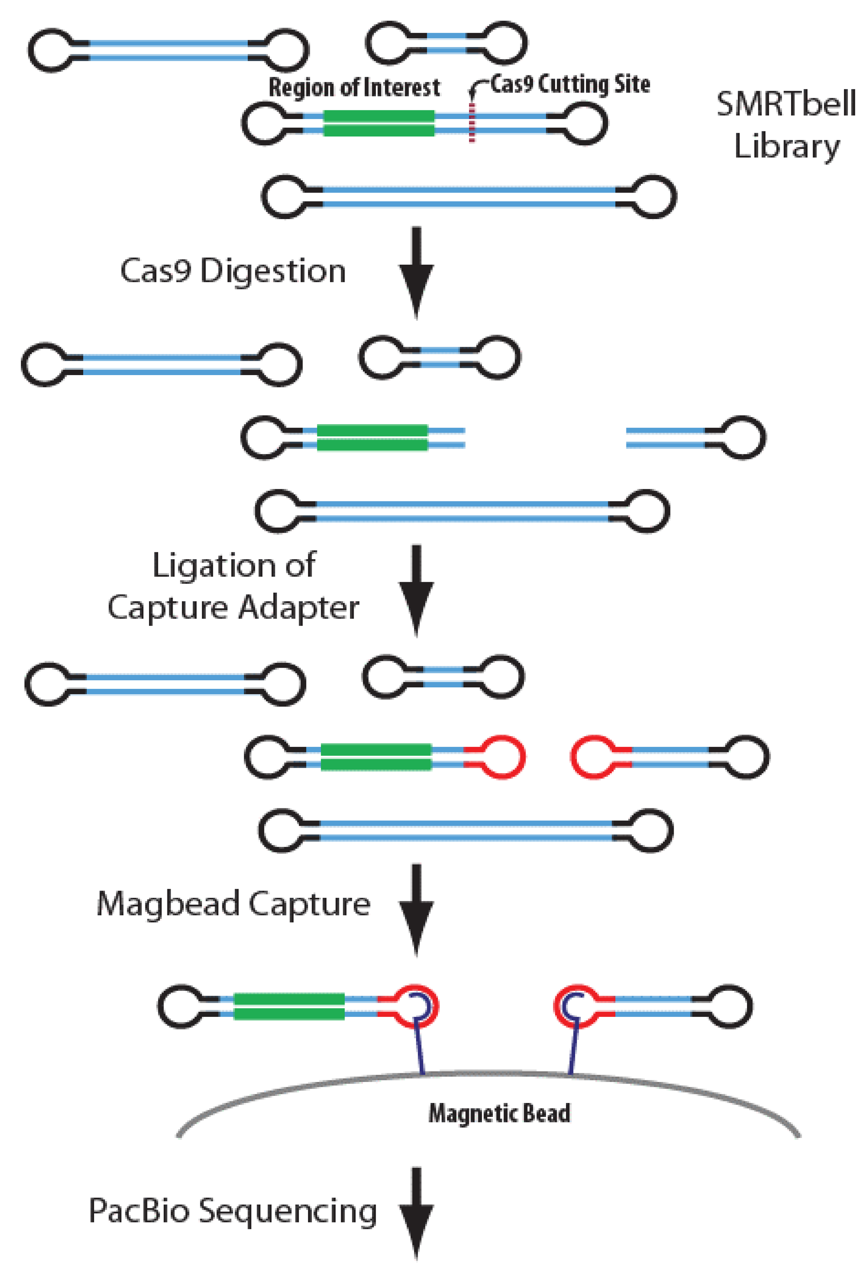


Figure 1. No-Amp workflow. A standard SMRTbell template library is created and a crRNA (guide RNA) is designed adjacent to the region of interest. Digestion with Cas9 breaks open the SMRTbell molecules to enable ligation with a capture adapter. SMRTbell molecules that contain the capture adapter are enriched on magnetic beads and prepared for SMRT Sequencing on a Sequel System.

