Best Practices for Whole-Genome *De Novo* Sequencing with Long-read SMRT Sequencing

David Rank, Kristin Gleitsman, Wenlong Jiang, Kristi Spittle Kim, Nick Sisneros, Joan Wilson, Marty Badgett, Paul Peluso
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Abstract

With the introduction of P6-C4 chemistry, PacBio has made significant strides with Single Molecule, Real-Time (SMRT®) Sequencing . Read lengths averaging between 10 and 15 kb can be now be achieved with extreme reads in the distribution of > 60 kb. The chemistry attains a consensus accuracy of 99.999% (QV50) at 30x coverage which coupled with an increased throughput from the PacBio® RS II platform (500 Mb – 1 Gb per SMRT Cell) makes larger genome projects more tractable. These combined advancements in technology deliver results that rival the quality of Sanger "clone-by-clone" sequencing efforts; resulting in closed microbial genomes and highly contiguous *de novo* assembly of complex eukaryotes on multi-Gbase scale using SMRT Sequencing as the standalone technology.

We present here the guidelines and best practices to achieve optimal results when employing PacBio-only whole genome shotgun sequencing strategies. Specific sequencing examples for plant and animal genomes are discussed with SMRTbell™ library preparation and purification methods for obtaining long insert libraries to generate optimal sequencing results. The benefits of long reads are demonstrated by the highly contiguous assemblies yielding contig N50s of over 5 Mb compared to similar assemblies using next-generation short-read approaches. Finally, guidelines will be presented for planning out projects for the *de novo* assembly of large genomes.

Workflow Matrix

Pacific Biosciences Workflow Matrix

Input DNA (µg/100 Mb) (Target Organism)	1	3	3	6	10	25
Library Size (cut-off)	7 kb	15 kb	7 kb	15 kb	7 kb	15 kb
Assembly method	Gap Filling: PBJelly		Hybrid: EC Tools/ PBcR Spades		PacBio Only: HGAP Dazzler, Falcon MHAP	
Coverage Required	10X PacBio + short read		20X PacBio + short read		50X PacBio	
Assembly Quality						
Contig N50	Gap Filled		0.1-1.5 Mb		1-15 Mb	
QV	95%-99% (filled gaps)		99.995+%		99.999+%	

Improvements

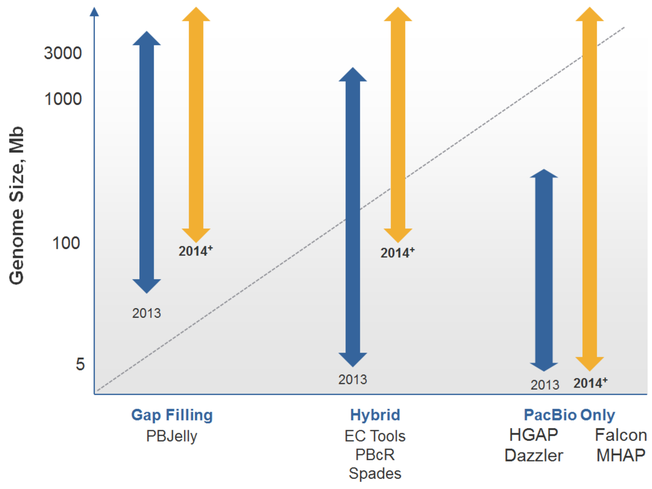


Figure 1
Improvements in SMRT Sequencing and assembly algorithms over time have facilitated the assembly of larger more complex organisms using all PacBio data. Furthermore, as template preparation, chemistry, and assembly methods evolve, the quality of finished genomes will improve substantially.

Sample QC & Shearing

Large-insert SMRTbell library preparation requires DNA of the highest quality and molecular weight. In order to characterize the DNA prior to starting, it is recommended to take a careful look at results from Nanodrop, Qubit and Pulsed-field Gel Electrophoresis (PFGE) runs to obtain the longest reads.

A) Genomic DNA QC & Test Shearing of DNA

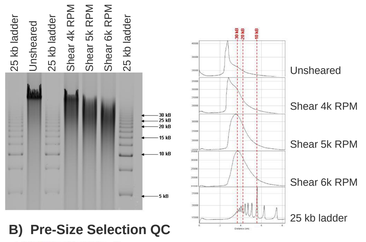
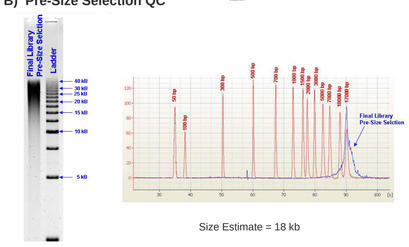


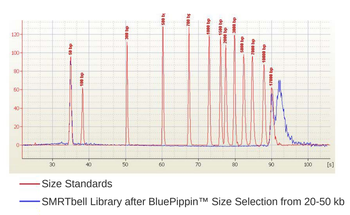
Figure 2
(A) For appropriate shearing and library construction, determine gDNA size distribution prior to shearing. PFGE is highly recommended to determine sample quality.
Ideal gDNA >50 kb mode
Test shearing is performed to evaluate ideal shear conditions resulting in appropriate library distribution.
Ideal mode ~20 kb (~90% size distribution <30 kb)
(B) Pre-library size selection. *Estimate >18 kb*
(C) *Step 3*: Assess final library insert size and quality.
Ideal library insert size >18 kb

B) Pre-Size Selection QC



Size Estimate = 18 kb

C) Final Library QC

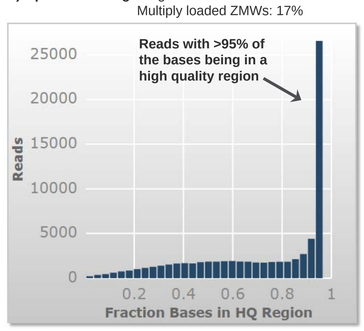


Size Estimate = 18 kb

Loading Titrations

For optimal loading and data output, a titration of the sample library is recommended. This ensures that the library is not underloaded resulting in a decrease in per-SMRT-Cell output, and is not overloaded resulting in a decrease in read quality. The following plots are results from primary analysis from the PacBio RS II.

