Automated, Non-Hybrid *De Novo* Genome Assemblies and Epigenomes of Bacterial Pathogens

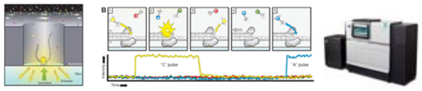
Tyson A. Clark, Khai Luong, Jason Chin, Matthew Boitano, Stephen W. Turner, and Jonas Korlach
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Abstract

Understanding the genetic basis of infectious diseases is critical to enacting effective treatments, and several large-scale sequencing initiatives are underway to collect this information[1]. Sequencing bacterial samples is typically performed by mapping sequence reads against genomes of known reference strains. While such resequencing informs on the spectrum of single-nucleotide differences relative to the chosen reference, it can miss numerous other forms of variation known to influence pathogenicity: structural variations (duplications, inversions), acquisition of mobile elements (phages, plasmids), homonucleotide length variation causing phase variation, and epigenetic marks (methylation, phosphorothioation) that influence gene expression to switch bacteria from non-pathogenic to pathogenic states[2]. Therefore, sequencing methods which provide complete, *de novo* genome assemblies and epigenomes are necessary to fully characterize infectious disease agents in an unbiased, hypothesis-free manner.

Hybrid assembly methods have been described that combine long sequence reads from SMRT® DNA Sequencing with short reads (SMRT CCS (circular consensus) or second-generation reads), wherein the short reads are used to error-correct the long reads which are then used for assembly. We have developed a new paradigm for microbial *de novo* assemblies in which SMRT sequencing reads from a single long insert library are used exclusively to close the genome through a hierarchical genome assembly process, thereby obviating the need for a second sample preparation, sequencing run, and data set. We have applied this method to achieve closed *de novo* genomes with accuracies exceeding QV50 (>99.999%) for numerous disease outbreak samples, including *E. coli*, *Salmonella*, *Campylobacter*, *Listeria*, *Neisseria*, and *H. pylori*. The kinetic information from the same SMRT Sequencing reads is utilized to determine epigenomes. Approximately 70% of all methyltransferase specificities we have determined to date represent previously unknown bacterial epigenetic signatures. With relatively short sequencing run times and automated analysis pipelines, it is possible to go from an unknown DNA sample to its complete *de novo* genome and epigenome in about a day.

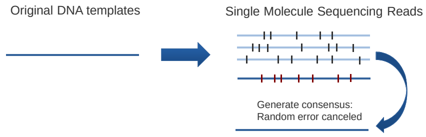
SMRT® Sequencing



Zero-Mode Waveguide (ZMW)

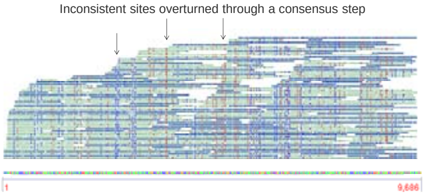
Using fluorescence to identify real-time nucleotide incorporations

Errors Are Random in SMRT® Sequencing, Not Correlated with Real Variants



Example: Generate a Highly Accurate Seed Read through Consensus

Inconsistent sites overturned through a consensus step

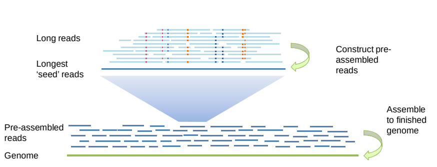


1. Start with 9.7 kb seed read
2. Align other reads to the seed read for construct mini-assembly
3. Construct accurate pre-assembled consensus sequence

- Utilizes every bit of data:
  - Longest reads for continuity
  - Shorter reads used for consensus accuracy
- Sequence Identity to the reference: 85.7% (seed read) ➔ 99.3% (pre-assembled long read), 9089 bp
- Chimera / low quality regions can be filtered out early
- Accurate long consensus reads easier to assemble

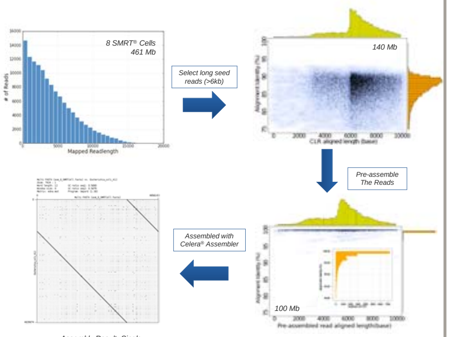
Hierarchical Genome Assembly Process (HGAP)

Overview



Bacterial Genome Assembly with HGAP

Finished genomes with >99.999% accuracy from long PacBio® reads



Assembly Result: Single contig spans the whole chromosome

Most of the PLRs are above 99.95% accuracy using only PacBio long reads.

