Data Release for Polymorphic Genome Assembly Algorithm Development

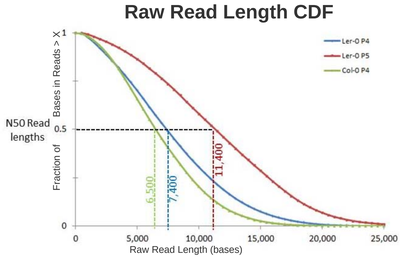
Marty R. Badgett[1], Jason Chin[1], David Rank[1], Paul Peluso[1,], Joseph Karaliuls[1], Maria Bernal[2], Ute Krämer[2], Sangtae Kim[3], Christa Lanz[3], Detlef Weigel[3], Chongyuan Luo[4], Ronan O'Malley[4], Joe Ecker[4]
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, [2]Dept. Of Plant Physiology, Ruhr University, 44801 Bochum, Germany, [3]Dept. of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany,[4]Salk Institute for Biological Studies, La Jolla, CA 92037

Abstract

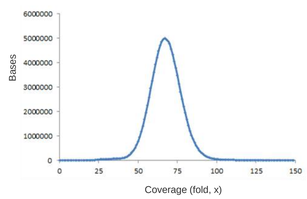
Heterozygous and highly polymorphic diploid (2n) and higher polyploidy (n > 2) genomes have proven to be very difficult to assemble. One key to the successful assembly and phasing of polymorphic genomics is the very long read length (9-40 kb) provided by the PacBio® RS II system. We recently released software and methods that facilitate the assembly and phasing of genomes with ploidy levels equal to or greater than 2n (presentation abstract by Jason Chin). In an effort to collaborate and spur on algorithm development for assembly and phasing of heterozygous polymorphic genomes, we have recently released sequencing datasets that can be used to test and develop highly polymorphic diploid and polyploidy assembly and phasing algorithms. These data sets include multiple species and ecotypes of *Arabidopsis* that can be combined to create synthetic in-silico F1 hybrids with varying levels of heterozygosity. Because the sequence of each individual line was generated independently, the data set provides a 'ground truth' answer for the expected results allowing the evaluation of assembly algorithms. The sequencing data, assembly of inbred and in-silico heterozygous samples (n=>2) and phasing statistics will be presented. The raw and processed data has been made available to aid other groups in the development of phasing and assembly algorithms.

Arabidopsis thaliana Ler-0 and Col-0 Data Sets

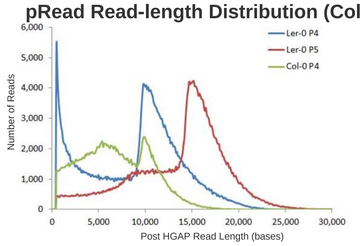
Raw Read Length CDF



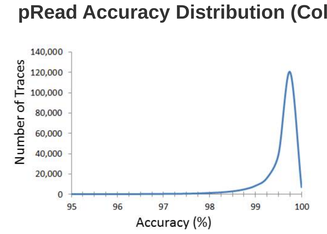
Typical Genome-coverage Distribution



pRead Read-length Distribution (Col-0)



pRead Accuracy Distribution (Col-0)



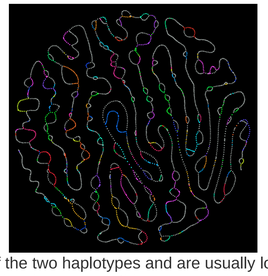
HGAP Assembly of *A. thaliana* inbreds

Strain	Inbred	Chemistry	Assembly Size	Contig N50	Max Contig
Col-0	yes	P4	119.7 Mb	6.2 Mb	10.2 Mb
Ler-0	yes	P4	120.8 Mb	5.9 Mb	13.3 Mb
Ler-0	yes	P5	124.6 Mb	6.2 Mb	13.0 Mb

Data Available:
More than 60 X raw PacBio data was collected and is available for algorithm development for both inbred homozygous and outbred heterozygous (by combining the data sets) assembly test data sets. The available data was error corrected and assembled[1]. Results shown above.

Example of *Arabidopsis thaliana* Ler-0 and Col-0 In-silico F1 Assembly

String Graph illustration of a portion of the Ler-0 x Col-0 in-silico F1 assembly



Bubbles represent separation of the two haplotypes and are usually locations of structural variation. Colored regions are phase-able by single nucleotide polymorphisms

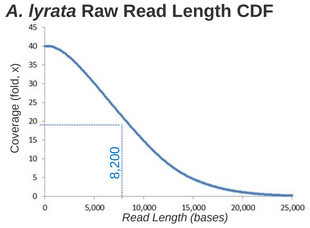
Falcon Assembly of *A. thaliana* in silico F1

Inbred	Chemistry	Assembly Size	Contig N50	Max Contig	Assembler
Col-0 x Ler-0	P4	130 Mb	6.1 Mb	10.8 Mb	Falcon Assembler
Col-0 x Ler-0	P4	220 Mb	110 kb	1.6 Mb	Celera® Assembler

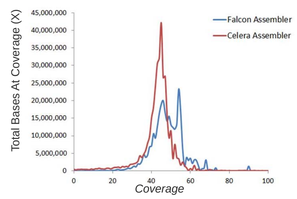
Data Used:
The Col-0 and Ler-0 sequences differ by ~ 0.5% identity with many structural variations[1]. 60X raw Col-0 and 60X raw Ler-0 P4 sequence data were mixed and co-assembled using the Falcon Assembler and the Celera Assembler. The assembly statistics for the Falcon Assembler are very similar to the single inbred genomes assembled independently, indicating the diploid heterozygous assembly is of similar quality to the inbred (haploid-like) assemblies of the single inbreds[2]. The Celera assembly, separates the two haplotypes into a ~2X larger and more fragmented assembly.

