

Abstract

Significant advances in bioinformatics tool development have been made to more efficiently leverage and deliver high-quality genome assemblies with PacBio® long-read data. Current data throughput of SMRT® Sequencing delivers average read lengths ranging from 10-15 kb with the longest reads exceeding 40 kb. This has resulted in consistent demonstration of a minimum 10-fold improvement in genome assemblies with contig N50 in the megabase range compared to assemblies generated using only short-read technologies.

This poster highlights recent advances and resources available for advanced bioinformaticians and developers interested in the current state-of-the-art large genome solutions available as open-source code from PacBio and third-party solutions, including HGAP, MHAP, and ECTools. Resources and tools available on GitHub are reviewed, as well as datasets representing major model research organisms made publically available for community evaluation or interested developers.

For more information, see pacb.com/pag

PacBio-only *de novo* Assembly

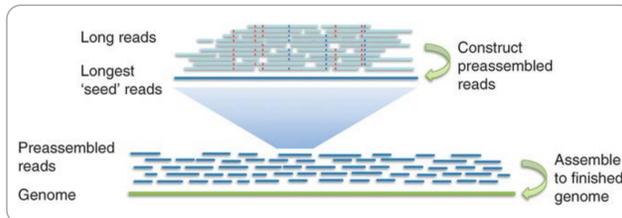


Figure 3. Overview of the Hierarchical Genome Assembly approach.

HGAP.3: Hierarchical Genome Assembly Process

As a fully-integrated protocol in SMRT Analysis, the maximum supported genome assembly is 130 MB. However, larger genomes can be assembled using the makefile-based workflow, **smrtmake**.

github.com/PacificBiosciences/smrtmake

github.com/PacificBiosciences/smrtmake#hgap3

FALCON is an experimental hierarchical genome assembly tool kit. Utilizes a forked version of Gene Meyer's **DALIGNER** for pre-assembly, and a string graph-based contig assembly. Haploid or diploid genomes up to ~100 MB can be assembled.

github.com/PacificBiosciences/falcon

github.com/thegenemyers/DALIGNER

MHAP: MinHash Alignment Process

Uses a probabilistic algorithm for detecting overlaps in long reads, resulting in drastic speedups in pre-assembly. Has been validated on genomes up to 3 GB.

github.com/marbl/MHAP

PBcR self-correction: A mode within PBcR (aka pacBioToCA) to do self-correction in the same style as HGAP. Celera® Assembler 8.2 uses the MHAP algorithm for faster overlap calculation during the self-correction phase.

wgs-assembler.sf.net/wiki/index.php/PBcR

Successful PacBio-Only *De Novo* Genomes

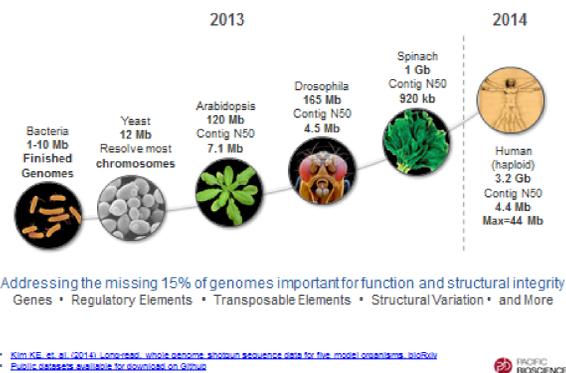


Figure 1. Advances in PacBio-only *de novo* genome assembly have reached the range of many plant and animal genomes.

Overview of Assembly Approaches

PacBio SMRT Sequencing has proven useful for generating high-quality genome assemblies without an existing reference (*de novo*), and improving the contiguity and accuracy of existing reference genomes. Several common assembly approaches are summarized here:

- 1. PacBio-only *de novo* assembly.** Using just PacBio reads from a long-insert library, the reads are often preprocessed before being assembled using an Overlap-Layout-Consensus algorithm. The best known implementation of this is HGAP.
- 2. Hybrid *de novo* assembly.** Using a combination of PacBio and short-read data, the reads are used together during assembly to generate a hybrid assembly.
- 3. Gap-filling.** Starting with an existing mate-pair based assembly, the internal gaps (consisting of Ns) inside the scaffolds are filled using PacBio sequences.
- 4. Scaffolding.** Using an existing assembly (such as an assembly based on short-read data), PacBio reads are used to join contigs.

