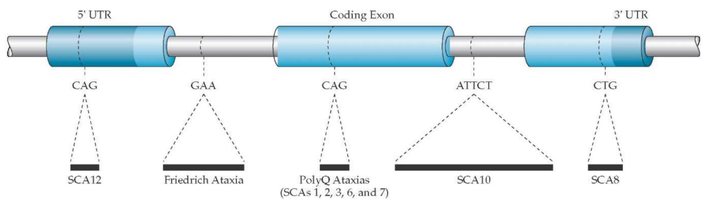
Enrichment of unamplified DNA and long-read SMRT Sequencing to unlock repeat expansion disorders

Jenny Ekholm, Yu-Chih Tsai, David Greenberg, and Tyson A. Clark
PacBio, 1380 Willow Road, Menlo Park, CA 94025

Introduction

There are at least 22 inherited neurological disorders that are known to be caused by repeat expansions. These disorders are signified by repeats that expand anywhere from below 1 kb to 100s of kbs and the number of repeats determines the clinical phenotype. For many of these disorders, while the disease gene has been discovered, the underlying biological mechanisms have not yet been fully understood. This is mainly due to technological limitations that do not allow for the needed base-pair resolution of the long, repetitive genomic regions. However, thus far it is known that interruption sequences in the midst of the repetitive elements affect DNA stability and seems to affect phenotype severity and cause disease anticipation. Epigenetics is also believed to play a role in several of these diseases.



http://what-when-how.com/wp-content/uploads/2012/04/tmp5F23.jpg

We have developed a novel, amplification-free enrichment technique that uses the CRISPR/Cas9 system to target genomic regions of interest. This method, in conjunction with PacBio's long reads and uniform coverage, enables sequencing of complex genomic regions such as repeat expansions. It allows access to the needed base-pair level resolution to accurately count the repeats present for diagnostic purposes. In addition, single molecule real-time (SMRT) sequencing also enables identification of the medically relevant interruption sequences and epigenetic information that seems to play a key role in the underlying disease mechanisms. Moreover, any PCR bias that could effect the results is removed from the experiment.

SMRT Sequencing overview

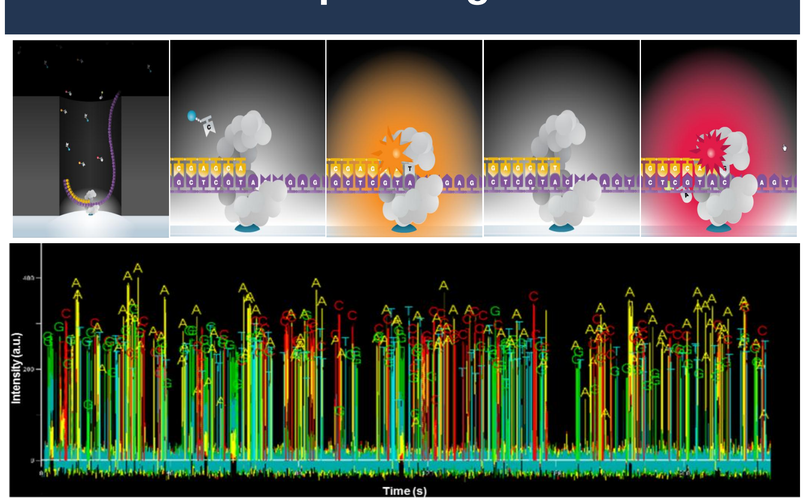
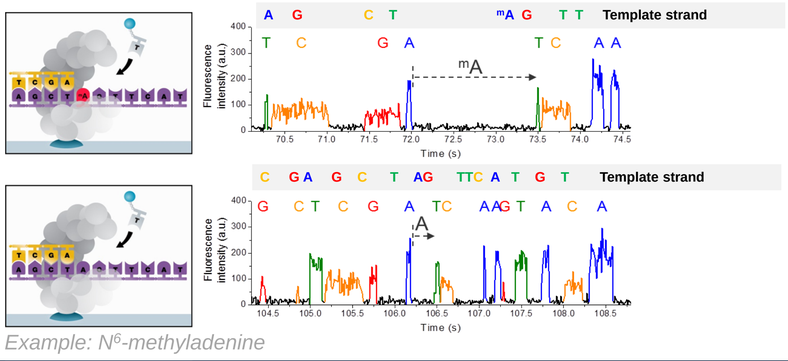


