Long-Read, Single-Molecule Applications for Protein Engineering

David Alexander, Tyson Bowen, Brett Bowman, Michael Brown, James Bullard, John Harting, Lance Hepler, Regina Lam, Walter Lee, Patrick Marks, Erik Miller
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

The long read lengths of PacBio's SMRT® Sequencing enable detection of linked mutations across multiple kilobases of sequence. This feature is particularly useful in the context of protein engineering, where large numbers of similar constructs are generated routinely to explore the effects of mutations on function and stability.

We have developed a PCR-based barcoded sequencing method to generate high quality, full-length sequence data for batches of constructs generated in a common backbone. Individual barcodes are coupled to primers targeting a common region of the vector of interest. The amplified products are pooled into a single DNA library, and sequencing data are clustered by barcode to generate multi-molecule consensus sequences for each construct present in the pool.

As a proof-of-concept dataset, we have generated a library of 384 randomly mutated variants of the Phi29 DNA polymerase, a 575 amino acid protein encoded by a 1.7 kB gene. These variants were amplified with a set of barcoded primers, and the resulting library was sequenced on a single SMRT Cell. The data produced sequences that were completely concordant with independent Sanger sequencing, for a 100% accurate reconstruction of the set of clones.

Methods

Error-prone PCR resulted in the generation of randomly mutated constructs of the Phi29 DNA polymerase gene. Of these mutants, 384 were isolated as a test case for amplicon sequencing.

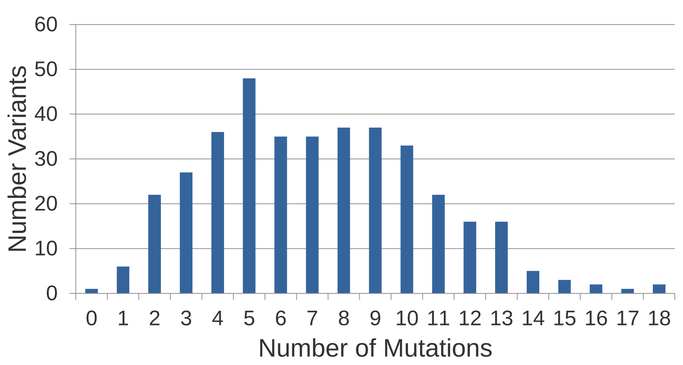

Figure 1. Random mutagenesis of Phi29 DNA polymerase. We generated a panel of randomly mutated constructs using the GeneMorph® II kit (Stratagene) and the Phi29 DNA polymerase gene (1.7 kB) as the template. DpnI was used to eliminate the template plasmid, mutagenized inserts were cloned into the pET11a vector using the NdeI and BamHI sites, and individual colonies were isolated after transformation. Plasmids were purified from the isolates using Plasmid Miniprep kits (Qiagen). A mean of 7.2 mutations per construct were obtained.

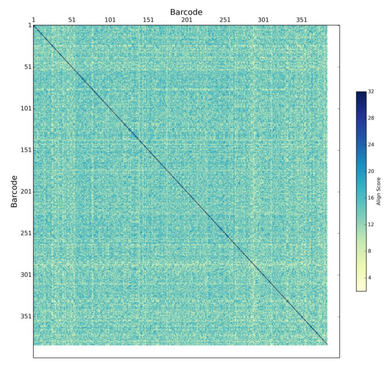

Figure 2. Barcodes for SMRT Sequencing. A set of 384 16-base barcodes was optimized for detection and discrimination in SMRT Sequencing. Pairwise alignment scores for the 384 barcodes are shown.

An optimized set of barcodes was incorporated into primers directed to common vector sequences flanking the cloning site.

