# Best Practices for Whole Genome Sequencing Using the Sequel System

**Justin Blethrow, 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 8 Gb per SMRT Cell for long insert libraries >20 kb and up to 10 Gb 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 per preparation from as littles as 3 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 DNA by Pulsed Field Gel Electrophoresis
↓
Test and Large Scale Shear
↓
Exo VII Treatment
↓
DNA Damage Repair
↓
End-Repair/A-Tailing
↓
Adapter Ligation
↓
Inactivate Enzyme
↓
AMPure Purification
↓
Size Selection
↓
AMPure Purification
↓
Primer Annealing/Polymerase Binding
↓
SMRT Sequencing

Large-insert library construction success increases with:

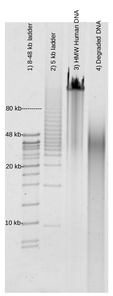
- High-molecular weight DNA
- Pulsed Field Gel Electrophoresis (PFGE) quality control
- Optimization of shearing parameters
- Proper size-selection cutoff
- Damage repair after size selection
- Following loading recommendations

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



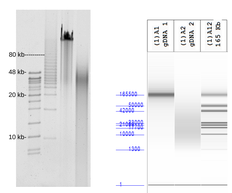

**C.** High Molecular Weight vs. Degraded DNA run on CHEF Mapper

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 >15 kb usually generates suitable 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**).

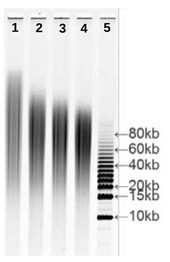
## Library Construction Recommendations

### Recommended Shearing Devices for Large-insert Fragments

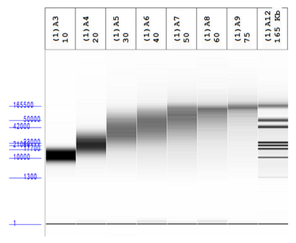
For shearing DNA, PacBio recommends either: 1) needle shearing with a 26 G needle, which allows for flexibility in number of shearing pulses with the needle or 2) the Megaruptor, a simple, automated, and highly reproducible system to fragment DNA up to 75 kb.

Demonstration of Needle Shearing

Megaruptor® DNA Shearing System

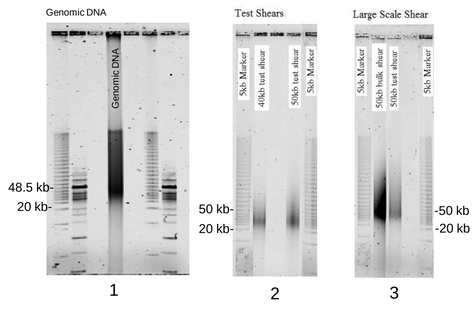



Lane 1: Input K12 gDNA
Lane 2: 26G needle shear, 5x shears
Lane 3: 26G needle shear, 10x shears
Lane 4: 26G needle shear, 20x shears
Lane 5: Bio-Rad 5 kb DNA ladder



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



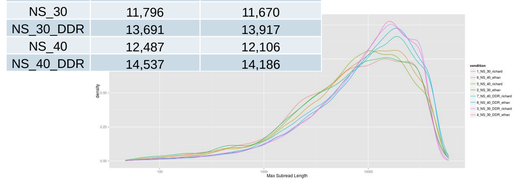

Recommended DNA shearing steps:

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

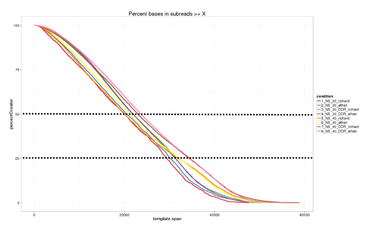
In the example shown on the left, a plant gDNA sample was sheared to 40 and 50 kb. While the 40 kb shear generated good distribution, the 50 kb shear condition was selected for the large-scale shear because it provided slightly larger fragments.

### 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 Yield | >60-70 |
| SMRTbell Library Yield (Post Size Selection)* | >20 |

*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 µg, 500 ng, and 100 ng sheared DNA into library construction.

