A Comparison of Assemblers and Strategies for Complex Large-Genome Sequencing with PacBio® Long Reads

Jenny Gu[1], Richard Hall[1], Cheryl Heiner[1], Brian Sogoloff[2], James Meldrim[2], Kristen Connolly[2], Terrance Shea[2], Carsten Russ[2], Christina Cuomo[2], Les J. Szabo[3]

[1]Pacific Biosciences, Menlo Park, CA; [2]Broad Institute, Cambridge, MA; [3]USDA, Agricultural Research Service

Introduction

PacBio® sequencing holds promise for addressing large-genome complexities, such as long, highly repetitive, low complexity regions and duplication events that are difficult to resolve with short-read technologies. Several strategies, with varying outcomes, are available for *de novo* sequencing and assembling of larger genomes. Using a dikaryotic fungal genome, estimated to be ~80 Mb in size, as the basis dataset for comparison, we highlight assembly options when using only PacBio sequencing or a combined strategy leveraging data sets from multiple sequencing technologies. Comparisons of results generated from different assemblers available for large-genome assembly using data generated from SMRT® Sequencing will be shown. These include results generated with HGAP, Celera® Assembler, MIRA, PBJelly, and other assembly tools currently in development. Improvements observed include a near 50% reduction in the number of contigs coupled with a doubling of contig N50 size, at the minimum, in genome assemblies incorporating SMRT Sequencing data. We further show how incorporating long reads also highlights new challenges and missed insights of short-read assemblies arising from heterozygosity inherent in multiploid genomes.

Methods

PacBio® RS II Sequencing Chemistries Provide Long Read Lengths over 10 Kb

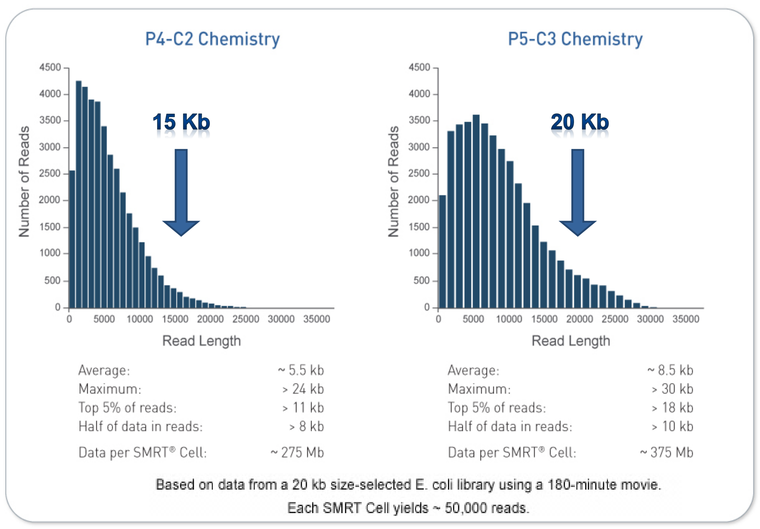
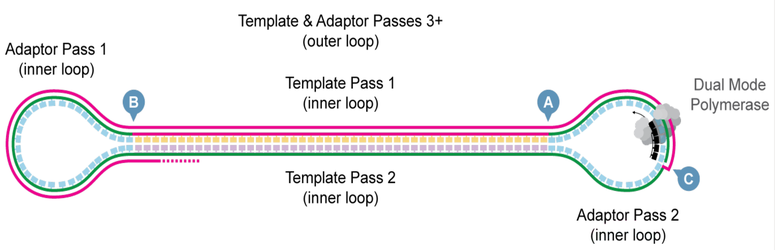


Figure 1. Example read length distribution from a SMRT® sequencing run with 20 kb size-selected *E. coli* library using a 180 min movie. Average throughput of 300 MB per SMRT cell with ~50,000 reads.

Universal SMRTbell™ Template



Continuous Long Reads (CLR)

- Insert size: 2 kb to > 10 kb
- Movie collection time: 120 min

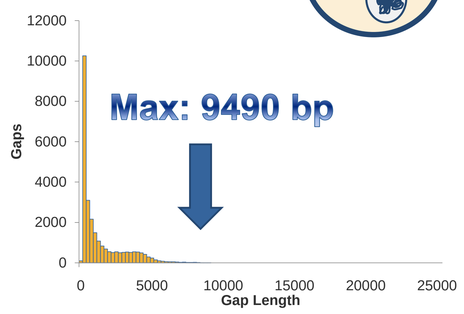
A: Template Pass 1 = Strand-Displacement Mode
B: Adaptor Pass 1 = Single-Stranded Mode
Template Pass 2 = Single-Stranded Mode
C: Adaptor Pass 2 = Single-Stranded (some ds) Mode
Template & Adaptor Passes = Strand-Displacement Mode

Figure 2. Schematic of SMRTbell sequencing to generate Continuous Long Reads > 10kb.

