Expanding Genomic Knowledge of Under-Sequenced Organisms Using Petabase-Scale Data Mining

Harold P. Hodgins, Briallen Lobb, Mackenzie Peck, and Andrew C. Doxey

Department of Biology, University of Waterloo, Waterloo Ontario

Seventy percent of previously sequenced species have only one representative genome available in the NCBI Genbank database. This limits our understanding of the pan-genome, evolution, and function of these organisms.

Using species-specific *k*-mer signatures, we can predict which of the 26,847,286 datasets in the Sequence Read Archive might contain genomic information for any under-sequenced organism. Using this data we have found that it might be possible to add one or more draft genomes (at 1X coverage or greater) for 29,450 (30%) of species which have only one Genbank dataset.

As a proof of principle, we used SRA datasets to expand the pan-genome for *Clostridium tarantellae*, which has only one representative genome available. Our analysis found several metagenomic datasets containing *C. tarantellae* which we used to assemble draft genomes for *C. tarantellae*, thus improving our understanding of its pan-genome and evolutionary history. Analysis of the associated metadata revealed an association with a fish host niche. This association was previously predicted but never demonstrated.

This work demonstrates the potential of large-scale data mining for expanding the genomic knowledge of under-sequenced organisms and for revealing novel functional and environmental associations.