Integrating data and algorithms from the ENIGMA project into KBase

D.M. Needham[1], A. Zhang[1], J-M. Chandonia[2], D. Chivian[2], L.M. Lui[2], W. Zheng[3], S. Zhao[1], Y. Yin[1], D.A. Weitz[3], T.C. Hazen[4,5], P.S. Novichkov[2], J. Zhou[6], E.J. Alm[1], A.P. Arkin[2,7], P.D. Adams[2,7]

[1]Massachusetts Institute of Technology, Cambridge MA; [2]Lawrence Berkeley National Lab, Berkeley CA; [3]Harvard University, Cambridge MA; [4]University of Tennessee, Knoxville TN; [5]Oak Ridge National Lab, Oak Ridge TN; [6]University of Oklahoma, [7]University of California at Berkeley

Project Goals: ENIGMA -Ecosystems and Networks Integrated with Genes and Molecular Assemblies use a systems biology approach to understand the interaction between microbial communities and the ecosystems that they inhabit. To link genetic, ecological, and environmental factors to the structure and function of microbial communities, ENIGMA integrates and develops laboratory, field, and computational methods.

We aim to obtain novel genomes with high quality (of completeness and contamination) from ENIGMA samples through single-cell sequencing and integrate them into KBase as good references for not-yet cultured bacteria in natural environments.

As sequencing becomes less expensive, researchers are turning from 16S rRNA surveys to metagenomics in order to add a new level of functional and phylogenetic resolution to their sequence-based analyses. Metagenomic data harbors additional layers of data on population structure, strain dynamics, and genome evolution that cannot be inferred from 16S alone. Yet, powerful and user-friendly tools for the analysis of this data are not publicly available.

Microbial communities across nearly all biological systems, from sediment to groundwater, are diverse, dynamic ecosystems comprised of genetically diverse populations. Such diversity can be broad, as between species, as well as, within populations, as in strains of species. Typical metagenomic approaches explore this diversity in a manner than loses information the genomes of individual cells. In contrast, characterizations at the individual cell level yield information about interactions between organisms, such as between bacteria and phage, as well as strain level differences within a population. However, Public available genomes are good resources as reference for functional and taxonomy annotation, while most of them are culturable species from host-associated environments. Genomes of good quality and novelty are lacking to serve as references for not-yet cultured bacteria in natural environments. Thus, population genetic and evolutionary data analysis tools within the KBase environment and single cell sequencing of ENIGMA samples could have an outsize impact on environmental microbiology research.

Here we report new functions that we add to the KBase environment to catalyze metagenomic data analysis. First, we have built a standard and comprehensive set of reference genomes to which metagenomic reads can be compared. We have designed the pipeline to compare metagenomes to references and tested it in our samples. We built the estimators of strain level

nucleotide diversity, and even inference of strain genomes. We designed the tools to study within-population genome rearrangements and mutations. We published one compact tool, the meta_decoder (https://github.com/caozhichongchong/meta_decoder), that automatically identifies and compares the bacterial strains, mobile genetic elements, and phase variation across samples. We have tested meta_decoder using simulated datasets and we are now applying it to ENIGMA genomes and metagenomes.

We profiled 15,343 single genomes by droplet microfluidics (Microbe-seq) of an enigma groundwater sample GW-FW-305. We assembled many genomes, including 16 high quality bacterial genomes, > 75% completion and < 5% genomic redundancy. Several genomes were >97% complete and < 1% redundant, characteristics that are unusual for traditional metagenomes from a single sample. We uploaded these novel genomes with high quality (of completeness and contamination) into KBase as good references for not-yet cultured bacteria in natural environments.

We collaborate with ENIGMA data management team (John-Marc Chandonia) and KBase team (Dylan Chivian). The ENIGMA data we use is the Pseudomonas genomes from Lauren M. Lui (Arkin Lab).

This material by ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies a Scientific Focus Area Program at Lawrence Berkeley National Laboratory is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research under contract number DE-AC02-05CH11231