ENIGMA: Management and Analysis of ENIGMA Data using KBase

John-Marc Chandonia[*,1] (JMChandonia@lbl.gov), Pavel S. Novichov[1], Alexey E. Kazakov[1], Adam P. Arkin[1,2], and Paul D. Adams[1,2]

[1]Lawrence Berkeley National Lab, Berkeley; [2]University of California at Berkeley

http://enigma.lbl.gov

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.

The overarching goal of data management in ENIGMA is to enable integration of data from complex high-throughput field and laboratory studies towards multi-scale, predictive and systems level understanding of complex microbial community structure, function and evolution. To that end, we must ensure that all data are shared internally between ENIGMA teams, and also with scientists involved in other DOE programs in order to facilitate collaborative research and increase the overall pace of scientific discovery.

To enable more powerful data management and analysis, we have established a strategic collaboration with DOE Systems Biology Knowledgebase (KBase), a fast growing computational infrastructure that provides a unique solution for sharing both data and computational modules/pipelines, conforms to the SC digital data management policy, and supports private data, data sharing and provenance.

KBase has proven to be extremely valuable as an organizational tool for ENIGMA. All ENIGMA data are linked in a series of narratives from a single "master narrative" that is shared with all ENIGMA researchers. We have also helped to test and prototype KBase tools for managing organizations, using these ENIGMA narratives as a starting point.

Several large ENIGMA datasets that are amenable to analysis by KBase tools are stored as objects in KBase. These include the genomes of 297 sequenced isolates from ENIGMA campaigns, amplicon data from 16S surveys of 100 ENIGMA wells, and metagenomic shotgun reads. We use tools and workflows available in KBase to perform various types of bioinformatic analyses, including genome and metagenome assembly, QA/QC, and annotation; RNA-Seq differential expression, etc. We have applied a suite of KBase tools to analyze metagenomic samples obtained from three sediment cores across several projects (ENIGMA Core Pilot and Sediment Core Projects). In addition to being stored in KBase, public ENIGMA datasets are also deposited in appropriate repositories such as NCBI and MG-RAST. Other data that cannot currently be modeled in KBase are stored in accessible locations such as our Fitness Browser (fit.genomics.lbl.gov), or MAGI (magi.nersc.gov). These datasets are currently linked to ENIGMA narratives, and will be ported to KBase objects when such data can be modeled in KBase.

The ENIGMA data management team collaborates with other ENIGMA researchers to port tools and new types of data developed within ENIGMA into KBase. These include the metagenomics tools developed

by the Alm Lab (see poster by Anni Zhang) and metabolic analysis tools such as MAGI (see poster by Ben Bowen). We are also integrating into KBase a tool for functional and taxonomic profiling of shotgun metagenomic datasets, developed by the data management team in collaboration with the Arkin Lab. We monitor new data types and tools as they are introduced by ENIGMA team members, in order to update our priorities appropriately. We prioritize data types and computational tools for integration into KBase according to metrics such as scientific impact, costs, and feasibility of integration.

The ENIGMA data management team has also found that approximately 2/3 of ENIGMA data types would be well represented in KBase using "Generic" data objects that contain experimental measurements, which could be linked in KBase to non-Generic data types that represent biological or environmental objects (e.g., isolates and samples). We prototyped Generic data types, uploaders, graphing tools, search tools, and ontologies in KBase, and are currently collaborating with the KBase project to harden and deploy these technologies for use by ENIGMA team members as well as all other KBase users.

ENIGMA is a Scientific Focus Area Program at Lawrence Berkeley National Laboratory and is supported by the Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.