

M-tools and iVirus: Software tools and a cyberinfrastructure for meta-omic analyses

Joel A. Boyd^{1*} (joel.boyd@uqconnect.edu.au), Benjamin Bolduc^{3*} (bolduc.10@osu.edu), Paul G. Dennis¹, Michael Imelfort¹, Timothy Lambertson¹, Donovan H. Parks¹, Simon Roux³, Connor T. Skennerton², Ben J. Woodcroft¹, Ken Youens-Clark⁴, Phil Hugenholtz¹, Bonnie L. Hurwitz⁴, **Gene W. Tyson¹, Matthew B. Sullivan²**

¹Australian Centre for Ecogenomics (ACE), School of Chemistry & Molecular Biosciences, University of Queensland, Queensland, Australia; ²Division of Geological & Planetary Sciences, California Institute of Technology, Pasadena, California, USA; ³Department of Microbiology & Department of Civil, Environmental, and Geodetic Engineering, The Ohio State University, Columbus, Ohio, USA; ⁴Department of Agricultural & Biosystems Engineering, University of Arizona, Tucson, Arizona, USA

URL: <http://ecogenomic.org/software>; <http://ivirus.us>

Project Goals: Over the last decade, analysis of microbial and viral communities has undergone a major shift as the result of improvements in sequencing technology. The transition from single gene amplicon to large-scale meta-omic studies has allowed us to capture not only the phylogenetic but functional diversity of a community. However, the size and complexity of meta-omic datasets often require that multiple specialized tools be applied to address the different aspects of microbial and viral community analysis. Here we present two software suites, M-tools and iVirus, that are being developed as part of an ongoing interdisciplinary project focused on exploring the ecological and biogeochemical implications of climate change induced permafrost thaw.

The M-tools suite is organised around the analysis of microbial metagenomes, with a primary goal of providing a user-friendly pipeline for extracting high quality population genomes. The M-tools suite also provides software for community profiling of unassembled metagenomic data. iVirus is focused on collecting viral datasets and deploying the most commonly used tools for viral meta-omics, creating a publicly available, community resource ideal for sharing and collaboration. M-tools is comprised of six programs that span the workflow from raw data to high quality population genomes and iVirus has four apps developed for interrogating meta-omic datasets.

M-tools:

- **GraftM:** Creates community profiles from raw meta-omic sequences using Hidden Markov Models (HMM) to identify genes of interest which are classified using

phylogenetic tree insertion methods. Provides the tools for the creation of custom gene packages for the analysis of sequence data.

- **SingleM:** Provides highly resolved community composition from metagenomes using conserved single copy marker genes. By not heavily relying on reference databases of sequenced genomes, SingleM can be used to accurately profile communities that contain novel lineages. It can also be used to determine how representative a set of population genomes is of a community.
- **GroopM:** Recovers population genomes from large metagenomic datasets using differences in population abundance across metagenomic samples (differential abundance binning).
- **CheckM:** Assesses the quality of isolate, single cell, and population genomes using lineage specific single copy marker gene sets. Includes utilities for comparing genomes and exploring features such as GC content, sequence length, and tetranucleotide signatures.
- **RefineM:** Refines isolate, single cell or population genomes using qualitative and quantitative features such as GC content, coverage and coding density.
- **OrfM:** Rapidly predicts ORFs in raw metagenomic reads.

iVirus:

- **vContact/vContact-PCs:** Generates Protein Clusters (PCs) using a Markov clustering algorithm and incorporates metadata annotations. Then assigns contigs to taxonomic groups using the presence or absence of shared PCs along the length of the contig.
- **PCpipe:** Compares ORFs from user-defined datasets to existing viral PCs as a means to organize viral sequence space into functional units that can serve as (i) a universal functional diversity metric for viruses, (ii) a scaffold for iterative functional annotations, and (iii) input for ecological comparisons.
- **Fizkin:** Performs Bayesian network analyses based on the amount of shared sequence content in viromes and contextual data about the sample's environment.
- **VirSorter:** Identifies viral sequences in microbial genomes and metagenomic datasets

This work is supported by the U.S. Department of Energy under funding opportunity announcement number DE-FOA-0000866. Pathways to carbon liberation: a systems approach to understanding carbon transformations and losses from thawing permafrost and iVirus is supported by the US Department of Energy Office of Biological and Environmental Research under the Genomic Science program (Award DE- SC0010580).