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

Chen-Shan Chin, David H. Alexander, Patrick Marks, Aaron Klammer, James Drake, Cheryl Heiner, Tyson A. Clark, Khai Luong, Matthew Boitano, Stephen W. Turner & 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, high-accuracy reads (SMRT (circular consensus sequencing) CCS or second-generation reads) to generate long, highly accurate reads that are then used for assembly. We have developed a new paradigm for microbial *de novo* assemblies in which long SMRT sequencing reads (average readlengths >5,000 bases) 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%) to 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. The process has been automated and requires less than 1 day from an unknown DNA sample to its complete *de novo* genome and epigenome.

SMRT® Sequencing

PacBio *RS* II

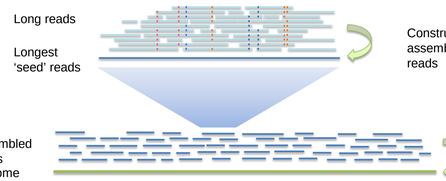
- Double the throughput of the previous model, the PacBio *RS*
- Industry's highest consensus accuracy and longest read lengths

Requirements for 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

Hierarchical Genome Assembly Process (HGAP)

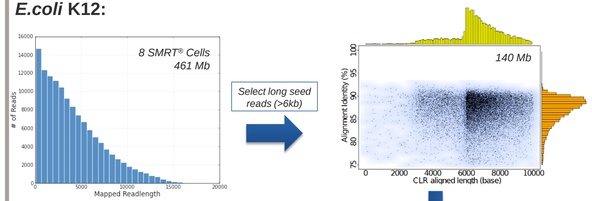
Overview



Bacterial Genome Assembly with HGAP

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

***E.coli* K12:**



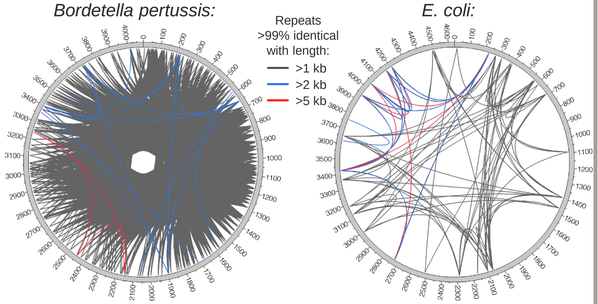
Assembly result: single contig spans the whole chromosome 99.99951% concordance with reference

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

- High concordance (>QV50) of *de novo* assembly with reference
- 99.8% ORF prediction concordance

Application: Whooping Cough

The Pertussis Genome is Very Repetitive

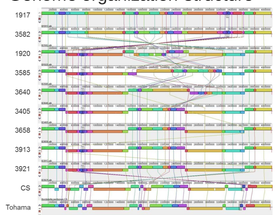


Finished Pertussis Genomes

Year	Strain	Sequencing	Genome size	Reference
2003	Tohama	Sanger: • 87,500 paired-end reads (1-4kb shotgun libraries) • 2,560 paired-end reads (10-20kb pBAC library) • 41,700 sequencing reads during finishing	4,086,186 bp	Parkhill et al. Nature Genetics 35: 32-40
2011	CS	454 & Sanger: • 329,480 454 reads yielding 287 contigs • 11,444 paired-end ABI3730 reads • Filled gaps through sequencing of PCR products	4,124,236 bp	Zhang et al. J Bacteriology 193: 4017-4018
2013	B1917	6 SMRT Cells	4,102,176 bp	this study*
2013	B1920	8 SMRT Cells	4,114,613 bp	this study*
2013	B3405	6 SMRT Cells	4,109,986 bp	this study*
2013	B3582	8 SMRT Cells	4,104,315 bp	this study*
2013	B3585	8 SMRT Cells	4,106,397 bp	this study*
2013	B3640	8 SMRT Cells	4,110,999 bp	this study*
2013	B3658	6 SMRT Cells	4,103,245 bp	this study*
2013	B3913	6 SMRT Cells	4,109,515 bp	this study*
2013	B3921	4 SMRT Cells	4,111,519 bp	this study*

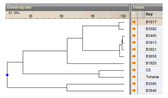
Comparative Genomics

Genome organization structure



Virulence genes

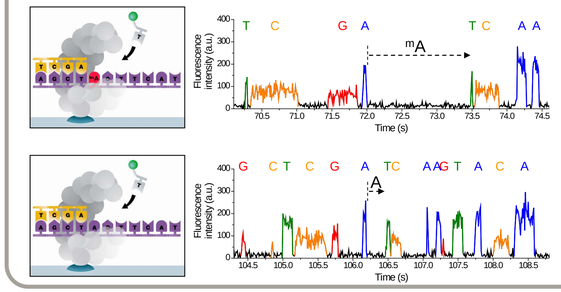
Phylogeny:



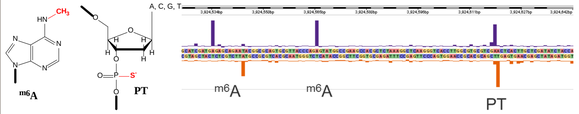
Collaboration with A. Zeddeman, H. van der Heide, M. Bart & F. Mooi
National Institute for Public Health and the Environment (RIVM), Netherlands

Epigenome Analysis

Base Modifications and Polymerase Kinetics[3]



Example: *Salmonella* Epigenomes



Methyltransferase Specificity	Modified Base	S. Bareilly [SA122881]	S. Heidelberg [SL10_04]	S. Heidelberg [2969]	S. Typhimurium [2048]	S. Javiana [1992_73]	S. Cubana [2050]	S. Sa.Paul SF3	S. Sa.Paul SF48
5′–GATC–3′ 3′–CTAG–5′	m6A	■	■	■	■	■	■	■	■
5′–CAGAG–3′ 3′–GTCTC–5′	m6A	■	■	■	■	■	■	■	■
5′–ATGCAT–3′ 3′–TACGTA–5′	m6A	■	■		■	■	■		
5′–CACTG–3′ 3′–GTGAC–5′	m4C	■		■	■	■	■	■	■
5′–ACGANGC–3′ 3′–TGCTNCG–5′	m6A		■	■					
5′–CCGANNNNNGTC–3′ 3′–GGCTNNNNNCAG–5′	m6A	■							
5′–GANNNNNNNTAYG–3′ 3′–CTCNNNNNNYATRC–5′	m6A		■	■				■	■
5′–GRNTAYNNNNNRTCG–3′ 3′–CNNATRNNNNNYAGC–5′	m6A					■			
5′–RAACNNNNNTGA–3′ 3′–YTTGNNNNNACT–5′	m6A						■		
5′–GGANNNNNNATTA–3′ 3′–CCTNNNNNNTAAT–5′	m6A						■		
5′–C_{PT}AAC–3′ 3′–CTT_{PT}G–5′	PT							■	■

Collaboration with M. Allard, E. Brown, E. Strain, M. Hoffman, T. Muruvanda, S. Musser (FDA), R. Roberts (NEB), B. Weimer (UC Davis)

■ complete modification
■ partial modification

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.

Acknowledgments

We would like to thank the Joint Genome Institute.