KBase: A case study illustrating the derivation and testing of mechanistic connections between geochemistry and the microbiome

Adam P. Arkin[1], **Robert Cottingham**[3], **Chris Henry**[2], Paramvir S. Dehal (psdehal@lbl.gov)[1*], Benjamin Allen[3], Jason Baumohl[1], Kathleen Beilsmith[2], David Dakota Blair[4], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Zachary Crockett[3], Ellen Dow[1], Meghan Drake[3], Janaka N. Edirisinghe[2], José P. Faria[2], Jason Fillman[1], Tianhao Gu[2], AJ Ireland[1], Marcin P. Joachimiak[1], Sean Jungbluth[1], Roy Kamimura[1], Keith Keller[1], Vivek Kumar[5], Sunita Kumari[5], Miriam Land[3], Sebastian Le Bras[1], Zhenyuan Lu[5], Filipe Lui[2], Dan Murphy-Olson[2], Erik Pearson[1], Gavin Price[1], Priya Ranjan[3], William Riehl[1], Boris Sadkhin[2], Samuel Seaver[2], Alan Seleman[2], Gwyneth Terry[1], Charles Trenholm[1], Sumin Wang[1], Doreen Ware[5], Pamela Weisenhorn[2], Elisha Wood-Charlson[1], Ziming Yang[4], Shinjae Yoo[4], Qizhi Zhang[2]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Argonne National Laboratory, Argonne, IL; [3]Oak Ridge National Laboratory, Oak Ridge, TN; [4]Brookhaven National Laboratory, Upton, NY; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
http://kbase.us

Project Goals: The Department of Energy Systems Biology Knowledgebase (KBase) is a knowledge creation and discovery environment designed for both biologists and bioinformaticians. KBase integrates a large variety of data and analysis tools, from DOE and other public services, into an easy-to-use platform that leverages scalable computing infrastructure to perform sophisticated systems biology analyses. KBase is a publicly available and developer-extensible platform that enables scientists to analyze their own data within the context of public data and share their findings across the system.

DOE is investing in multiple projects that attempt to understand the mechanistic processes of environmental ecology. These large-scale projects typically have multiple field sites that are gathering diverse forms of data including amplicons, metagenomes, extracted isolates, and geochemistry across time and space. For example, the ENIGMA project has field sites at Oak Ridge National Laboratory aimed at mechanistically understanding the microbial biogeochemical processes within contaminated subsurface sediment and has collected data across hundreds of wells and sediment cores resulting in over 4000 samples. ENIGMA and similar projects cross correlate taxa and geochemistry to infer the taxa and genomic functions that are constrained in their locational abundance and activity by environmental conditions and which in turn transform those environments. This type of research generally involves teams performing different measurements across the same samples collaboratively integrating and analyzing the resultant data using a plethora of complex computational approaches.

Here, we illustrate how innovations within the KBase platform can help support projects like ENIGMA. This analysis leverages new functionality delivered over the past year, including a new representation of Samples (See poster ***KBase: Significant Improvements***) with the ability to upload complex organized sets of Samples with detailed sample attributes describing the environment and sampling process. These attributes are validated and ontologically described to facilitate data sharing, reuse and downstream analysis. For each sample, data representing measurements of the Sample, like geochemistry, 16S amplicon reads, shotgun metagenome reads

and isolates are directly linked, in system, to their corresponding sample. This allows us to cross-correlate geochemistry, taxa, and gene function, as well as to evaluate relationships using isolates prioritized in our analyses. Over the last year we have also added analytical functionality for many of these data types that enable rich exploration of microbial ecological questions. With the ENIGMA data we demonstrate how to: 1) capture and organize multiple forms of measurement data linked to samples; 2) explore this data to identify ubiquitous taxa; 3) correlate across amplicons and geochemistry to identify key taxa linked to biogeochemical processes of interest; 4) leverage metagenomic data to understand the gene functions, genome organization and composition of key taxa, and the identification of isolates of interest for follow-up in laboratory experiments; and 5) build metabolic models of these isolates and MAGs that allow us to test these relationships. Together, these operations drive towards a generalized ability to mechanistically link environmental attributes to genes and genomes using a formal modeling framework to generate and evaluate hypotheses. Furthermore, the use of KBase enables the ENIGMA team to to share and collaborate on analysis, ensuring good stewardship of their products. They can also publish their data, tools, analysis and final results and evaluation of these results in a larger context of work done by the environmental microbiology community that is shared on the KBase system and thus can be included in their analysis.

Samples collected from the ENIGMA project were organized by field campaign and uploaded into KBase using the newly developed Samples framework. This capability allows KBase to capture critical metadata associated with samples that can be used to make important inferences in downstream analysis. These samples have a variety of extracted measurement data that have also been uploaded into KBase. Metagenomic reads and assemblies, metagenome assembled genomes, geochemistry, and 16S amplicon data are linked in KBase to corresponding samples. This allows users to easily see all of the measured data of a sample. Related Samples are grouped together in Sample Sets.

We introduced new analysis tools in KBase to assign taxonomy to marker gene sequences, filter and transform taxonomic abundance matrices, and correlate amplicon abundances with chemical abundance measurements in environmental samples. An amplicon from the prevalent Rhodanobacter genus was highly correlated with nitrous oxide concentration in ENIGMA samples, prompting us to examine denitrification in Rhodanobacter and co-occurring microbes. To further our understanding of these communities, we have assembled shotgun metagenomic reads, extracted metagenome assembled genomes, and scanned their functional repertoires to identify key lineages involved in nitrate reduction.

We constructed draft metabolic models of the identified key isolates and merged these models into a compartmentalized community model to mechanistically explore the microbial community dynamics. Model predictions are able to capture individual species' contributions to dynamics in nitrate reduction relating to energy biosynthesis and co-factor recycling with changes in the abundance of nitrate in the environment.

This work is supported as part of the BER Genomic Science Program. The DOE Systems Biology Knowledgebase (KBase) is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.