***Escherichia coli* (K12 MG1655) Assembly Results**

SMRT® Cells	CLR bases	CLR Cov.	Seed read Cov.	PLR Cov.	PLR nReads	PLR mean read length	Assembly size	# of contigs >10 kb (all)	Genome covered	N50	Concordance with Sanger reference	QV	% full-length matched ORF predicted
8	460M	99.4	30.2	21.5	17,232	5,777	4.66M	1(2)	100.3%	4.65M	99.99951%	53.1	99.8%
6	340M	73.4	22.6	15.7	13,090	5,566	4.70M	10(14)	101.3%	1.16M	99.99938%	52.1	100.0%
4	232M	50.0	14.9	10.1	8,610	5,422	4.69M	17(21)	101.1%	0.39M	99.99876%	49.1	98.8%

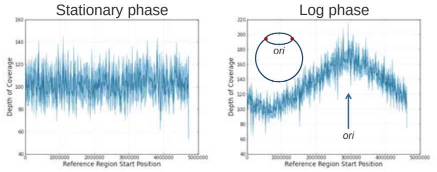
- High concordance (>QV50) of *de novo* assembly with reference
- 21.1X pre-assembled long reads (PLRs) with average length of 5.7 kb resolve all ~5.5 kb rRNA repeats to give a single contig assembly
- We reach 99.8% ORF prediction concordance

Genome Assembly Summary

Requirements for Achieving High-Quality Finished Genomes:

1. High-consensus accuracy
   - Lack of systematic bias
2. Long sequence reads to resolve repeats
3. Lack of sequence context bias
   - GC content
   - Low complexity sequence

Shotgun Sequencing Coverage Across Assembly:



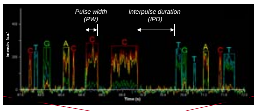
References

[1] *e.g.*, the 100K Foodborne Pathogen Genome Project (www.100kgenome.vetmed.ucdavis.edu/)
[2] Srikhanta et al. (2010) *Nat Rev Microbiol* 8: 196-206.
[3] Flusberg et al. (2010) *Nat Methods* 7: 461-465.
[4] Murray et al. (2012) *Nucleic Acids Reseach* 40: 11450-62.

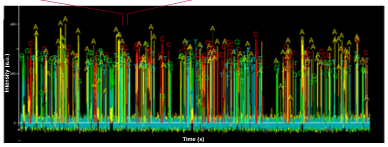
Acknowledgements

The authors would like to thank the Joint Genome Institute for DNA samples and discussions on these analyses and results.

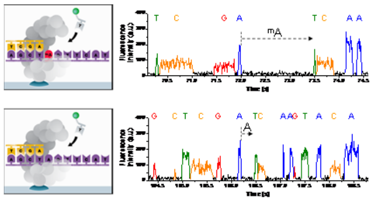
Detection of Base Modifications with SMRT® Sequencing



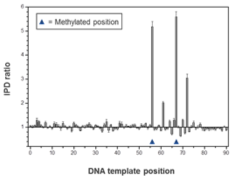
- Pulse width (PW) = duration of fluorescence pulse
- Interpulse duration (IPD) = time between successive pulses

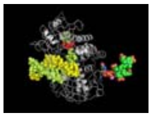


Effects of Base Modifications on Polymerase Kinetics



IPD is Increased Before T Incorporation Across from N6-methyladenine (mA)

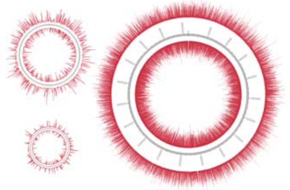




Polymerase has contacts with the DNA modification before and after incorporation of the base opposite the modification

Methylome Analysis

Kinetic Variation Across a Bacterial Genome

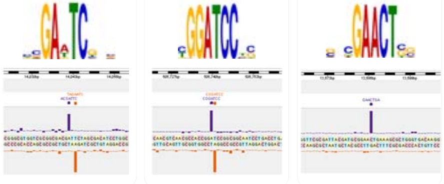


Methyltransferase Motif Identification

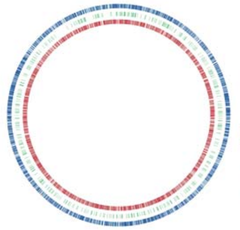
Motif #1: 6mA
5′-GANTC-3′
3′-CTNAG-5′

Motif #2: 6mA
5′-GGATCC-3′
3′-CCTAGG-5′

Motif #3: 4mC
5′-GAACT-3′
3′-CTTGA-5′



Methylome Summary



Motif	# in genome	# modified	% modified
5'-GANTC-3' 3'-CTNAG-5'	9,036	8,559	94.7%
5'-GGATCC-3' 3'-CCTAGG-5'	1,022	1,020	99.8%
5'-GAACT-3' 3'-CTTGA-5'	4,605	4,267	92.7%