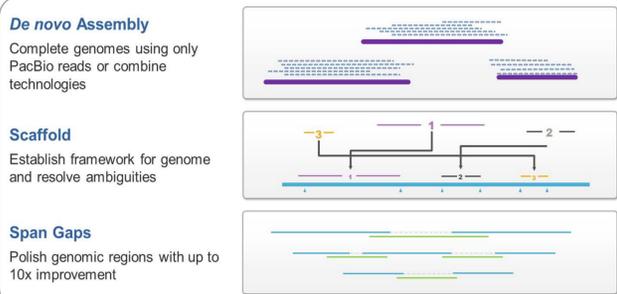
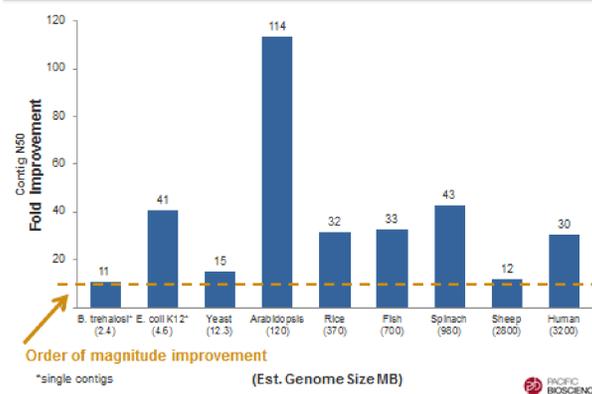


Figure 2. Illustration of genome assembly approaches leveraging PacBio data.

Summary: Comparison PacBio vs. 2nd Gen Draft Assemblies



Hybrid *de novo* Assembly

ECTools: A long-read correction algorithm utilizing unitigs generated from a short-read assembly

github.com/jgurtowski/ectools

schatzlab.cshl.edu/presentations/2013-06-18.PBUserMeeting.pdf

PBcR self-correction :

Assembly pipeline integrated with Celera Assembler which error corrects PacBio CLR reads with higher accuracy short reads, typically either PacBio CCS or other technologies. Also known as **PacBioToCA**.

wgs-assembler.sf.net/wiki/index.php/PBcR

Gap-Filling

PBJelly 2: Demonstrated improvements in existing reference genomes >1 GB by filling gaps in scaffolds with PacBio reads.

sf.net/projects/pb-jelly/

Scaffolding

PBJelly 2: The new version of PBJelly has support for joining scaffolds.

sf.net/projects/pb-jelly/

Additional Resources

Expectations for PacBio® Genome Assemblies from the Customer Perspective

Figure 4. Customer perspective of PacBio-only genome assemblies.

PacBio Devnet

The primary starting point for accessing many of the tools and resources described here, including software, documentation, and datasets.

pacbiodevnet.com

GitHub

PacBio's GitHub repository houses source code for analysis tools and documentation created by our developers, and much more.

github.com/PacificBiosciences

PacBio bioinformaticians have curated a training wiki detailing many analysis techniques and tips on experimental design.

github.com/PacificBiosciences/Bioinformatics-Training

Publicly-available datasets can be accessed for interested parties from a central point.

github.com/PacificBiosciences/DevNet/wiki/Datasets

References

- Berlin K, et al. (2014) Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing. bioRxiv doi: 10.1101/008003
- Chin, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods, 10, 563-569.
- English AC, et al. (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS ONE 7(11): e47768. doi:10.1371/journal.pone.0047768
- Lee H, et al. (2014) Error correction and assembly complexity of single molecule sequencing reads. bioRxiv doi: 10.1101/006395

Acknowledgements

The author would like to thank everyone who helped generate data for the poster.

