Amplification-free Targeted Enrichment and SMRT Sequencing of Repeat-Expansion Genomic Regions

Tyson A. Clark, Yu-Chih Tsai, Ting Hon, Brett Bowman, Janet Ziegle, Jenny Ekholm, and Jonas Korlach
Pacific Biosciences, 1305 O'Brien Dr, Menlo Park, CA 94025

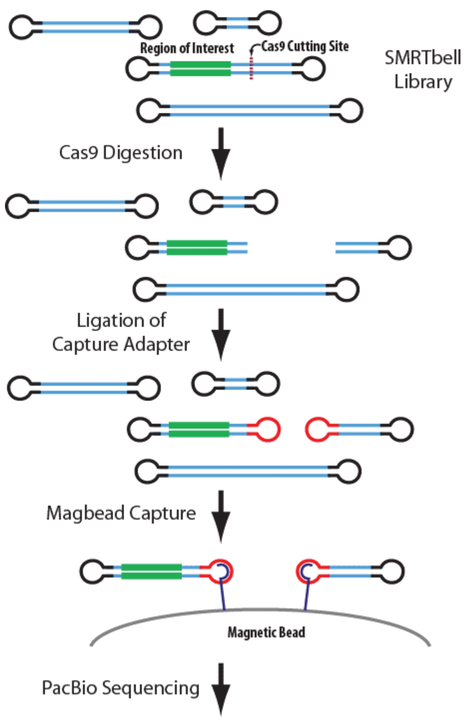
Abstract

Targeted sequencing has proven to be an economical means of obtaining sequence information for one or more defined regions of a larger genome. However, most target enrichment methods are reliant upon some form of amplification. Amplification removes the epigenetic marks present in native DNA, and some genomic regions, such as those with extreme GC content and repetitive sequences, are recalcitrant to faithful amplification. Yet, a large number of genetic disorders are caused by expansions of repeat sequences. Furthermore, for some disorders, methylation status has been shown to be a key factor in the mechanism of disease.

We have developed a novel, amplification-free enrichment technique that employs the CRISPR/Cas9 system for specific targeting of individual human genes. This method, in conjunction with SMRT Sequencing's long reads, high consensus accuracy, and uniform coverage, allows the sequencing of complex genomic regions that cannot be investigated with other technologies. Using human genomic DNA samples and this strategy, we have successfully targeted the loci of a number of repeat expansion disorders (*HTT*, *FMR1*, *ATXN10*, and *C9orf72*).

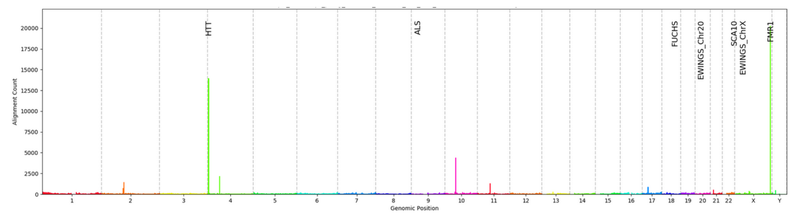
With this data, we demonstrate the ability to isolate hundreds of individual on-target molecules and accurately sequence through long repeat stretches, regardless of the extreme GC-content, followed by accurate sequencing on the PacBio RS II or Sequel instruments. Analysis algorithms that use the Arrow model for SMRT Sequencing have been developed to quantify the number of repeats and generate high-quality consensus sequences for each sample. The method is compatible with multiplexing of multiple targets and multiple samples in a single reaction. Furthermore, this technique also preserves native DNA molecules for sequencing, allowing for the possibility of direct detection and characterization of epigenetic signatures. We demonstrate detection of 5-mC in the CGG repeat of the *FMR1* gene that is responsible for Fragile X syndrome.

Method Overview



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 PacBio RS II or Sequel System.

Targeted Sequencing of 4 Repeat Expansions



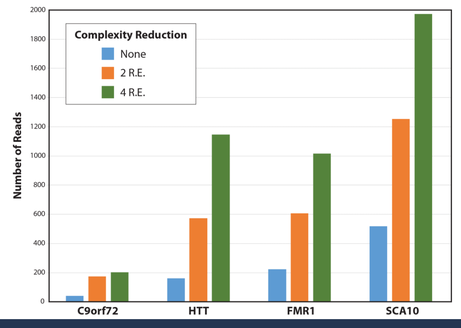
Target Gene	Associated Disease(s)	Chr	crRNA Coordinates	Strand	Target Size	Repeat
HTT	Huntington's Disease	Chr 4	3075105-3075086	-	2700bp	CAG
FMR1	Fragile X and Fragile X-associated Tremor/Ataxia Syndrome (FXTAS)	Chr X	147911587-147911606	+	2800bp	CGG

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.