Dikaryotic Fungal Genome (~80 MB)

Profile of Gaps in Draft Assembly:

Gap Stats	
#Seqs	19649
Min	25
1st Qu.	377
Median	970
Mean	1647
3rd Qu.	2731
Max	9490
Total	32372470
n50	3305
n90	810
n95	506



Sequencing results dependent on quality of DNA library:

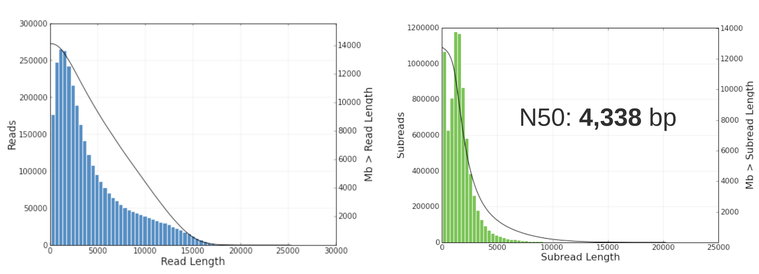


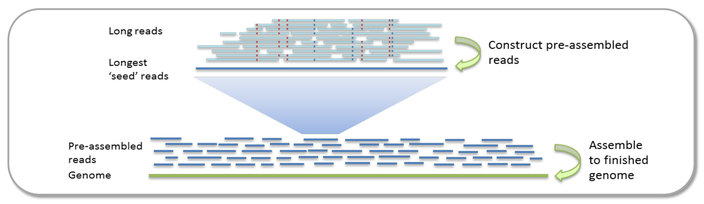
Figure 3. (Top) Distribution of gaps in draft assembly of fungal genome. (Bottom) Sequenced read length distribution for dikaryotic fungal genome library using the P4-C2 chemistry with 10 kb size selection and a data recording time of 180 min movie. The results of the filtered subreads used for genome assembly is impacted by the quality of the starting library.

SMRT Analysis and compatible third party software is available from PacBio DevNet: **http://pacbiodevnet.com/**

Bioinformatics Strategies for Large Genome Assembly (*De novo* vs. Gap Filling)

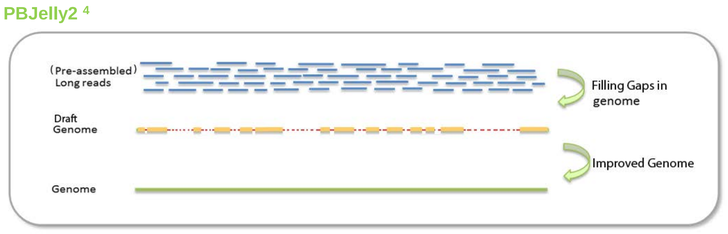
***De Novo* Genome Assembly (PacBio reads Only):**

HGAP[1], Celera Assembler[2], MIRA Assembler[3]



Gap Filling for existing draft genomes:

PBJelly2 [4]



Impact of Read Coverage on Gap Filling with PBJelly 2 Assembly:

	Draft Genome	1x	2.5x	5x	10x	20x	Preassembled Reads (100x)
# Scaffolds	3538	0	3425	3374	3303	3228	3196
# Contigs	22184	0	19447	17756	15629	15564	12825
# Gaps	19649	0	16976	15330	13284	13243	10587
% Gap Reduction	-	-	**14%**	**22%**	**32%**	**33%**	**46%**
Total Scaffolds (MB)	114.10	0	114.83	116.12	117.55	119.82	117.67
Total Contigs (MB)	78.88	0	84.73	88.61	92.42	95.61	93.82
Total Gaps (MB)	32.37	0	26.84	23.95	20.90	19.69	18.84
% Gap Reduction	-	-	**17%**	**26%**	**35%**	**39%**	**42%**

0 = PBJelly2 software failure

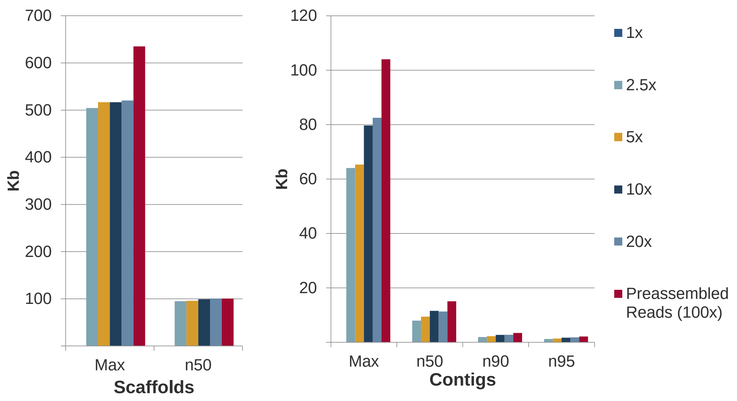


Figure 4: (Top) Metrics table of genome improvements with increasing PacBio read coverage. (Bottom) Improvements observed in scaffolds and contigs when using PBJelly2 for gap filling and interscaffolding to improve fungal draft genome.

