ENIGMA: Building a reference-based metagenomics workflow in KBase

An-Ni Zhang[1*] (anniz44@mit.edu), Shijie Zhao[1] and Eric J. Alm[1], A.P. Arkin[2,3] and P.D. Adams[2,3]

[1]Massachusetts Institute of Technology, Cambridge; [2]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory; [3]University of California, Berkeley, CA.

Project Goals: ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies uses 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.

Here we aim to develop a powerful and user-friendly tool for the analysis of genetic and evolutionary traits by integrating metagenomics and genomes. The tools developed will be embedded into the KBase platform and demonstrated using ENIGMA data.

Abstract: As sequencing becomes less expensive, researchers are turning from 16S rRNA surveys to shotgun sequence metagenomics in order to add new levels 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. Powerful and user-friendly tools for the analysis of these data are not yet widely available. We believe population genetic and evolutionary data analysis tools made available via the KBase platform will have an outsize impact on environmental microbiology research.

Here we report five new functions that we will 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 are now building the estimators of strain level diversity, and even inference of strain genomes. We will design the tools to study within-population genome rearrangements and mutations. Finally, we will design a new statistical approach that would allow data generated by different researchers using different protocols to be compared on equal footing.

In this work, we are collaborating with the ENIGMA data management team (John-Marc Chandonia) and the KBase team (Dylan Chivian). The ENIGMA data we are initially using are 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.