#921T

# Sequencing of expanded CGG repeats in the *FMR1* gene

Erick W. Loomis[1], John Eid[2], Paul Peluso[2], Jackie Yen[2], John Major[2], David Rank[2], Flora Tassone[3], Paul J. Hagerman[1,3]

[1]Univ. of Calif. Davis School of Medicine, Biochemistry and Molecular Medicine, Davis, CA;
[2]Pacific Biosciences of California Inc., Menlo Park, CA; [3]Univ. of Calif. Davis School of Medicine, MIND Institute, Sacramento, CA

## Background:

- The Fragile X Mental Retardation 1 *(FMR1)* gene contains a trinucleotide CGG repeat sequence in the 5' untranslated region (5'UTR)
- Moderate CGG repeat expansions (**55-200** CGG repeats) can lead to Fragile X-Associated Tremor Ataxia Syndrome (FXTAS), a late onset neurodegenerative disorder characterized by intention tremor, gait ataxia, and other features [1]
- Large CGG repeat expansions (**>200** CGG repeats) result in Fragile X Syndrome (FXS) due to transcriptional silencing and loss of *FMR1* protein (FMRP) function [1]

## Problem:

- **Large expanded CGG repeats (>100) produce incoherent signal from methodologies reliant on reading from a population of identical copies (Fig. 1)**
  - **This has prevented acquisition of base-resolution sequence data within the causative element for multiple distinct human disorders**

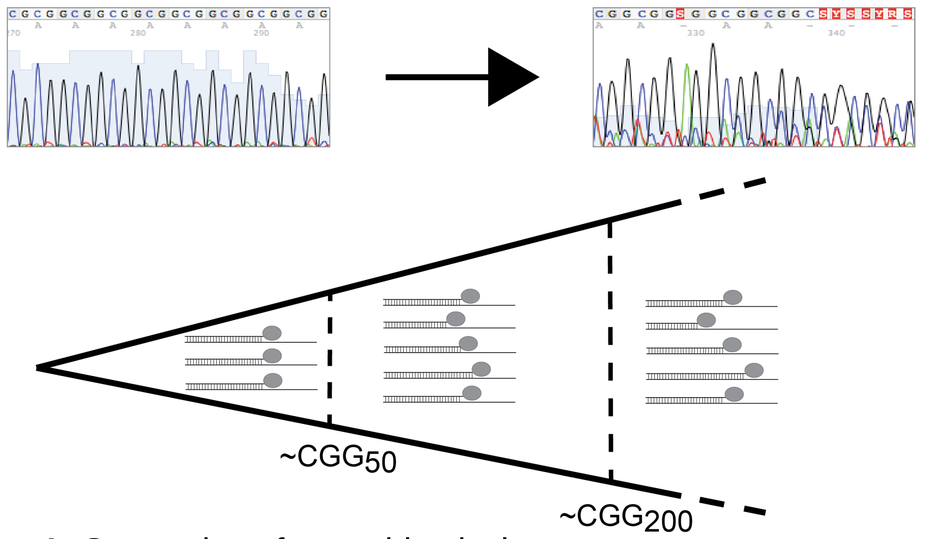


**Figure 1:** Generation of many identical copies becomes increasingly difficult with increasing repeat expansion. This prevents accurate sequencing of repeat expansions in most of the disease relevant range (>100 repeats).

**Hypothesis:** Single molecule, real-time (SMRT) sequencing can produce high quality sequence data through large expanded CGG repeats.