Targeted Sequencing *FMR1* and *HTT* Repeat Expansions

Target Gene	Associated Disease(s)	Chr	crRNA Coordinates	Strand	Target Size	Repeat
HTT	Huntington's Disease	Chr 4	3075105-3075086	-	2700bp	CAG
FMR1	Fragile X and Fragile X-associated Tremor/Ataxia Syndrome (FXTAS)	Chr X	147911587-147911606	+	2800bp	CGG

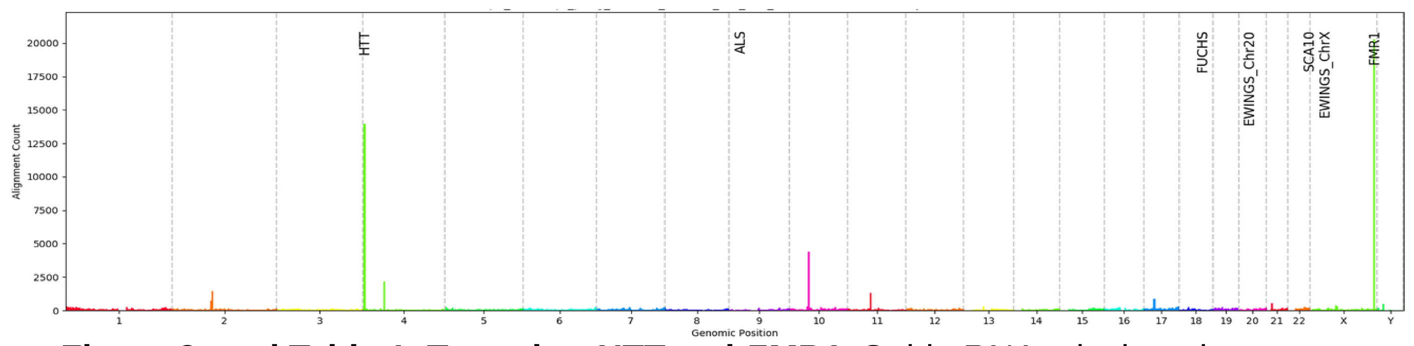
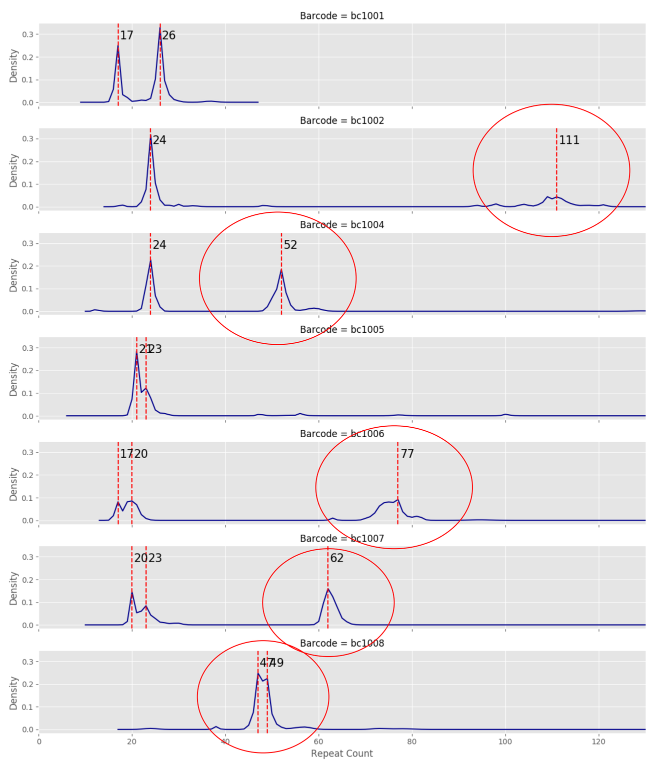


Figure 2. and Table 1. Targeting *HTT* and *FMR1*. Guide RNAs designed to capture two repeat expansion loci were multiplexed in a single experiment. Molecule coverage across the entire genome is shown above. Off-target signal can be explained by homology of the guide RNA sequence to other regions in the human genome.