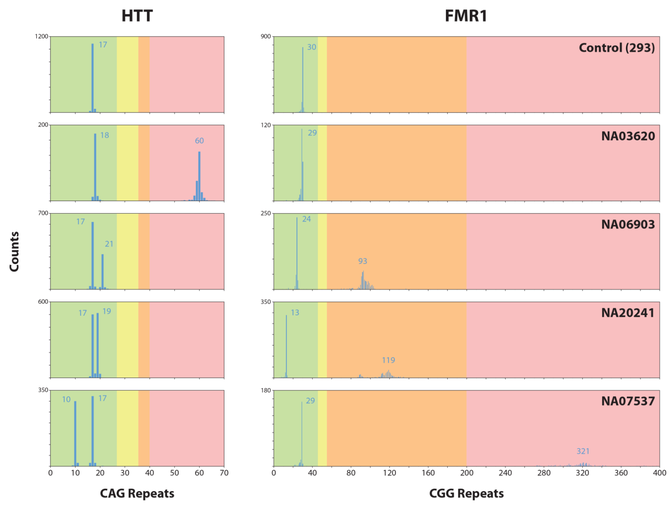
Complexity Reduction Improves On-Target Rate

Complexity Reduction	Input Genomic DNA	Final SMRTbell Yield	% Yield	CCS Reads	On-Target Reads	% Reads On-Target
2 R.E.	20.0 µg	1 µg	5.0%	144,057	6265	2.2%
4 R.E.	80.0 µg	1 µg	1.3%	152,383	17,629	10.1%

Several restriction enzymes that do not cut within the regions of interest were chosen to remove unwanted SMRTbell templates prior to Cas9 digestion and capture. Inclusion of 2 or 4 restriction enzymes predictably reduces the SMRTbell template yield, but dramatically increases the number of on-target reads and the percentage of reads that come from targeted regions.



Repeat Count Histograms



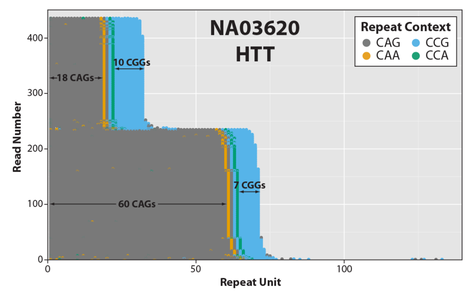
Repeat counts are plotted for the HTT (left) and *FMR1* (right) loci across all 5 Coriell samples with count numbers on the y-axis and CAG (HTT) or CGG (*FMR1*) repeat numbers on the x-axis. Mode values for each allele are labeled. Shaded background in each plot represents risk ranges for developing disease.

Repeat Structure Variation

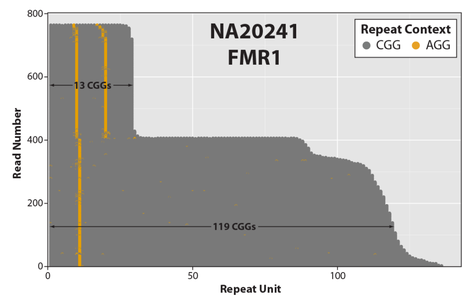
Targeted sequencing of multiple repeat expansion loci was carried out on DNA from two Coriell cell lines from patients with known expansions.

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.

A

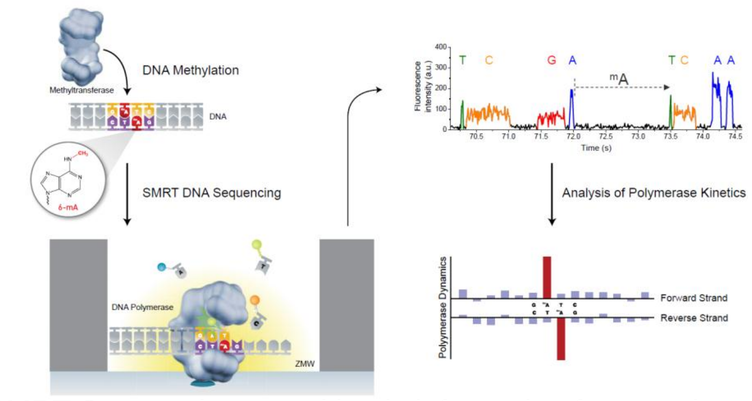


B

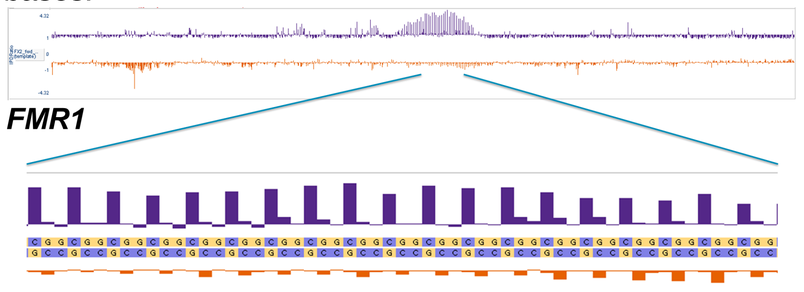


Methylation Detection

Direct Detection of DNA Modifications During SMRT Sequencing



SMRT Sequencing uses kinetic information from each nucleotide to distinguish between modified and native bases.



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

Authors
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025