A) Optimal Loading: Single loaded ZMWs: 48%
Multiply loaded ZMWs: 17%



B) Over Loading: Single loaded ZMWs: 49%
Multiply loaded ZMWs: 38%

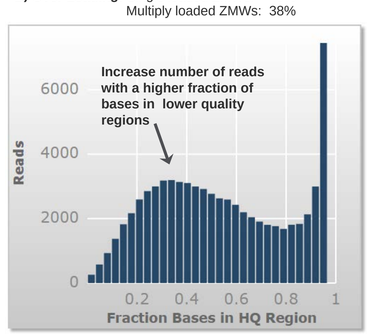


Figure 3
(A) Example of ideal loading. Maximizing overall throughput and maintaining quality reads.

(B) Overloading leads to a higher number ZMWs occupied by more than one molecule. Fewer of these ZMWs will result in reads with very long regions of high quality bases. Instead these reads will have regions of lower quality that are filtered out during post primary processes.

Mapped Read Length Statistics

An increase in longer reads can be obtained with long-insert libraries and size selection on the BluePippin™ system (Sage Science) to exclude short fragments.

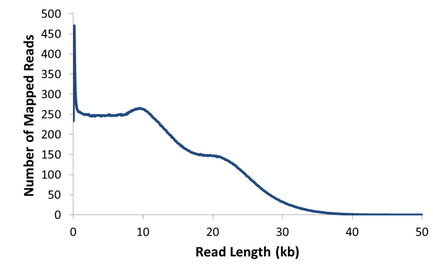
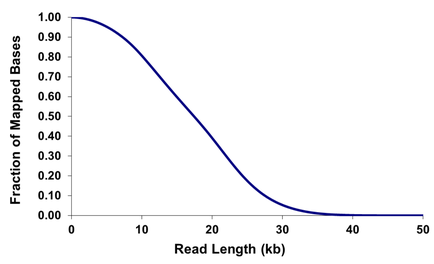


Figure 4
Shown is an example of the mapped sub-read length distribution from a SMRT Cell loaded with ~50,000 singly loaded ZMWs.



gure 5
umulative distribution plot of the ume mapped data from Figure 4.

Production Runs

Example production run from one library sample processed over multiple instruments and days. Library was created using the latest sample preparation improvements.

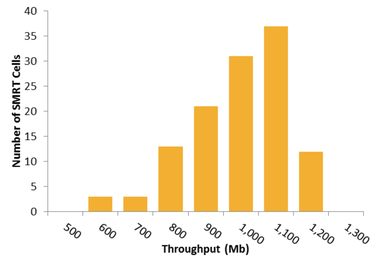


Figure 6
Histogram of total throughput for 120 SMRT Cells. Runs were from a single library including titration runs.

Production Run Statistics	
Average Mapped Read Length (bases)	11,900
Mapped Read Length N50 (bases)	*17,000
Mapped Read Length Max (bases)	56,000
[On Chip] pM	10
Input SMRTbell Library per chip (ng)	8
Yield per chip (Mb)	1000

Table 1
Production run summary statistics on mapped reads.

*N50 raw reads 20 kb.

Assembly Methods & Considerations

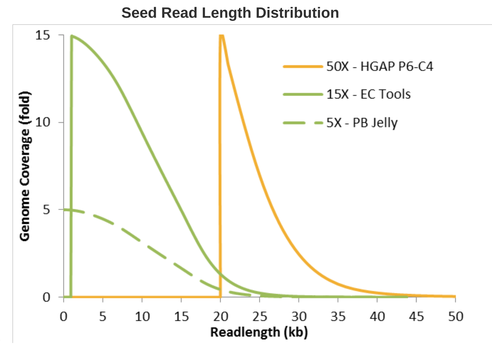


Figure 7
Read-length distribution for raw read used for inputs in each of the assemblies, EC Tools and PB Jelly and pre-assembly seed reads for HGAP. As coverage levels increase, the number of reads >20 kb increases substantially.

Assembly Method	Avg Read Length (kb)	N50 Read Length (kb)	Coverage @ 20 kb	% Genome Coverage @ 20 kb
HGAP P6-C4	24.9	24.5	15X	100
EC Tools	8.4	12.1	1.3X	73.3
PB Jelly	8.4	12.1	0.4X	32.9

Table 2
Summary comparisons of the assembly methods by increasing coverage levels demonstrates an increase in assemble-able bases in reads >20 kb.

Conclusions

With the recent improvements in PacBio's long-read chemistry, longer insert libraries are required to take full advantage of these advances. Shown here are the most up-to-date approaches towards generating long libraries leveraging the BluePippin™ size-selection system. Also provided are considerations for assessing a large genome project. Factors such as input DNA quality, desired coverage, genome size and *de novo* assembly method(s) are all important components that impact the project results. The benefits of long reads are characterized by highly contiguous assemblies with contig N50s of over 5 Mb. Therefore, we recommend the approaches outlined here to maximize your success.

Acknowledgements

The authors would like to thank everyone who helped generate data for the poster and have contributed to the ongoing improvements in generating long-insert libraries.