The higher throughput of the Sequel System allows for higher multiplexing capacity for the No-Amp method. Here, an enrichment factor achieved of *>49,000* for *HTT* and *>37,000* for *FMR1* were achieved when multiplexing 7 samples and 2 loci in one Sequel SMRT Cell.

Coriell sample	Reported HTT rpt	HTT mot*	FMR1 mot*
NA20246	15,24	372	149
NA20253	22, 108	310	142
NA13505	22, 50	601	333
NA14044	19, 205	158	105
NA13509	15, 70	289	281
NA03620	16, 60	84	58
NA13511	45, 47	129	56
Total		1943	1124



* Molecules on target

Table 2. **Multiplexing results of 7 samples and two loci on one SMRT Cell.** The seven *HTT* samples were purchased from Coriell and for demonstration purposes the DNA samples were enriched for both the *HTT* expansion as well as *FMR1*. Column 2 represents for reported repeat lengths based on PCR. Columns 3 and 4 shows the number of enriched molecules on target for each locus.

Figure 3. *HTT* repeat lengths per allele. The figure shows the repeat expansion lengths captured for each of the Coriell samples in the same order as represented in the table. The mutated expanded allele is indicated with a red circle.

We have developed software solutions for No-Amp that enables easy visibility of expansion classifications and interruption sequence detection.

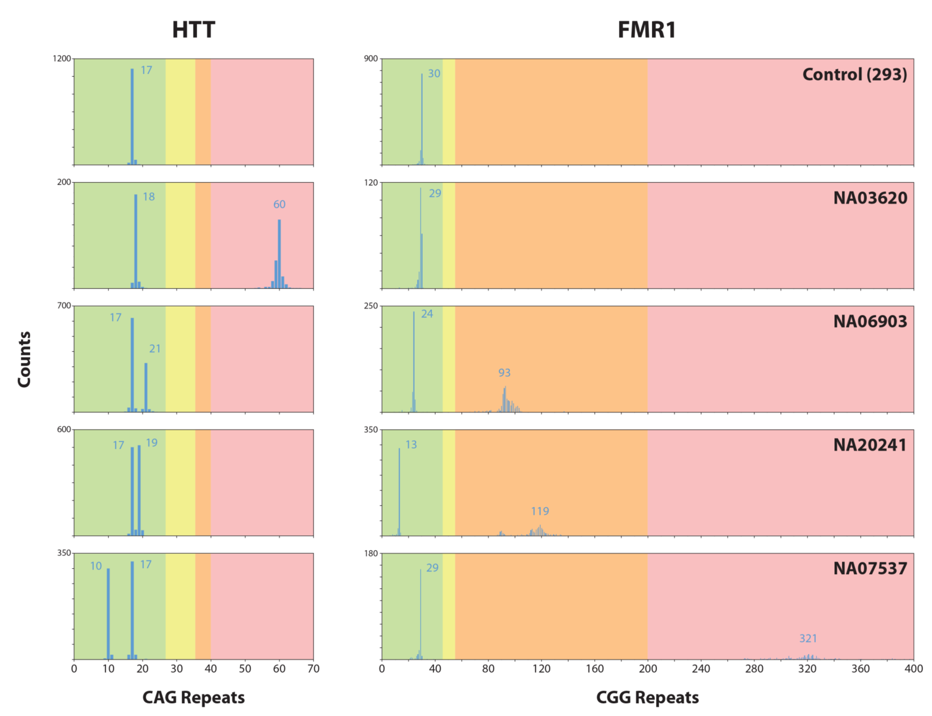


Figure 4. Expansion classes. Repeat counts are plotted for the *HTT* (left) and *FMR1* (right) loci across all samples with count numbers on the y-axis and CAG (*HTT*) or CGG (*FMR1*) repeat numbers on the x-axis. Mode values for each allele are labeled. Shaded background in each plot represents risk ranges for developing disease.

