Best Practices for Whole Genome Sequencing Using the Sequel System

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

PacBio, 1380 Willow Road, 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 which enable more complete, accurate, and contiguous assemblies of these large, complex genomes. The latest Sequel chemistry can produce 5 – 8 Gb per SMRT Cell with reduced input SMRTbell libraries (as low as 5 pM). Read lengths averaging 10 – 15 kb can be routinely achieved, with the longest reads >60 kb. Furthermore, 50% of useable bases are in reads greater than 20 kb.

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 using the latest Sequel chemistry are presented. We also describe ways to maximize library yield per preparation from as littles as 5 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 Library Construction Workflow

QC DNA by Pulsed Field Gel Electrophoresis
↓
Test and Large Scale Shear
↓
Exo VII Treatment
↓
DNA Damage Repair
↓
End-Repair
↓
Adapter Ligation
↓
Exo III and Exo VII Treatment
↓
Size Selection
↓
Post Size Selection DNA Damage Repair
↓
Primer Annealing
↓
Polymerase Binding
↓
SMRT Sequencing

Large-insert library construction success increases with:

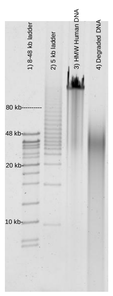
- High-molecular weight DNA
- Pulsed Field Gel Electrophoresis (PFGE) quality control
- Optimization of shearing parameters
- Proper size-selection cutoff
- Damage repair after size selection
- Following loading recommendations

Sequel System

Sample QC Highly Recommended

A. BIO-RAD® CHEF Mapper® XA Pulsed Field Electrophoresis System

B. Sage Science™ Pippin Pulse Electrophoresis Power Supply System



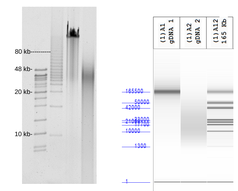
C. High Molecular Weight vs. Degraded DNA run on CHEF Mapper

A PFGE run provides information on sample quality and fragment size. PacBio recommends using either Bio-Rad's CHEF Mapper **(A)** or Sage Science Pippin Pulse **(B)** as PFGE instruments.

The gel image **(C)** shows high molecular weight DNA at ~150 kb which can be sheared to the desired size (>30 kb). Lane 4 shows a less than ideal gDNA with a smear up to 80 kb. Depending on the severity of degradation, the sample may be used directly for library construction. A size selection cutoff of 10 kb usually generates good subread lengths.

D. Advanced Analytical FEMTO Pulse™ Automated Pulsed-Field CE Instrument

While both CHEF Mapper and Pippin Pulse are reliable systems for characterizing genomic DNA, electrophoresis run times are intensive (>16 hrs) and require significant amount of DNA as input. Advanced Analytical's FEMTO Pulse instrument **(D)** is a fast high-resolution capillary based electrophoresis system able to resolve fragments up to 165 kb in one hour, ideal when constructing large-insert libraries. More importantly, the system requires picogram (pg) quantities of DNA.



E. CHEF Mapper

F. FEMTO Pulse

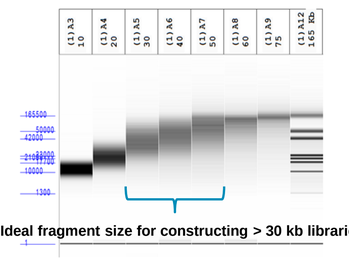
Human genomic DNA was also loaded on the CHEF Mapper and FEMTO Pulse. Separation observed in CHEF Mapper (**E**) exhibits comparable performance as the Femto Pulse (**F**).

Library Construction Recommendations

Recommended Shearing Devices for Large-insert Fragments

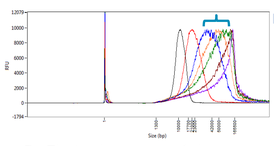
Megaruptor® DNA Shearing System

The g-TUBE device from Covaris cannot shear gDNA to >30 kb fragments. PacBio recommends the Megaruptor - a simple, automated, and highly reproducible system to fragment DNA up to 75 kb.



