Automated finished microbial genomes and epigenomes to understand infectious diseases

Ralph Vogelsang
Pacific Biosciences, United States

Motivation and objectives

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 samplesis 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 states2 (Srikhanta *et al.*, 2010). 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.

Methods

Hybrid assembly methods have been described that combine long sequence reads from SMRT® DNA sequencing with short reads (SMRT CCS 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 long SMRT sequencing reads (average read lengths>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.

Results and discussion

We have applied this method to achieve finished 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 epi-genomes. Approximately 70% of all methyltransferase specificities we have determined to date represent previously unknown bacterial epigenetic signatures.

Conclusions

Our method allows for rapid and comprehensive elucidation of the genetic and epigenetic basis of infectious disease agents. The process has been automated and requiresless than 16 hours from an unknown DNA sample to its complete *de novo* genome and epigenome.

References

Srikhanta YN, Fox KL, Jennings MP (2010) The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat Rev Microbiol.* **8**(3), 196-206. doi: 10.1038/nrmicro2283

1 e.g., the 100K Foodborne Pathogen Genome Project (www.100kgenome.vetmed.ucdavis.edu/)