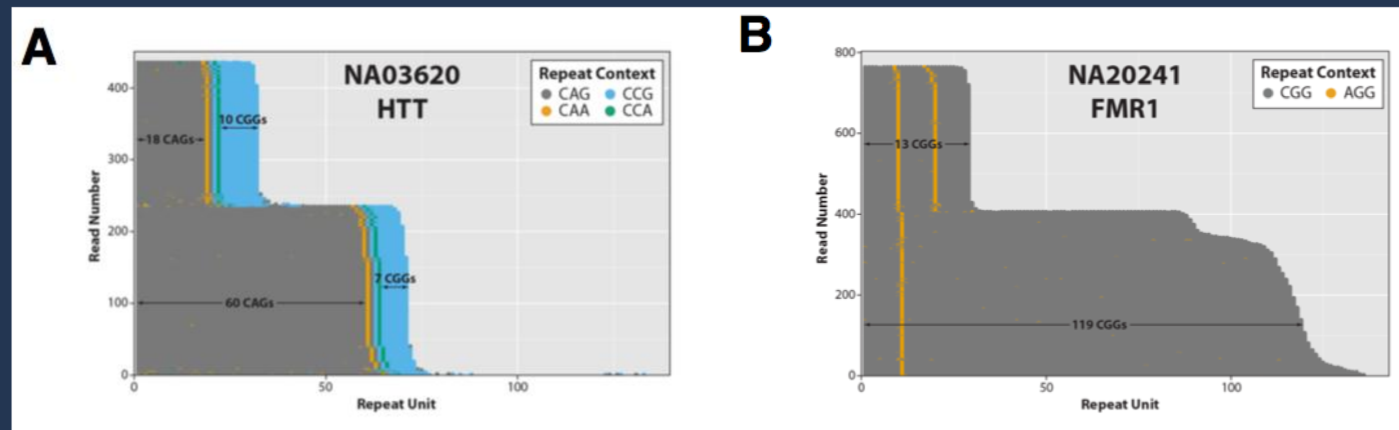


Figure 5. Characterization of repeat structure. Individual Circular Consensus Sequencing (CCS) reads are trimmed of flanking sequence to include only the relevant repeat region. Trimmed repeat sequences are sorted from shortest to longest. Each individual molecule is represented by a series of colored dots on a horizontal line with each dot representing a single repeat unit, color coded based on the repeat content. **(A)** *HTT* region in NA03620: Two alleles are visible with varying numbers of CAG and CCG repeats. **(B)** *FMR1* region in NA20241: Two alleles with varying numbers of CGG repeats and AGG interruptions.