Ideal fragment size for constructing > 30 kb libraries

A. Megaruptor shears loaded on FEMTO Pulse



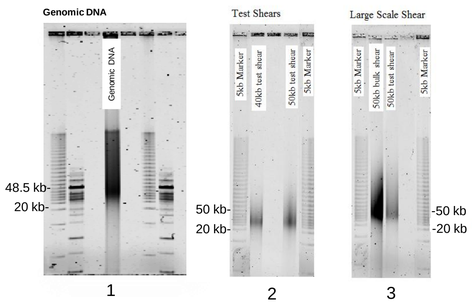
B. Electropherograms of shears

Label	Peak Max	Average
10 kB	11304	12245
20 kB	21245	24123
30 kB	36399	46143
40 kB	49495	60600
50 kB	78351	75416
60 kB	134525	92133
75 kB	149225	104512

C. Sizing report of shears

To demonstrate shearing performance of the Megaruptor, a high molecular weight human genomic DNA was sheared to 10, 20, 30, 40, 50, 60, and 75 kb fragments. In this example, 30, 40, and 50 kb shears are best conditions for constructing >30 kb libraries.

Recommend Shearing Optimization



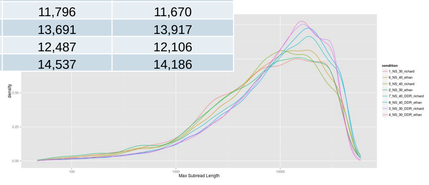
Recommended DNA shearing steps:

1. Assess genomic DNA quality using PFGE
2. Perform test shears, run on PFGE and determine optimal shearing condition
3. Perform large-scale shear

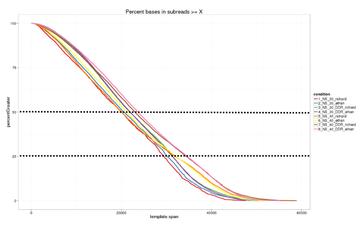
In the example shown on the left, a plant gDNA sample was sheared to 40 and 50 kb. While the 40 kb shear generated good distribution, the 50 kb shear condition was selected for the large-scale shear because it provided slightly larger fragments.

Post Size-selection DNA Damage Repair Improves Read Length

	Mean Mapped Subread Length	
	Sequel System #1	Sequel System #2
NS_30	11,796	11,670
NS_30_DDR	13,691	13,917
NS_40	12,487	12,106
NS_40_DDR	14,537	14,186



Mean mapped subread length is improved by ~2 kb for both 30 and 40 kb libraries after treatment of the size selected library with DNA Damage Repair (DDR) enzymes



Post size-selection DDR increases N50 and N25 subread lengths up to ~2 kb.

	N50 Subread Length		N25 Subread Length	
	Sequel System #1	Sequel System #2	Sequel System #1	Sequel System #2
NS_30	21,043	21,118	29,853	30,476
NS_30_DDR	21,740	21,303	30,733	30,234
NS_40	20,625	21,297	32,096	31,762
NS_40_DDR	23,103	23,036	34,086	33,405

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, estimate the starting DNA requirement by using the following library yield assumptions.

	% Yield
SMRTbell Library After Exo III/VII Treatment	30-40
SMRTbell Library Yield (Post Size selection)*	5-10

*Yield depends on fragment distribution and size-selection cutoff

When DNA is limiting (<500 ng) or low quality, you may need to opt for a non-size selected library. The table below summarizes results from an experiment comparing yield from 5 ug, 500 ng, and 100 ng sheared DNA into library construction.

DNA Input into Exo VII treatment	5,000 ng	500 ng	100 ng
Post Damage Repair Yield	2,300 ng (46%)	290 ng (57%)	65 ng (65%)
Final SMRTbell Library Yield (Non Size Selected)	1,400 ng (30%)	105 ng (21%)	28 ng (28%)

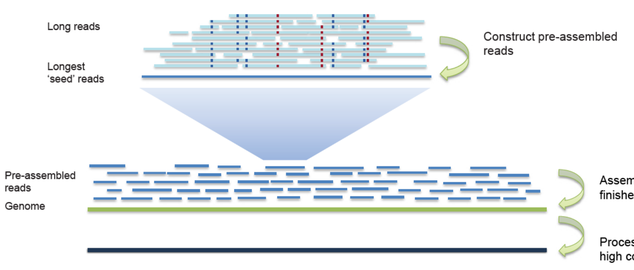
Loading Recommendations

Recent advances in chemistry (Sequel Sequencing Kit 2.0) and improved instrument workflow (Instrument Control Software 4.0) significantly reduced the amount of SMRTbell library required for loading in the Sequel System. The system generates average read lengths between 10 – 15 kb with throughput of 5 – 8 Gb per SMRT Cell while requiring lower loading concentration compared to the PacBio RS II.

