Best Practices for Whole Genome Sequencing Using the Sequel System

Justin Bliethrow, Nick Sisneros, Shreyasee Chakraborty, Sarah Kingan, Richard Hall, Joan Wilson, Christine Lambert, Kevin Eng, Emily Hatas and Primo Baybayan

PacBio, 1305 O’Brien Dr., Menlo Park, CA, 94025

Abstract

Plant and animal whole genome sequencing has proven to be challenging, particularly due to genome size, high density of repetitive elements and heterozygosity. The Sequel System delivers long reads, high consensus accuracy and uniform coverage, enabling more complete, accurate, and contiguous assemblies of these large complex genomes. The latest Sequel chemistry increases yield up to 8Gb per SMRT Cell for long insert libraries >20 kb and up to 10Gb per SMRT Cell for libraries >40 kb. In addition, the recently released SMRTbell Express Template Prep Kit reduces the time (~3 hours) and DNA input (~3 µg), making the workflow easy to use for multi-SMRT Cell projects.

Here, we recommend the best practices for whole genome sequencing and de novo assembly of complex plant and animal genomes. Guidelines for constructing large-insert SMRTbell libraries (>30 kb) to generate optimal read lengths, and yields using the latest Sequel chemistry are presented. We also describe ways to maximize library yield preparaion from as little as 1 µg of sheared genomic DNA. The combination of these advances makes plant and animal whole genome sequencing a practical application of the Sequel System.

Large-insert Workflow: DNA to Sequence

QC Data: Pulse Field Gel Electrophoresis

- Post and Large Scale Shear
- Eco RV Treatment
- DNA Damage Repair
- Gen Repair/Cutting
- Adaptor Ligation
- Isolation/Extraction
- AMF Purification
- Size Selection
- AMF Purification
- Primer Annealing/Fragment Binding
- Sequel System

Recommended Shearing Optimization

Recommended Shearing Devices for Large-insert Fragments

Pulsed Field Gel Electrophoresis (PFGE) quality control
- Optimization of shearing parameters
- Proper size selection cutoff
- Damage repair after size selection
- Following bogus recommendations

DNA Requirements for Whole Genome Sequencing

- The total amount of DNA required for whole genome sequencing depends on project requirements (e.g. genome size, coverage, genome complexity, etc.).
- When designing experiments, the starting DNA yield by using the following library yield equations:

Sample QC Highly Recommended

- SMRTbell
- Pippin
- Fragment Mapping
- SMRTbell Sequencing

Data Analysis

Hierarchical Genome Assembly Process (HGAP) and Polishing

- Error-corrected reads are assembled with a string graph or noise spanning, generating primary and alternate contigs that represent the alternative alleles, between the haplotypes. FALCON Unzip identifies heterozygous SNPs in FALCON contigs and uses these SNPs to phase reads. The phased reads are then used to refine the assembly graph, resulting in an extension of the haplotype phasing originally captured in FALCON assembly graph bubbles.

- De novo assembly using either HGAP, FALCON, or FALCON-Unzip algorithms

Diploid Genome Assembly with FALCON and FALCON-Unzip

- Primary Contigs
- Longest Contig
- Subread Coverage
- Subread N50

Summary and Resources

- The Sequel System achieves 8-10 Gb of data per SMRT Cell with long insert libraries (>30 kb)
- SMRTbell Express template preparation reduces time and DNA input needed to generate long insert libraries
- Follow best practices to improve performance and overall project results
- Pulsed Field Gel Electrophoresis is important for assessing input genomic DNA, sheared DNA, SMRTbell library and final size selected SMRTbell library
- The Megaruptor, a high molecular weight fragmentation instrument, may be used directly for library preparation if necessary
- Post shearing optimization allows for library quality control and determining optimal shearing parameters
- Follow best practices to improve performance and overall project results
- Error-corrected reads are assembled with a string graph or noise spanning

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2018 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRT Cell, Bio-Shear, and Sequel are trademarks of Pacific Biosciences. Bio-Pippin and SageELF are trademarks of Sage Science. H3K-9 and H3K-18 are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of Advanced Analytical Technologies. All other trademarks are the sole property of their respective owners.