Methylation Detection

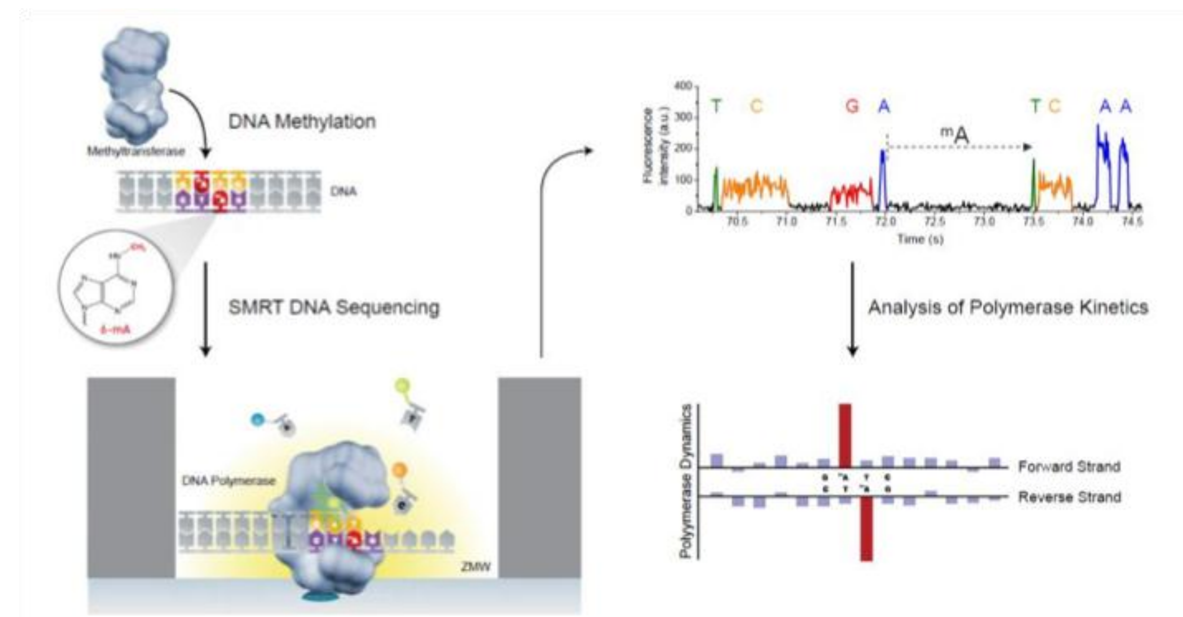


Figure 6. Direct Detection of DNA Modifications During SMRT Sequencing. SMRT Sequencing uses kinetic information from each nucleotide to distinguish between modified and native bases.

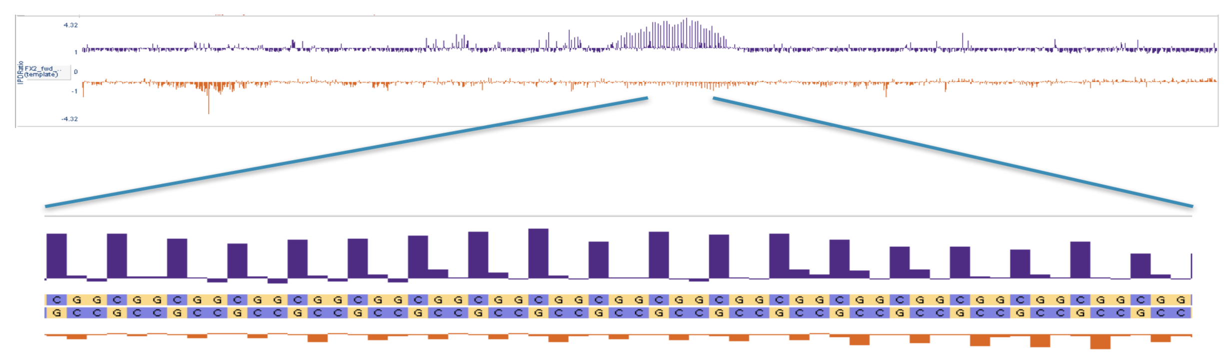


Figure 7. *FMR1* example. Kinetic information from a targeted region of the *FMR1* gene shows heavy methylation (5mC) of the CGG repeat.

Conclusion

Enrich for targeted genomic regions without amplification

- Avoid PCR bias
- Preserve epigenetic modification signals
- Target any genomic region regardless of sequence content

Achieve base-level resolution required to understand the underlying biology of repeat expansion disorder

- Accurately sequence through long repetitive and low-complexity regions
- Count repeats and identify interruption sequences
- Detect mosaicism with single-molecule sequencing

References

1. Tsai Y.C. et al. (2017) Amplification-free, CRISPR-Cas9 Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Disease Causative Genomic Regions *bioRxiv* doi:http://dx.doi.org/10.1101/203919
2. Hoijer I. et al. (2018) Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing *Hum Mutat.* doi: 10.1002/humu.23580
3. Schüle B. et al. (2017) Parkinson's disease associated with pure ATXN10 repeat expansion. *NPJ Parkinsons* Dis;3:27
4. Ebbert MTW. et al. (2018) Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol Neurodegener*. 13(1):46