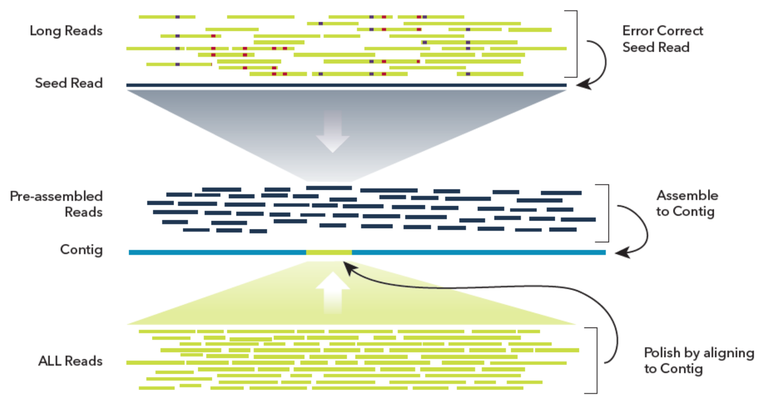
| DNA Input into Exo VII treatment | 5,000 ng | 500 ng | 100 ng |
|---|---|---|---|
| Final SMRTbell Library Yield (Non size-selected) | 3,000 ng (60%) | >300 ng (60%) | >60 ng (60%) |

### Loading Recommendations

PacBio recommends Diffusion Loading when using the Sequel Binding Kit 2.0 and 2.1 and Sequencing Kit 2.1. Note that the SMRTbell Express Template Prep Kit is only compatible with Diffusion Loading. For all library sizes, PacBio recommends a 2 – 8 pmol on-plate loading concentration. Please note that sample quality may influence optimal loading concentrations.

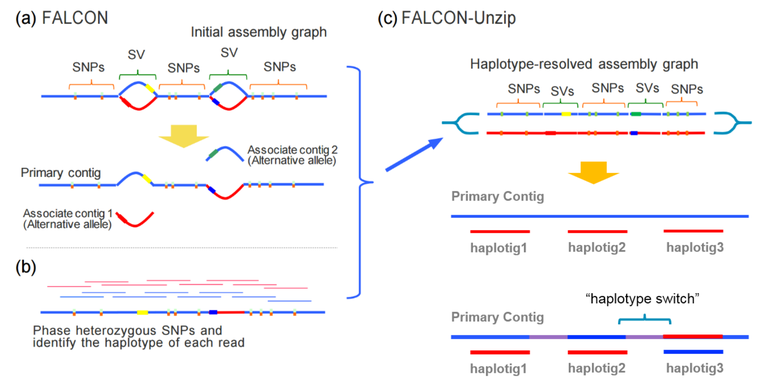
## Data Analysis

### Hierarchical Genome Assembly Process (HGAP) and Polishing




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).

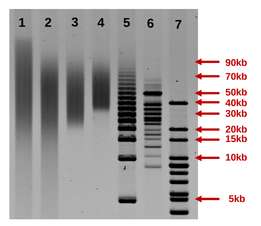
### Diploid Genome Assembly with FALCON and FALCON-Unzip




Error-corrected reads are assembled with a string graph of read overlaps, generating primary and alternate contigs that represent the alternative alleles, between the haplotypes[2]. FALCON-Unzip identifies heterozygous SNPs in FALCON contigs and uses these SNPs to phase reads. The phased reads are then used to redraw the assembly graph, resulting in an extension of the haplotype phasing originally captured in FALCON assembly graph bubbles.

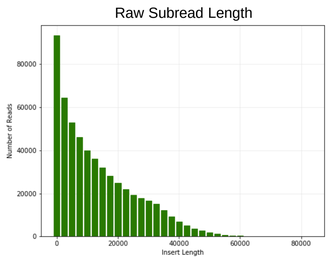
### Case Study: Plant Genome

**Organism:** Plant
**Genome size:** 400 Mb
**SMRTbell library size:** >30 kb, 30-80 kb size selection
**SMRTbell library prep:** SMRTbell Express kit
**Sequel SMRT Cells 1M**: 4
**Chemistry:** Sequel Sequencing Kit v.2.1

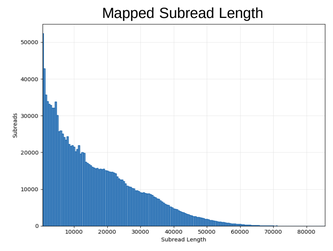



Lane 1: gDNA
Lane 2: non-size selected library
Lane 3: 20-50 kb size-selected library
Lane 4: 30-80 kb size-selected library
Lanes 5-7: DNA ladders

Raw Subread Length




Mapped Subread Length




| HGAP Assembly | |
|---|---|
| **Contig N50** | **1.2 Mb** |
| **Total Length** | **389 Mb** |
| **No. Contigs** | **884** |
| **Longest Contig** | **5.2 Mb** |
| **Subread Coverage** | **62-fold** |
| **Subread N50** | **27,397 bp** |

## 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 system or needle shearing is recommended for shearing DNA >30 kb
- Optimize conditions by performing test shears prior to large-scale shearing
- Treat size-selected libraries with DNA Damage repair enzymes
- *De novo* assembly using either HGAP, FALCON, or FALCON-Unzip 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.