

Management and Analysis of ENIGMA Data using KBase

Pavel S. Novichkov^{*,1} (PSNovichkov@lbl.gov), John-Marc Chandonia¹, Alexey E. Kazakov¹ and Paul D. Adams^{1,2}

¹Lawrence Berkeley National Lab, Berkeley; ²University 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 date, we have deposited ENIGMA data into a number of publicly accessible locations, including NCBI and MG-RAST. We have also built specialized repositories that are accessible to the public (e.g., our fitness data browser available at fit.genomics.lbl.gov). To enable even more powerful data management and analysis in the future, 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.

We have performed project-wide surveys of ENIGMA data types and computational tools, and used this information to prioritize integration of our data and tools into KBase. Several large datasets, including 267 sequenced isolates from ENIGMA campaigns, as well as 16S surveys of 100 ENIGMA wells, have already been uploaded into KBase and shared for further computational analysis by ENIGMA groups. Such data have already been used for analysis of the genetic potential of isolates. We are continuing to monitor new data types and tools as they are introduced by ENIGMA team members, in order to update our priorities appropriately. We will prioritize all data types as well as key ENIGMA computational tools (e.g., the RB-TnSeq pipeline) for integration into KBase according to metrics such as scientific impact, costs, feasibility of integration.

The ENIGMA data management team has 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.

In addition to these “Generic Data” technologies, we are developing appropriate data models in KBase to describe other key ENIGMA data, such as environmental samples and sampling locations. We are also using KBase to organize and link to data that are stored elsewhere; e.g., MG-RAST, or data that are shared internally within the project using Google Drive. The data management team continues to facilitate work by ENIGMA scientists in uploading and sharing their data into KBase and other appropriate public repositories (e.g., GenBank). We are also providing assistance with project management (e.g., using JIRA when appropriate in order to track complex tasks that involve multiple laboratories).

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.