Predictive pathogen biology: Genome-based prediction of pathogenic potential and countermeasures targets

Debjit Ray, Joseph S. Schoeniger, Kelly Williams, Corey Hudson and **Christopher Polage**
Sandia National Laboratories, Canada
University of California Davis Medical Center, Canada

Horizontal gene transfer (HGT) and recombination leads to the emergence of bacterial antibiotic resistance and pathogenic traits. HGT events can be identified by comparing a large number of fully sequenced genomes across a species or genus, define the phylogenetic range of HGT, and find potential sources of new resistance genes. In-depth comparative phylogenomics can also identify subtle genome or plasmid structural changes or mutations associated with phenotypic changes. Comparative phylogenomics requires that accurately sequenced, complete and properly annotated genomes of the organism. Assembling closed genomes requires additional mate-pair reads or "long read" sequencing data to accompany short-read paired-end data. To bring down the cost and time required of producing assembled genomes and annotating genome features that inform drug resistance and pathogenicity, we are analyzing the performance for genome assembly of data from the Illumina NextSeq, which has faster throughput than the Illumina HiSeq (~one-two days versus ~one week), and shorter reads (150bp paired-end versus 300bp paired end) but higher capacity (150-400M reads per run versus ~5-15M) compared to the Illumina MiSeq. Bioinformatics improvements are also needed to make rapid, routine production of complete genomes a reality. Modern assemblers such as SPAdes 3.6.0 running on a standard Linux blade are capable in a few hours of converting mixes of reads from different library preps into high-quality assemblies with only a few gaps. Remaining breaks in scaffolds are generally due to repeats (e.g., rRNA genes) are addressed by our software for gap closure techniques, that avoid custom PCR or targeted sequencing. Our goal is to improve the understanding of emergence of pathogenesis using sequencing, comparative genomics, and machine learning analysis of ~1000 pathogen genomes. Machine learning algorithms will be used to digest the diverse features (change in virulence genes, recombination, horizontal gene transfer, patient diagnostics). Temporal data and evolutionary models can thus determine whether the origin of a particular isolate is likely to have been from the environment (could it have evolved from previous isolates). It can be useful for comparing differences in virulence along or across the tree. More intriguing, it can test whether there is a direction to virulence strength. This would open new avenues in the prediction of uncharacterized clinical bugs and multidrug resistance evolution and pathogen emergence.

e: debray@sandia.gov