Long-Read, Single-Molecule Sequencing Applications for Protein Engineering

Tyson Bowen, Senior Research Associate, Enzymology
August 27, 2014
Introduction to PacBio® SMRT® Sequencing Technology

SMRT® Sequencing for High-throughput Protein Engineering
General Idea: Eavesdrop on a DNA Polymerase in Real Time

- Nucleotides
  - Fluorescently labeled through phosphates

- Template

- DNA polymerase
  - Fast
  - Processive
  - Frugal
  - Faithful: 1 in $10^5$
  - Small

Newly Synthesized Strand
Challenge – Observing Enzyme Dynamics in Real Time

- In traditional TIRF microscopy the diffusion background increases noise significantly
Zero-Mode Waveguides Facilitate Single-Molecule Measurements

Very little light reaches bulk solution

Observation Volume Confinement
~20 x 10^{-21} liters

SMRT® Cell
Array of 150,000 ZMWs

PacBio® RS II

- Fast intrinsic sequencing rate
- Very long read length (P5-C3)
  - Average RL of 8.5 kb
  - Half of bases come from reads >10 kb
  - Longest reads ~30-40 kb
- Simple sample prep – no amplification
  - DNA modifications and damage have unique kinetic signatures
- Highest consensus accuracy

- De novo sequencing: The *E. coli* K12 genome can be assembled into a single contig at >99.999% accuracy from a single SMRT® Cell in a 2 hr movie
Overview of SMRTbell™ Template Preparation

- Isolate genomic DNA
- Shear or digest
- End repair
- Ligate adapters
- Anneal primer and bind Pol

SMRTbell™ DNA library

- Stalled Pol/DNA/Nucleotide complex
- Immobilize to chip surface by diffusion or magnetic beads

- Application determines fragment size
  - ~250 – 50,000 bases (CCS vs. CLR)
- Barcoding is supported for multiplexing
  - PCR: Introduce barcodes using primers
What Single Molecule Sequencing looks like

- The 4 colors/bases are easily distinguished
- 2-3 bases per second

~30 seconds
Read Length with SMRT® Sequencing

Typical Sanger Trace:
~800 bases
Read Length with SMRT® Sequencing

Typical Sanger Trace:
~800 bases
Introduction to PacBio® SMRT® Sequencing Technology

SMRT® Sequencing for High-throughput Protein Engineering
Phi29 DNA Polymerase

• DNA polymerase from *Bacillus subtilis* phage phi29
  
  – Well understood
  
  – B-family polymerase
  
  – Highly processive
  
  – High fidelity (3’ to 5’ exonuclease)
  
  – Good strand-displacement activity
  
  – Single subunit (66 kDa)

Enzyme Engineering Design Cycle

1. Formulate design hypotheses
2. High-throughput cloning & purification
3. Screening & data interpretation
4. Detailed studies on interesting mutants
High-throughput cloning & purification

Site Directed Mutagenesis

Transformation and Plasmid Purification

Sequence Verification

Protein Expression and Purification

Protein Studies

Easy to Multiplex

Evaluating PhyNexus Lysate Direct PhyTip® Columns for automated Plasmid DNA Purification

Sequence Verification: Sanger vs PacBio® SMRT® Sequencing

Use PhyTip® Affinity Columns to purify with automated liquid handling
Project Overview: Validating 384 Individual Clones 1.7 kb in Length

**Goal**
- To compare a Sanger workflow with a PacBio® workflow (including barcoding) for clone validation

**Scope**
- Validating 384 distinct phi29 clones, each 1.7 kb in length, with high sequence homology (~99%)

**Project Design**
- Sanger: Five amplicons of ~750 bp for each clone
- PacBio: One barcoded amplicon of 1.7 kb for each clone

**Results**
- Sequenced all clones with 100% accuracy using one barcoded library and one SMRT® Cell
Overview of PacBio’s SMRT® Sequencing Method

Primer Design

- Vector
- Universal Primer
- Barcode #1
- Mutation A
- Mutation B
- Vector
- Universal Primer
- Barcode #2
- Mutation C
- Mutation D
- Universal Primer

Pool Barcoded Amplicons

Prepare SMRTbell™ Library

Sequencing on 1 SMRT Cell

Analysis

X 384 Clones
Each SMRT® Cell Will Generate ~50,000 Reads

Barcodes Are Identified and Sequences Are Binned (384 Bins)

Sequences from Each Bin Are Aligned and Bases Are Called

De novo assembly for each barcode in a single fasta file
Coverage Required for Accurate Sequencing

- Assemblies performed with subsets of data at differing levels of coverage
- At 45X coverage, errors detected with a frequency of $10^{-5}$
- Above 50X coverage, no errors detected in set of 384 barcoded 1.7 kb amplicons
Simple pooling of PCR products produced >100X coverage for all 384 clones in a single run.
Complexity of Sanger Assembly Scales With Insert Size

Sanger Assembly

Vector 1kb insert Vector

6kb

PacBio – No Assembly Required!

Vector 1kb insert Vector

6kb
PacBio® Barcoded-Insert Sequencing for Protein Engineering

- PCR-based sample preparation is convenient for automated high-throughput pipelines

- Multi-molecule consensus provides high accuracy sequences at per-clone costs lower than Sanger validation

- Data are fully assembled in FASTA format, simplifying bioinformatics analysis

- Turnaround times (~24 hrs from PCR product to data) enable real-time sampling of selection rounds
# Acknowledgement Slide

## Enzymology
- Keith Bjornson
- Regina Lam
- Jeremiah Hanes
- Satwik Kamtekar
- Harold Lee
- Walter Lee
- Erik Miller
- Leewin Chern
- Phil Jiao
- Lukasz Joachimiak
- Thang Pham
- Simeon Pinaula
- Jessica Soriano
- Minh Ton
- Kathy Tran

## Bioinformatics
- David Alexander
- James Bullard
- Brett Bowman
- Michael Brown
- John Harting
- Lance Hepler
- Patrick Marks

## Pacific Biosciences
- Jonas Koolach
- Steve Kujawa
- Kristin Robertshaw
- Steve Turner