SMRT® Sequencing Solutions for Plant Genomes and Transcriptomes

Jenny Gu, Jason Chin, Paul Peluso, David Rank, Elizabeth Tseng, Tyson Clark, Primo Baybayan
Pacific Biosciences, Menlo Park, CA

Abstract

Single Molecule, Real-Time (SMRT) Sequencing provides efficient, streamlined solutions to address new frontiers in plant genomes and transcriptomes. Inherent challenges presented by highly repetitive, low-complexity regions and duplication events are directly addressed with multi-kilobase read lengths exceeding 8.5 Kb on average, with many exceeding 20 Kb. Differentiating between transcript isoforms that are difficult to resolve with short-read technologies is also now possible.

We present solutions available for both reference genome and transcriptome research that best leverage long reads in several plant projects including algae, *Arabidopsis*, rice, and spinach using only the PacBio® platform. Benefits for these applications are further realized with consistent use of size-selection of input sample using the BluePippin™ device from Sage Science. We will share highlights from our genome projects using the latest P5-C3 chemistry to generate high-quality reference genomes with the highest contiguity, contig N50 exceeding 1 Mb, and average base quality of QV50. Additionally, the value of long, intact reads to provide a no-assembly approach to investigate transcript isoforms using our Iso-Seq™ protocol will be presented for full transcriptome characterization and targeted surveys of genes with complex structures.

PacBio provides the most comprehensive assembly with annotation when combining offerings for both genome and transcriptome research efforts. For more focused investigation, PacBio also offers researchers opportunities to easily investigate and survey genes with complex structures.

Methods

PacBio® RS II Sequencing Chemistries Provide Long Read Lengths >20 Kb

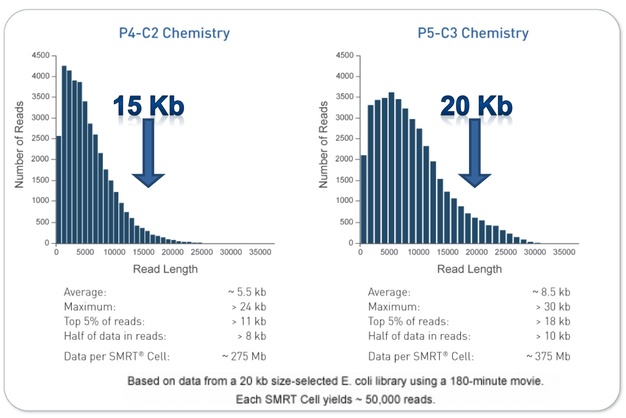


Figure 1. Example read length distribution from a SMRT Sequencing run with 20 kb size-selected *E. coli* library using a 180 min movie. Average throughput of 350 Mb per SMRT Cell with ~50,000 reads.

Hierarchical Genome Assembly Process (HGAP)[1]

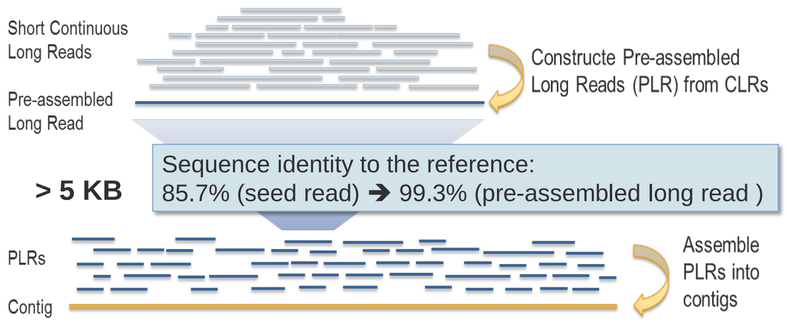


Figure 2. Schematic of Hierarchical Genome Assembly Process to make best use of all PacBio reads and improve consensus accuracy of pre-assembled long reads.