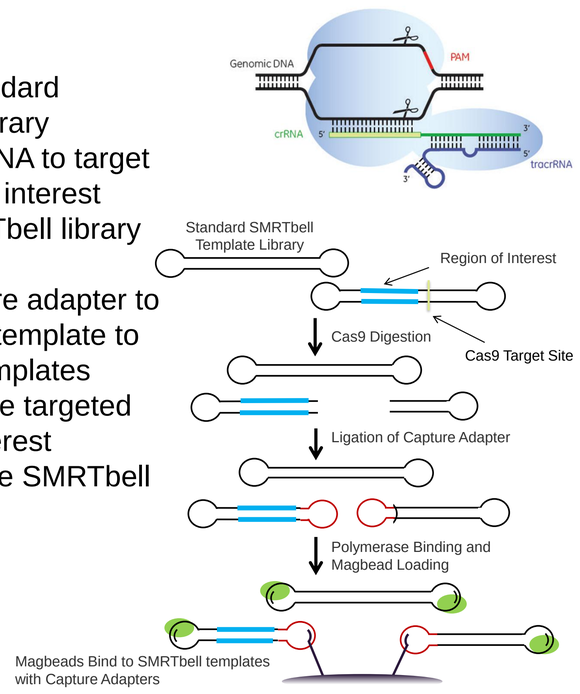
Figure 1. Synthesis by DNA polymerases immobilized in Zero Mode Waveguides (ZMWs) is captured in real-time. Fluorescence emitted from phospholinked nucleotide labels, which are released during incorporation, is converted into base calls with associated quality value using optimized algorithms.
Figure 2. In addition to calling the bases, SMRT Sequencing uses the kinetic information from each nucleotide to distinguish between modified and native bases.



Example: N^6-methyladenine

CRISPR/Cas9 system for targeted amplification-free enrichment

Figure 3.
- Prepare standard SMRTbell library
- Use guide RNA to target the region of interest
- Digest SMRTbell library using Cas9
- Ligate capture adapter to a SMRTbell template to enrich for templates containing the targeted region of interest
- Sequence the SMRTbell templates



CRISPR/Cas9 results for repeat expansion disorders

Guide RNA(s)	*HTT* (HD) Chr4 CAG	C9ORF72 (ALS) Chr9 GGGGCC	*FMR1* (FXS) ChrX CGG	Total Number of Molecules
HTT only	186	0	0	186
C9ORF72 only	0	93	0	93
FMR1 only	0	0	126	126
3x Multiplex	144	63	126	333

Figure 4. Molecules on target (CCS reads) targeting both a single region of multiple regions in the same reaction.
HTT: Huntingtin gene, HD: Huntington's disease, ALS: Amyotrophic Lateral Sclerosis, *FMR1*: Fragile X Mental Retardation 1 gene, FXS: Fragile-X Syndrome

Characterization of *HTT* repeat expansion for Huntington's Disease

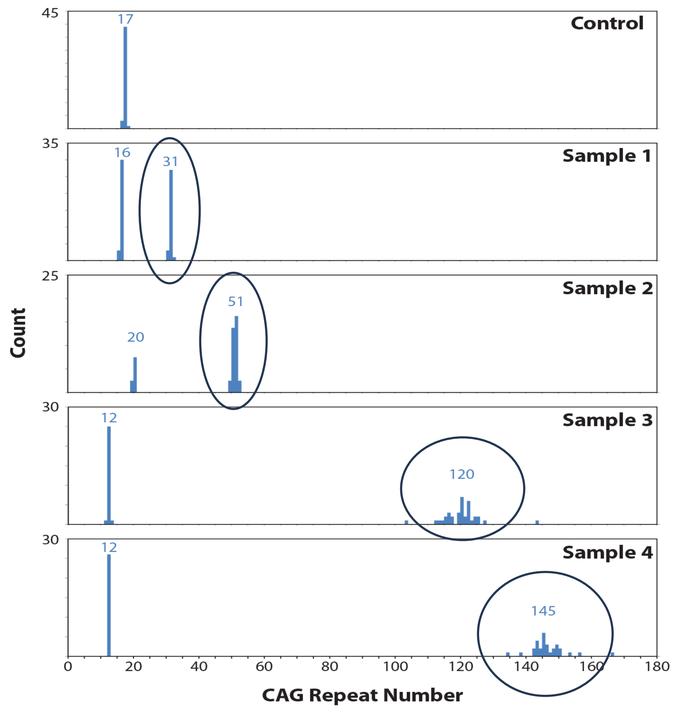


Figure 5.
- CRISPR/Cas9 method captured the *Huntingtin* gene in both control and HD samples
- Roughly equal number of sequenced molecules for normal and mutated alleles are captured despite size range differences
- Identified a widening of the repeat number distribution at the mutated allele, a known biological mechanism (mosaicism)

Characterization of *FMR1* pre-mutation sample

Figure 6a. Based on PCR results, sample was predicted to have ~80 CGG repeats but only 42 were found upon sequencing.