Library Size	Sequel System	PacBio RS II
10,000	5 - 20 pM	40 pM
>20,000	5 - 20 pM	>100 pM

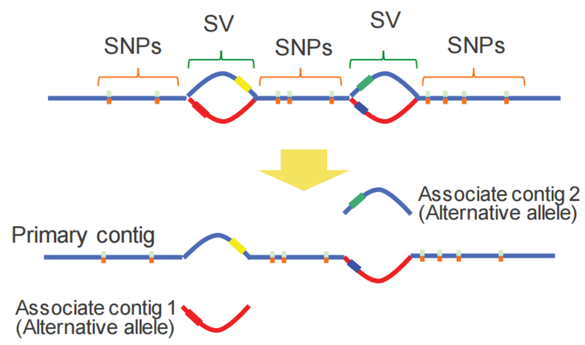
Data Analysis

Hierarchical Genome Assembly Process (HGAP)



HGAP[1] utilizes all PacBio data using the longest reads for contiguity and all reads to generate high-quality *de novo* assemblies with high consensus accuracy (>QV50).

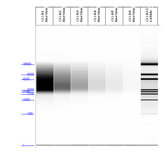
Diploid-aware Genome Assembly with FALCON



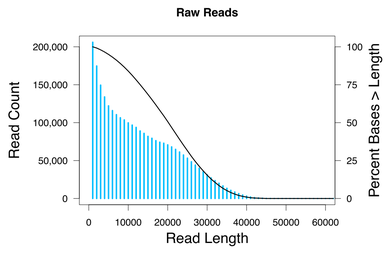
In diploid-aware assembly, error-corrected reads are assembled using a string graph of the read overlaps, generating primary and alternative contigs that represent the alternative alleles, or structural variants (SVs), between the haplotypes[2]. Recommended for heterozygous, non-inbred genome assembly. FALCON is available through Github.

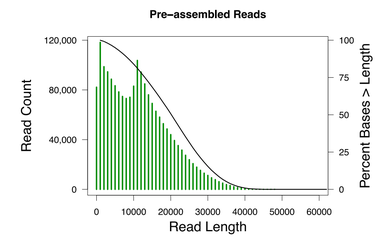
Case Study Example

Organism: Plant
Genome size: 1 Gb
SMRTbell library size: >30 kb, 20 kb size selection cutoff
Sequel SMRT Cells 1M: 6
Chemistry: Sequel Sequencing Kit v1.2.1



>30 kb library analyzed in FEMTO Pulse™ instrument





FALCON Assembly	
Contig N50	1.2 Mb
Total Length	900 Mb
No. Contigs	2,895
Longest Contig	18.9 Mb
Subread Coverage	37 fold
Subread N50	20,179 bp

Summary and Resources

- The Sequel System achieves avg read lengths of 10 – 15 kb with throughput of 5 – 8 Gb per SMRT Cell 1M
- Follow best practices to improve performance and overall project results
- Pulsed Field Gel Electrophoresis is important for assessing incoming genomic DNA, sheared DNA , SMRTbell library and final size-selected SMRTbell library
- The Megaruptor system is recommended for shearing DNA >30 kb
- Optimize shearing conditions by performing tests prior to large-scale shearing
- Treat size-selected libraries with DNA Damage repair enzymes
- *De novo* assembly using either HGAP of FALCON algorithms

Resources:

For all PacBio library prep and sequencing protocols, visit
http://www.pacb.com/support/documentation/
FALCON available on GitHub:
https://github.com/PacificBiosciences/FALCON/

References:

1. Chin, C.S. et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT Sequencing data. *Nature Methods.* 10(6), 563-569.

2. Chin, C.S. et. al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050-1054.