New pipeline for multi-omics data integration and discovery in KBase to identify the mechanism driving metabolic/OTU dynamics in a microbiome

Adam P. Arkin[1], **Bob Cottingham**[3], **Chris Henry**[2], and the KBase Team at the following institutions

[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 freely 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.

Increasingly microbiome systems are being interrogated using a combination of DNA sequencing and metabolome mass spectrometry with the ultimate goal of understanding how microbial communities shape (and are shaped by) the chemistry of their surrounding environments. Yet, the challenge of deciphering the biological mechanisms that give rise to observed dynamics in metabolite and species abundances in a given environment based on this data remains . Here we demonstrate a new workflow in KBase, comprised of many new data types and tools recently added to the KBase platform, that permit users to:
(1) identify metabolites and species that correlate based on cross-comparison of samples;
(2) search for isolates of species of interest that interact with metabolites of interest;
(3) predict biosynthesis pathways for a metabolite of interest in an isolate of interest;
(4) identify gene candidates for gap-filled steps within a predicted pathway;
(5) check phylogenetic neighbor genomes for evidence of conservation of pathway of interest;
(6) find and query available transcriptomes for evidence of expression of pathway of interest;
(7) identify other environments that involve similar species, metabolites, and pathways.

We demonstrate this new pipeline in action by analyzing a dataset from the Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA) SFA. In this study, a soil core was retrieved from a contaminated aquifer at the Oak Ridge National Lab Field Research Center (ORNL FRC), and fourteen biological samples were collected from this core at 9 inch vertical intervals. Each sample underwent both amplicon sequencing and metabolomic analysis, identifying a total of 3940 OTUs and 34 metabolites. Within KBase, we were able to correlate the OTUs to the metabolites. As we were particularly interested in applying a multi-omics approach mechanistically to understand these correlations, we applied the *Object Counts by Taxon* tool, which leverage the KBase Relation Engine that connects together related data entities in KBase, to identify which taxons in our dataset had the most multi-omics data available elsewhere in KBase. In this analysis, *Pseudomonas* emerged as the best candidate. Of the metabolites correlated to *Pseudomonas*, betaine had highest positive correlation. A search in

KBase revealed the Web of Microbes dataset uploaded by the Northen lab as part of a JGI collaboration with KBase, which demonstrated that a particular isolate of *Pseudomonas* is known to produce betaine as a byproduct. We applied the *Predict metabolite biosynthesis pathway* tool in KBase to identify 17 reactions and 30 genes involved in betaine biosynthesis in this strain of *Pseudomonas*, with three of the reactions having unknown genes (including the final step in betaine biosynthesis). We applied the *Find Candidate Genes for a Reaction* tool, which is another tool that leverages the KBase relation engine, to identify candidate genes for each of the three gap-filled steps in the betaine pathway. We then used the *Homolog Genome Context* to study how conserved all the betaine pathway genes (including our new candidate genes) are across all close *Pseudomonas* genomes. Close *Pseudomonas* genomes were determined using the *Mash Search* tool in KBase, which was developed in collaboration with JGI. This anlaysis revealed that the betaine pathway is broadly conserved in *Pseudomonas* genomes. We used data discovery tools (driven by the relation engine) in KBase to identify numerous public transcriptome profiles for some of the close *Pseudomonas* genomes identified by MASH. A search of our betaine gene families in these datasets revealed that they are most often expressed in antibiotic and protozoa induced stress conditions. Finally, we applied the *MAG (Metagenomic Assembled Genomes) Mash* tool, which was also developed in collaboration with JGI, to identify metagenome samples that contain similar pseudomonas genomes, revealing other environments where *Pseudomonas* and betaine production are potentially significant.

Overall, using the data discovery and analysis tools in KBase (and less than one hour of user-time), we were able to develop a mechanistic explanation for a metabolite-OTU correlation, identify missing genes in the betaine biosynthesis pathway in pseudomonas, determine that the betaine pathway is highly conserved in pseudomonas, pinpoint some conditions under which the betaine pathway is expressed, and explore some other environments where similar genomes are known to exist. The combination of tools applied in this workflow can be used to perform similar studies on numerous other microbiome and isolate systems, enabling the integration of multi-omics data to translate correlation to mechanistic understanding.

KBase is funded by the Genomic Science program within 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.