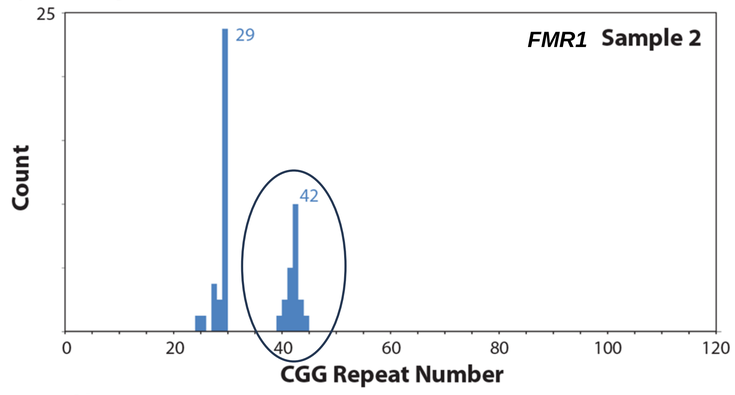


Figure 6b.
Normal allele: 100% concordance
Mutant allele: ~110 bp duplicated region preceding the CGG repeats

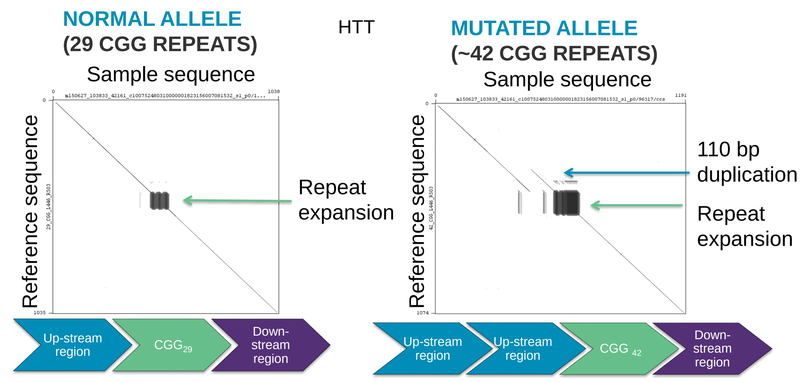


Figure 6c. *FMR1* expansion region captured and sequenced in three different individuals shows variable expansion sizes and differing interruptions

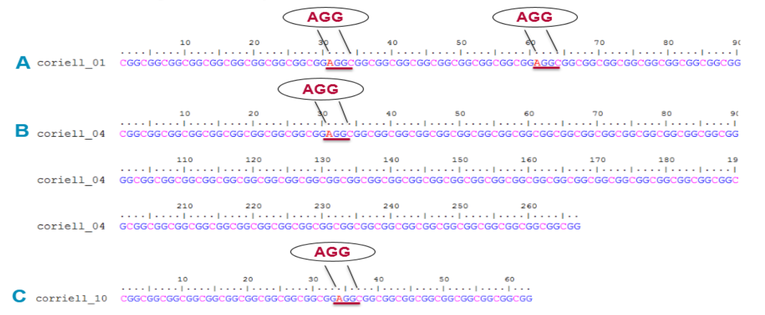
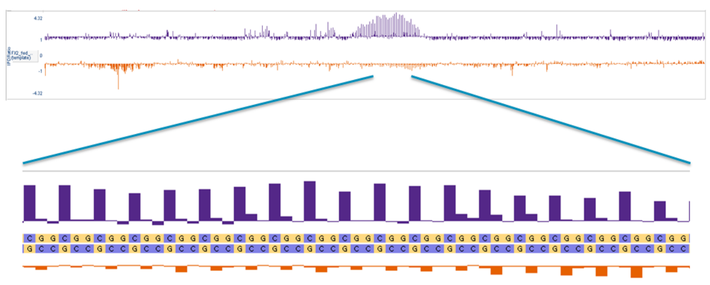


Figure 6d. Direct methylation detection of *FMR1* gene



CGG repeat region appears to be heavily methylated (5mC)

Conclusions

Enrich for targeted genomic regions without amplification
- No PCR bias
- Preserve epigenetic modification signals
- Target any genomic region within reach of Cas9

Achieve base-level resolution required to understand the underlying biology of repeat expansion disorder
- Accurately sequence through long repetitive and low-complexity regions
- Enables repeat counting and interruption sequence identification
- Single molecule sequencing detects mosaicism

Acknowledgements

Samples provided by Vanessa Wheeler (Harvard Medical School) for *HTT* and Paul Hagerman (UC Davis) for *FMR1*

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2016 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. Fragment Analyzer is a trademark of Advanced Analytical Technologies.
All other trademarks are the sole property of their respective owners.