Arabidopsis lyrata Data Set

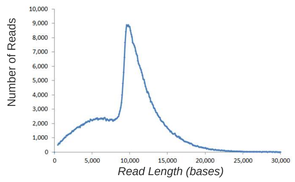
A. lyrata Raw Read Length CDF



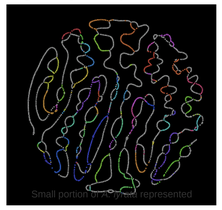
A. lyrata Diploid Collapse for Phasing



A. lyrata HGAP-Corrected pRead Length Distribution



String Graph Representation



Assemblies of *Arabidopsis lyrata*

Strain	Inbred	Chemistry	Assembly Size	Contig N50	Max Contig	Assembler
A. lyrata ssp. *lyrata* (MN47)[3]	**yes**	Sanger	183[A]/207[B] Mb	225 kb	1.2 Mb	Arachne
A. lyrata ssp. *petraea*	no	PB P5-C3	263[A] Mb	546 kb	8.9 Mb	Falcon
A. lyrata ssp. *petraea*	no	PB P5-C3	353[A] Mb	252 kb	1.7 Mb	Celera Assembler

[A] Total Length of Contigs [B] Total Length of Scaffolds

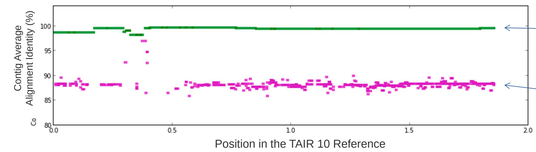
Data Available:
More than 40X raw PacBio data was collected and is available for algorithm development for outbred heterozygous or in-silico tetraploid (when mixed with *A. thaliana* data sets(s)) assembly test data sets. The available data was error corrected and assembled as a heterozygous diploid genome. Raw, error-corrected and assembly data metrics and results are presented above.

Example of *A. thaliana* Col-0 x *A. lyrata* In-silico Allo-tetraploid Assembly

Falcon Assembly of *A. thaliana* (Col-0) x *A. lyrata* in-silico tetraploid

Portion of assembly		Assembly Size	Contig N50	Max Contig
Col-0 x *A. lyrata*	(full in-silico tetraploid)	*Segregated into two parental sources below*		
Col-0 subset	(inbred diploid subset)	121 Mb	4.9 Mb	12.1 Mb
A. lyrata subset	(outbred diploid subset)	303 Mb	341 kb	5.1 Mb

Co-Assembly of Col-0 and *A. lyrata* Aligned to the TAIR-10 (Col-0) Reference



Data Used:
A. thaliana Col-0 and *A. lyrata spp. lyrata* genomes differ by ~ 13% identity with many structural variations[3]. 40X of raw Col-0 and 40X of raw *A. lyrata* sequence data were mixed and co-assembled using the Falcon Assembler. The co-assembly statistics of the in-silico tetraploid are very similar to the single inbred Col-0 and the outbred *A. lyrata* assemblies done independently indicating the tetraploid assembly is of similar quality to the assemblies of the inbred Col-0 and outbred *A. lyrata* when done independently[2].

Haploid, Diploid and Tetraploid Assembly Combinations Available

Strain	Ploidy/polymorphic	Reference
col-0	Inbred appears haploid	TAIR 10 Reference
ler-0	Inbred appears haploid	PBI Reference**
A. lyrata	Outbred - diploid	PBI references**
ler-0 x col-0	Diploid insilico	haplotype references available (above)
A. lyrata x col-0	Allotetraploid	haplotype references available (above)
A. lyrata x col-0+ler0	Outbred tetraploid	haplotype references available (above)

Data Availability

Arabidopsis thaliana Ler-0
P5 Chemistry raw data + Assembly
P4 Chemistry raw data + Assembly
Arabidopsis lyrata spp. lyrata
P4 Chemistry raw data + Assembly

https://github.com/PacificBiosciences/DevNet/wiki/Datasets

Arabidopsis thaliana Col-0
P4 Chemistry raw data + Assembly

Contact: Joe Ecker/Chongyuan Luo
Joseph Ecker <ecker@salk.edu>
Chongyuan Luo <cluo@salk.edu>

Conclusions

We encourage algorithm developers to download the raw data and polished assembly data sets for use in improvement, evaluation and development of haploid, and polymorphic diploid and tetraploid assemblies and algorithms.

Build your own string graph

References

[1]*Luo/Ecker in preparation (2014)*
[2]*Chin J. "Assembly and Phasing of Polymorphic Heterozygous Diploid Genomes" Presentation SFAF 2014 Santa Fe NM*
[3]*Hu et al. "The Arabidopsis lyrata genome sequence and the basis of rapid genome size change." Nature Genetics **43**, 476-481 (2011)*