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 enabling effective treatments, and several large-scale sequencing initiatives are underway to collect this information.  Sequencing bacterial samples is typically performed by mapping sequence reads against genomes of known reference strains.  While such spurious alignments inform 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 (e.g., duplications, inversions), acquisition of mobile elements (e.g., plasmids, transposons), homonucleotide length variation causing phase variation, and epigenetic marks (methylation, phosphorothioation) that influence gene expression to switch bacteria from non-pathogenic to pathogenic states. 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.

Bacterial Genome Assembly with HGAP

Finished genomes with >99.999% accuracy from long PacBio® reads E. coli K12:

<table>
<thead>
<tr>
<th>Year</th>
<th>Strain</th>
<th>Sequencing Genome size</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>Tohama</td>
<td>4,124,236 bp</td>
<td>Zhang et al. (2003)</td>
</tr>
<tr>
<td>2011</td>
<td>CS</td>
<td>4,113,518 bp</td>
<td>Zhang et al. (2011)</td>
</tr>
<tr>
<td>2013</td>
<td>100K</td>
<td>4,114,613 bp</td>
<td>Zhang et al. (2013)</td>
</tr>
</tbody>
</table>

Epigenome Analysis

References


Acknowledgments

We would like to thank the Joint Genome Institute.

Hierarchical Genome Assembly Process (HGAP)

Overview

- Long reads
- Longest read reads
- Construct pre-assembled maps
- Assemble to finished genome

Application: Whooping Cough

The Pertussis Genome is Very Repetitive

Bordetella pertussis

E. coli

Comparative Genomics

Genome organization structure

Virulence genes

Example: Salmonella Epigenomes

Base Modifications and Polymerase Kinetics

SMRT® Sequencing

- Double the throughput of the previous model, the PacBio® RS II
- Industry’s highest consensus accuracy and longest read lengths

Requirements for finished genomes

1. High-consensus accuracy
2. Lack of systematic bias
3. Long sequence reads to resolve repeats
4. GC content
5. Low complexity sequence context bias
6. High concordance (>95%) of de novo assembly with reference
7. 99.9% ORF prediction concordance

References

- e.g., the 100K Foodborne Pathogen Genome Project (www.100kgenomes.vetmed.ucdavis.edu)