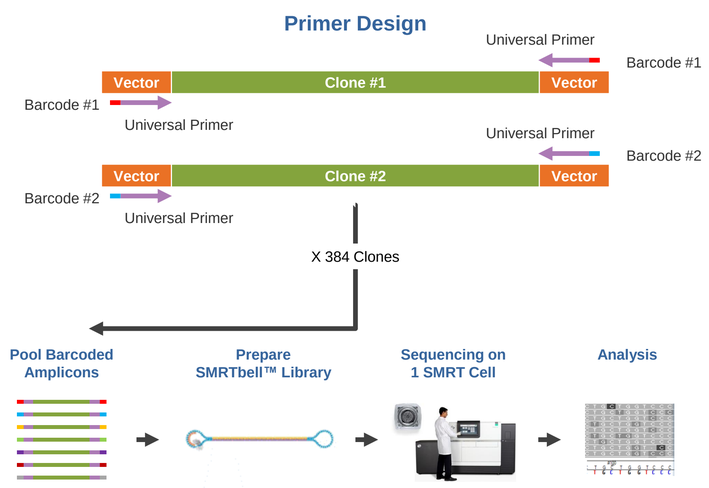

Figure 3. Library prep and sequencing: A panel of 384 primers with unique barcodes was generated to amplify inserts within the pET11a vector. PCR amplification was performed using the Phusion® II polymerase (Thermo Scientific). PCR products were pooled and purified with AMPure® PB magnetic beads (Pacific Biosciences), and 15 μg of DNA was prepared with the PacBio Template Prep Kit. Data collection was performed using P4-C2 Chemistry on the PacBio RS II in a single 90-minute movie. Data were analyzed with SMRT Analysis version 2.2.0 using pbbarcode to demultiplex barcodes and Long Amplicon Analysis (LAA) protocol for consensus. LAA was used with clustering and phasing algorithms turned off, resulting in a per-barcode *de novo* consensus sequence using subreads of minimum read quality 75 and read length 1500 bp. Consensus sequences in multi-FASTA format were compared against the corresponding consensus sequence derived from Sanger Sequencing for concordance using BLASR (Pacific Biosciences).

To obtain independent data for comparison, plasmids were also Sanger sequenced using five priming sites per plasmid (Sequetech).

Results

The sequences derived from SMRT Sequencing and Sanger sequencing were 100% identical. Single-molecule reads were down sampled to measure requirements for accurate sequence determination.

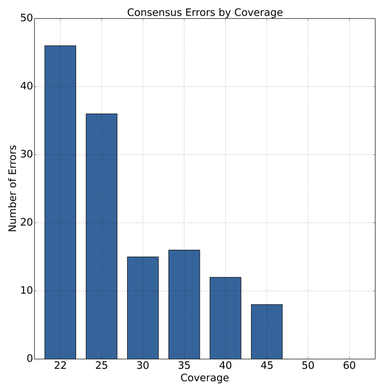

Figure 4. Coverage requirements for assembly. Sequences were determined against a subset of subreads. The total number of errors for the set of 384 constructs (~700 kB of sequence) is shown. An error rate of 10^{-5} is obtained with 45X coverage, and no errors were detected above 50X coverage.

Over 100 subreads were obtained for each of the 384 barcoded constructs pooled in this sequencing run.

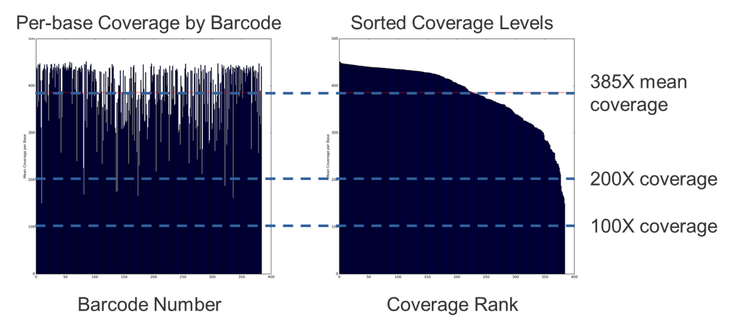

Figure 5. Coverage levels by barcode. Simple pooling of PCR products resulted in >100X coverage for all clones in the dataset. 50X coverage was sufficient for QV > 50 assembly.

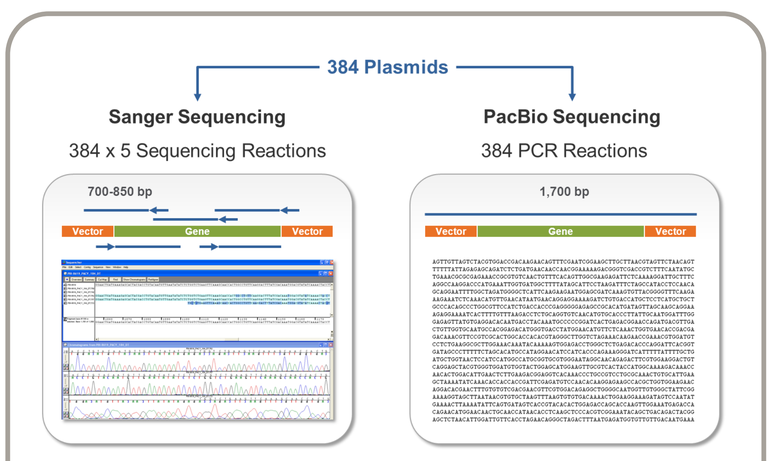

Figure 6. Downstream data analysis. For Sanger analysis, 1920 reactions were assessed for quality and reassembled into complete sequences. A single SMRT® cell produced identical, fully assembled sequences from a single pooled library.

Conclusions

The combination of long read length and high consensus accuracy makes the PacBio® RS II an excellent system for high-throughput clone validation in protein engineering and display pipelines. Full-length reads allow easy sorting of highly similar sequences and confident linkage of multiple mutations within individual clones. Here, we have used a barcoding strategy to demonstrate perfect sequencing of 384 mutated constructs of the Phi29 DNA polymerase using a single SMRT Cell. The low per-run cost and rapid turnaround of this method make it an excellent sequencing tool for selection pipelines.

Acknowledgements:
The authors would like to thank Kristin Robertshaw and Steve Kujawa for their assistance in preparing this poster.