Poster H-28

On the the fidelity of assembly, binning and gene calling of metagenomes using simulated datasets

Authors:
Konstantinos Mavromatis (*Joint Genome Institute*)
Natalia Ivanova (*Joint Genome Institute*)
Kerrie Barry (*Joint Genome Institute*)
Harris Shapiro (*Joint Genome Institute*)
Eugene Goltsman (*Joint Genome Institute*)
Alice McHardy (*IBM Thomas J. Watson Research Center.*)
Isidore Rigoutsos (*IBM Thomas J. Watson Research Center.*)
Asaf Salamov (*Joint Genome Institute*)
Frank Korzenieski (*Joint Genome Institute*)
Miriam Land (*Oak Ridge National Laboratory*)
Philip Hugenholtz (*Joint Genome Institute*)
Nikos C Kyrpides (*Joint Genome Institute*)

Short Abstract: We constructed three synthetic metagenomic datasets of increased complexity by combining reads from a selection of 113 isolate genome sequencing projects available through the Joint Genome Institute. The datasets were used to evaluate assembly, binning and gene calling methods used for metagenomic analysis

Long Abstract:
In an efford to evaluate methods used to analyse metagenomes, we constructed three synthetic metagenomic datasets of increased complexity by combining reads from a selection of 113 isolate genome sequencing projects available through the Joint Genome Institute. Isolate genomes were selected to represent populations in real metagenomic datasets based on similar patterns of genome size, GC content and relative phylogenetic position. Reads were randomly sampled from the selected genomes to match the read depth of their corresponding populations in the metagenomic assemblies. Sampled reads were then assembled using three programs used to assemble the real metagenomic data (Phrap, Arachne and JAZZ). Assembled contigs were binned using three different methods (oligonucleotide frequency, pattern discovery, best blast hit) and genes were called using two gene prediction pipelines (fgenes, Critica/Glimmer). Based on the sequence assembly of the isolate genomes we evaluated the quality of each step in the process and explored the role of the population composition as well as the algorithms used in the quality of the final metagenomic dataset used for further analysis.