Results

Comparison of Assembly Results:

Data	Assembly Coverage	Assembly Size (Mb)	Total Contig Length (Mb)	Scaffolds	Scaffold N50 (Kb)	Contigs	Contig N50 (Kb)	Contig Max (Kb)
Draft Reference Assembly (AllPaths)								
Illumina® Data	100X	114.07	81.63	3,538	93.38	30,122	4.5	60.89
HGAP v.2								
PacBio® Reads	6.5x	168.5	168.5	-	-	14,882	12.6	97.186
Celera Assembler (HGAP – Custom Analysis & Quiver)								
PacBio Reads	49.6x	167.47	167.47	-	-	8,182	23.7	140.23
MIRA Assembler								
PacBio Reads	6.46x	152.69	152.69	-	-	21,357	8.83	39.21
PBJelly 2								
Illumina Data + PacBio Reads	-	117.7	93.8	3,196	100.35	12,825	15.0	104.0

Blast QC results: (Total Predicted Reference Fungal Gene Set: 18,790; BLAST threshold: e= $10e^{-04}$

Unique BLAST hits	No Hits	Length (Mean)	Mismatch (Mean)	Alignment Gap (Mean)	Length (Max)	Mismatch (Max)	Alignment Gap (Max)
Draft Reference Assembly (AllPaths)							
18,790	94	876.45	33.47	1.84	12742	1190	84
HGAP v.2							
18,734	26	806.65	41.98	3.00	12189	1217	79
Celera Assembler (HGAP – Custom Analysis & Quiver)							
18,589	111	738.68	41.20	3.916	8669	925	111
MIRA Assembler							
18,686	215	786.50	41.573	4.08	13871	1320	129
PBJelly 2							
18,784	55	1034.89	39.15	2.160	16646	1185	84

Separation of the two haplotypes?

See also Poster P1057 - Diploid Genome Assembly

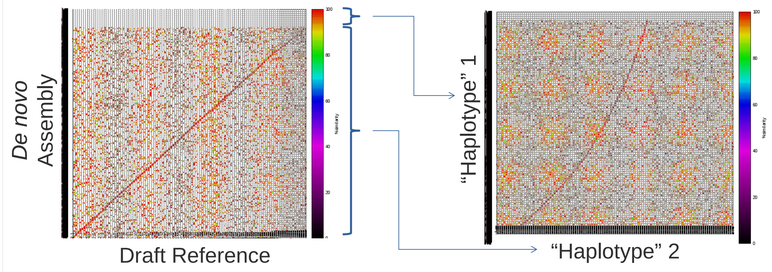
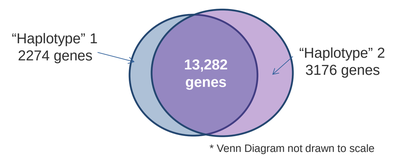


Figure 5:
Dotplots were used for initial assessment of *de novo* genome assembly when compared to the draft reference. Possible detection of the two putative haplotypes in this genome is suggested using this reference-based strategy.



Conclusion

Advances in SMRT® sequencing on the PacBio® RS II allow for improved read lengths and throughput that yields more than 50% of the data in read lengths greater than 10 kb. However, obtaining these long reads is very much dependent on the input library quality, which also greatly impacts the outcomes of downstream analysis. Several bioinformatics algorithms are available to leverage the benefits of long reads to improve genome assemblies when appropriate coverage and read lengths are obtained to address the research challenge. The available tools enable researchers several options when pursuing either a *de novo* or gap-filling strategy to improve genomes of interest. Comparing the assembly results, challenges potentially introduced by heterozygosity originating from the different haplotypes become increasingly evident when using long reads to overcome limitations of short-read technologies. Alternatively, long reads may potentially provide a new avenue to dissect genomes between the constitutive haplotypes and understand the contributions of each respectively when appropriate tools are developed to conduct such analyses.

References

[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] Myers EW, *et al.* "A Whole-Genome Assembly of Drosophila." *Science*, 287(5461):p. 2196-2204 (2000).

[3] Chevreux, B. *et al.* "Genome Sequencing Assembly Using Trace Signals and Additional Sequence Information." *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, 99, pp.45-56 (1999).

[4] English AC, *et al.* "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology." *PLoS ONE*, 7(11):e47768 (2012).

Acknowledgements

The authors would like to thank the collaborators and others at PacBio® for their contributions to this poster.

Secondary analysis source code is available from PacBio DevNet (http:// http://www.smrtcommunity.com/).