PacBio-Only Arabidopsis Genome

Optimize Long Read Lengths with Library Size Selection:

Electropherogram of SMRTbell™ library
HMW DNA and BluePippin™ Size Selection

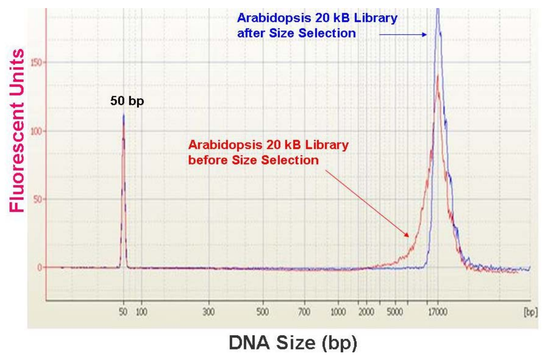
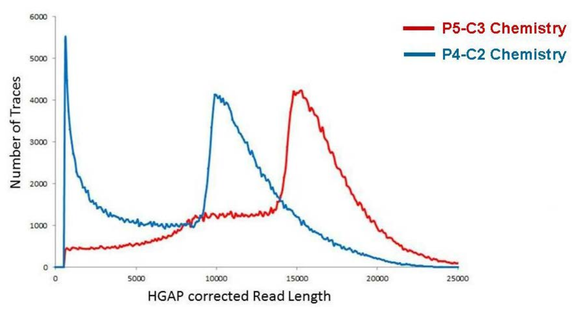


Figure 3. Size distribution of SMRTbell library before and after size selection using the BluePippin system from Sage Science. The size-selected library (blue) was used for sequencing an *Arabidopsis* genome using only PacBio reads.

A



B

Assembly Statistics Summary

Method	Ler-0 P4-C2	Ler-0 P5-C3	Ler-1 Short Read*
Assembly Size (Mbp)	123	122	110
# Contigs	983	322	4,662
Contig N50 (Mbp)	5.3	7.6	0.066
Max Contig Length (Mbp)	13.3	15.6	0.462

*Source: http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/current/Ler-1/assemblies/Allpaths_LG/Ler-1.allpaths_lg.final.assembly.fasta

C

Ler-0 Assembly (P5-C3) Mapped to Col-0 (TAIR-10) Reference

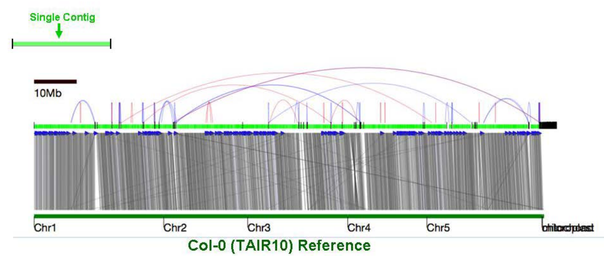


Figure 4 (A) Pre-assembled read-length distribution comparisons between two PacBio sequencing chemistries **(B)** Comparison of Ler-0 assemblies with P5-C3 and P4-C2 chemistries. P5-C3 demonstrate longer N50 and max contig sizes. **(C)** Graph demonstrating the contiguity, synteny, and the low number of relevant contigs (~35) in the final P5-C3 assembly.

PacBio-Only Spinach Genome

Spinach Genome (P5-C3 Chemistry)

Time to completion: 3 months

Draft Assembly: 173.4 Mb from 108 scaffolds > 1 Mb (7400 contigs, total 21 Mb bases in the gaps of the scaffolds)

PB Assembly: 446 Mb from 252 contigs > 1 Mb

	Draft Scaffolds	Draft Contigs	PacBio® Draft Contigs
# of Sequences	1,075,770	1,128,199	5,059
Total Size (bases)	969 Mb	817 Mb	952 Mb
Max Length	5,885,170	469,176	5,439,550
N50	389,694	21,463	920,046
N90	135	135	138,942

Figure 5 - Summary table of Spinach assembly results.

