Microbes Persist: Building a KBase Foundation for Viral Ecogenomics In Soil

Jeffrey A. Kimbrel*[1] (kimbrel1@llnl.gov), Benjamin Bolduc[2], **Matthew B. Sullivan[2,3]**, **Jennifer Pett-Ridge[1]**

[1]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA; [2]Ohio State University, Department of Microbiology, Columbus OH; [3]Ohio State University, Department of Civil, Environmental and Geodetic Engineering, Columbus OH

Project Goals: Microorganisms play key roles in soil carbon turnover and stabilization of persistent organic matter via their metabolic activities, cellular biochemistry, and extracellular products. Microbial residues are the primary ingredients in soil organic matter (SOM), a pool critical to Earth's soil health and climate. We hypothesize that microbial cellular-chemistry, functional potential, and ecophysiology fundamentally shape soil carbon persistence, and we are characterizing this via stable isotope probing (SIP) of genome-resolved metagenomes. We focus on soil moisture as a 'master controller' of microbial activity and mortality, since altered precipitation regimes are predicted across the temperate U.S. *Our SFA's ultimate goal is to determine how microbial soil ecophysiology, population dynamics, and microbe-mineral-organic matter interactions regulate the persistence of microbial residues under changing moisture regimes.*

The LLNL Soil Microbiome SFA is focused on the microbial biochemistry, functional potential and physiology of the soil microbiome. To measure and model the functioning of uncultivated soil communities, we are targeting specific ecophysiological 'traits' via analysis of genomes reconstructed from soil (informed by stable isotope tracing), as well as the role of phage in population turnover. Our approach to characterizing the role of viruses on microbial communities aims to identify, characterize and ecologically contextualize viruses in large-scale sequence datasets, all within the KBase environment. While metagenomic analytical platforms are now becoming widely available for microbial metagenomes, viral metagenomic analyses are critically different because:

1) Viruses lack universal genetic markers (they share no single gene), making taxonomic classification complicated with no standard available
2) Viral hosts must be predicted *in silico* to be able to evaluate virus impact on the ecosystem
3) Viral genomes can be extrachromosomal (episomal) or integrated (endogenised) into host genomes, which can confound 'viral' inferences if the genome 'termini' are not well identified
4) Coding sequences in viral genomes are governed by different rules that render prokaryote-centric structural and functional annotations less accurate

Currently, KBase, like many other platforms, does not contain opportunities for identification or exploration of viral and phage genomes. Our SFA team is currently adapting two viral workflow suites (iVirus and PhATE) into the KBase ecosystem, bringing these novel resources to the DOE research community and providing a foundation for future viral ecogenomic informatics development.

iVirus represents a collection of community and Sullivan Lab-based tools and datasets designed to enable cutting-edge viral ecology research, i.e., establishing and studying the patterns and processes that impact viral genes and genomes. This includes establishing 'bins' or 'units' that

represent functional sequence space (protein clusters), species- (viral populations) or genera- (viral clusters in network space) level taxonomy, and virus-tuned analytical tools to query these units once established. The incorporation of the iVirus suite of tools into KBase has already begun--with the vContact viral taxonomy tool (available now)—and more tools (such as identifying viral contigs in metagenomes) are coming soon. A challenge is that KBase uses a completely complementary relational database that is object-based instead of file-based (as in CyVerse where iVirus was 'born'). This requires significant time for establishing new objects and their relationships within the KBase environment. While this has slowed development time in the interim, as momentum builds, we expect to benefit for further downstream development.

The Phage Annotation Toolkit and Evaluator (PhATE) is a complementary LLNL-developed workflow for structural and functional viral genome annotation. PhATE begins by running multiple bacterial gene callers (GeneMarkS, Glimmer, Prodigal) and PHANOTATE, a phage-centric gene caller, and provides summary statistics for the resulting genes. Gene sequences (nucleotide and protein) are next BLASTed against several databases, including NCBI virus genomes and proteins, PhAnToMe, pVOGs, KEGG virus proteins, Swissprot, and NR. For KEGG virus proteins, annotation tags for each hit are queried and recorded (e.g., Pfam, taxonomy, uniprot identifiers). pVOG group identifiers are extracted from each top hit and used to construct an alignment-ready FASTA file data type containing the phage peptide of interest plus all of the members of a given pVOG group. In KBase, all annotations will be combined for each phage peptide and summarized in multiple downloadable output formats, including summary figures for visualization.

This research is based upon work supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research Genomic Science program under Award Number SCW1632 to the Lawrence Livermore National Laboratory, and a subcontract to the Ohio State University.