## Approach:

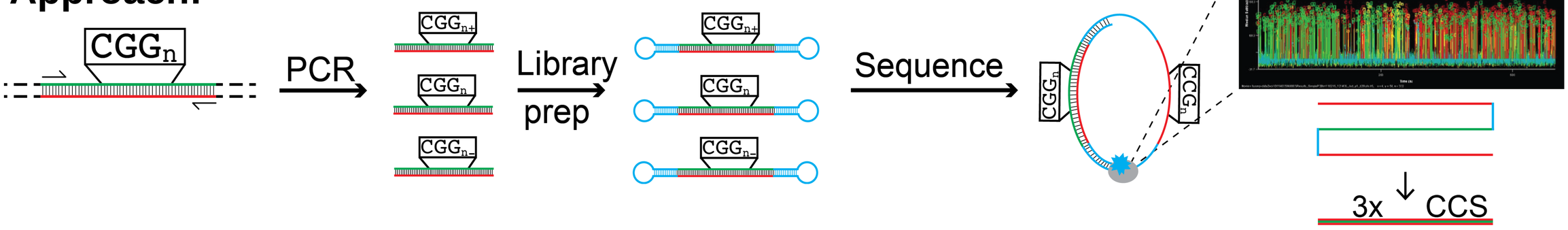


**Figure 2:** Test libraries were prepared using *FMR1* sequence with various CGG repeat sizes PCR amplified from human genomic DNA. Hairpin sequencing adapters were attached to create closed circular sequencing templates. Samples were sequenced using the Pacific Biosciences RS. Long reads with at least 3x coverage of the target sequence were assembled into Circular Consensus Sequence (CCS) reads using the BLASR algorithm.

## Results:

- CCS reads produced across physiologically relevant range of repeat expansion
- Ability to sequence large repeat alleles appears limited only by readlength
- Raw readlength >10kb results in CCS coverage of >750 CGG repeats
- Complete characterization at base-resolution is now possible for all disease-relevant alleles

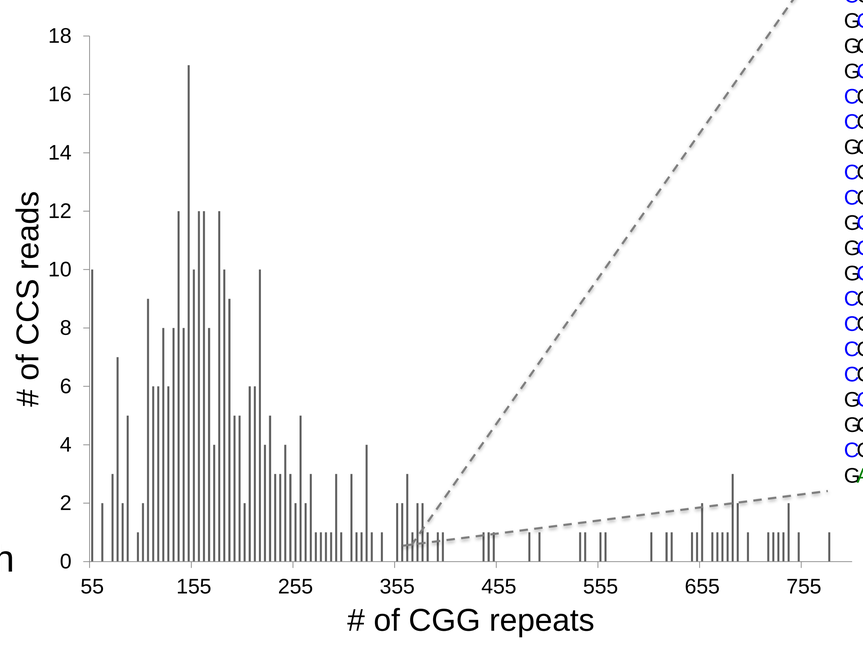

**Figure 3:** Plot of CCS reads across a range of CGG repeat sizes. Peak near 155 repeats is a combined result of SMRT cell passive loading bias and the advantage for smaller inserts to achieve the 3x coverage needed for CCS alignment. Example: CCS read containing 372 CGG repeats

## References:

1. Oostra, B.A. and R. Willemsen, FMR1: a gene with three faces. Biochim Biophys Acta, 2009. 1790(6): p. 467-77.
2. Eid, J., et al., Real-time DNA sequencing from single polymerase molecules. Science, 2009. 323(5910): p. 133-8.

You may contact the first author (during and after the meeting) at ewloomis@ucdavis.edu

Disclosure: This work was presented at the 15th International Workshop on Fragile X and Early-Onset Cognitive Impairment, but has not been previously published.

**Funding:** This project was funded by the NICHD and in collaboration with Pacific Biosciences of California Inc.