Survey Genes with Complex Structures

Overview of Detailed Gluten Gene Content and Composition from 10 Wheat Cultivars[2]

- 424 proteins characterized
- Full-length sequencing unambiguously resolves complex gene structures without assembly
- Obtained with a single SMRT® Cell run.
- High-throughput multiplex option available

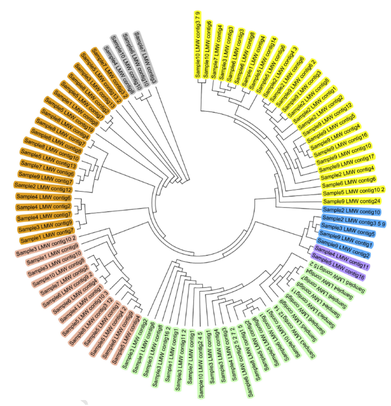


Figure 6 - Phylogenetic tree of full-length gluten gene structures with no assembly needed.

Achieve Complete BAC Sequencing

Gene Triplication Event Detected with BAC Sequencing for Aluminum Tolerance in Maize[3]

- Segregating population localized a QTL on a BAC, but unable to genotype with short-read sequencing because of high repeat content and GC skew
- BAC assembly with PacBio long reads resolved a triplication event of the ZnMATE1 membrane transporter
- Important for drought resistance and protecting against nutrient deficiencies

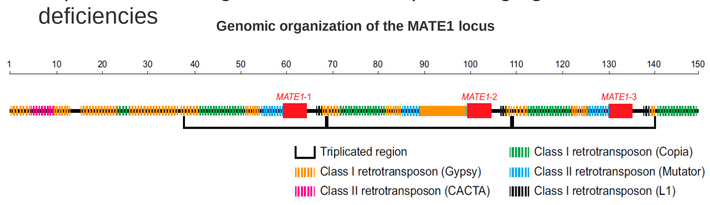
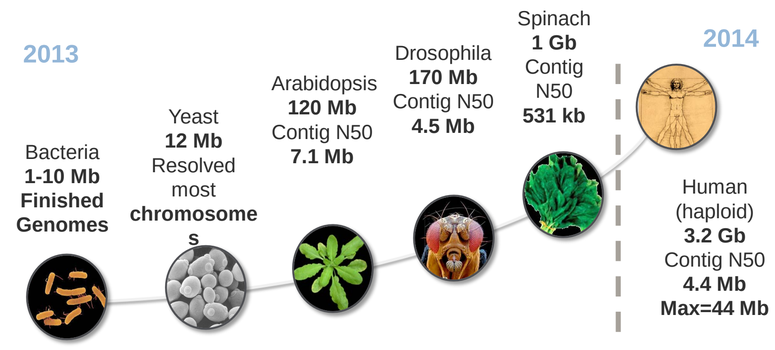


Figure 7. Schema of resolved adaptive gene-structure variation.

Conclusions

- PacBio RS II generates the most comprehensive, accurate and contiguous assemblies for plant genomes compared to existing reference genomes.
- P5-C3 chemistry and size selection of libraries help to capitalize on long read-length sequencing capabilities to generate high-quality genome assemblies.
- Complete BACs and targeted sequencing is possible for characterizing complex regions and full-length isoform surveys.



Public datasets, SMRT Analysis and compatible third party software are available from PacBio DevNet: **http://pacbiodevnet.com/**

References and Resources

[1] Chin CS., et. al. "Nonhybrid , finished microbial genome assemblies from long-read SMRT sequencing data." *Nat Methods*. Jun;10(6):563-9 (2013).

[2] Zhang W et al. (2014) PacBio sequencing of gene families – A case study with wheat gluten genes. *Gene* doi:10.1016/j.gene.2013.10.009

[3] Maron, LG et al. (2012) A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils. *PNAS* doi: 10.1073/pnas.1220766110

PAG 2014: Michael Schatz, "De novo assembly of complex genomes using single molecule sequencing"

PacBio Blog: *A. thaliana* genome assembly effort

PacBio Blog: PAG PacBio Workshop Showcases User's Large Genome Efforts

PacBio Webinar: Optimizing Eukaryotic De Novo Genome Assembly with Long-read Sequencing

BioRxiv: Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing

PacBio Applications: Plant and Animal Genomes

ACKNOWLEDGEMENTS

We thank Joe Ecker and Chongyuan Luo for the joint effort on the Arabidopsis genome and Allen van Deynze (UC Davis) for the joint collaboration to sequence the Spinach genome. We also thank the community for their continued interest in Pacific Biosciences.