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. 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 variation (duplications, inversions), acquisition of mobile elements (plasmids, prophages), homonucleotide length variation causing phase variation, and epigenetic marks (methylation, phosphorylation) 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. Hybrid assembly methods have been described that combine long sequence reads from SMRT® DNA Sequencing with short 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.

Hierarchical Genome Assembly Process (HGAP)

Bacterial Genome Assembly with HGAP

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

Methylozone Analysis

Kinetic Variation Across a Bacterial Genome

References


Acknowledgements

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

Pacific Biosciences, PacBio, SMRT, SMRTbell and the Pacific Biosciences logo are trademarks of Pacific Biosciences of California, Inc. All other trademarks are the property of their respective owners. © 2013 Pacific Biosciences of California, Inc. All rights reserved.