KBase : Omics driven discovery of novel functional capabilities in biological systems

Janaka N. Edirisinghe*[2] (janakae@anl.gov), Benjamin Allen[3], Jason Baumohl[1], Jay Bolton[1], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Zachary Crockett[3], Paramvir Dehal[1], Meghan Drake[3], José P. Faria[2], Annette Greiner[1], Tianhao Gu[2], James Jeffryes[2], 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], Akiyo Marukawa[1], Sean McCorkle[4], Cheyenne Nelson[1], 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], Gwyenth Terry[1], James Thomason[5], Doreen Ware[5], Pamela Weisenhorn[2], Elisha Wood-Charlson[1], Shinjae Yoo[4], Qizhi Zhang[2], **Robert Cottingham[3], Chris Henry[2] and Adam P. Arkin[1]**

[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.

Project Goals: Discovery and characterization of novel biochemistry is key to understanding the behavior of complex biological systems. To explain a microbe's role in biogeochemical cycles, or synthesis of various biofuel products from plant biomass, or to better understand community interactions within complex biological systems, it is essential to identify poorly explained or novel biochemical pathways. Though recent advances in system biology and the exponential growth of reference data suggest that there is an enormous amount of untapped enzyme potential, the traditional approaches for discovery of new functions and pathways are still mostly trial and error experimental processes. Using KBase, we demonstrate a mechanistic-modeling-based approach coupled with multi-omics data to support high throughput discovery of new metabolic pathways.

While there are system-biology approaches that have been developed in isolation for novel pathway identification, it has been a great challenge to translate raw data into an improved understanding of microbe-mediated chemical transformations in degradation and biosynthesis pathways. In KBase, users will be able to combine genomic, metabolic modeling and cheminformatic predictions reconciled with omics-data to predict novel pathway reactions as depicted in Figure 1. Currently, users are able to upload sequencing data then assemble, annotate and build metabolic models from isolate genomes or from metagenomes. Multi-omics data such as chemical abundance data (e.g.; FTICR data, MS2 metabolomics data) (Fig 1D), RNA-seq data (Fig 1E), and proteomics data (Fig IF) can then be mapped onto models and applied to predict active known pathways, reactions, and genes. This is done by optimizing the extent to which active reactions in the models are associated with genes with high expression, quantitatively measured enzyme levels and positively identified metabolites. However, it is often the case that much of the chemical abundance data does not get mapped onto the existing biochemistry databases, which suggests that many functioning pathways remain unidentified.

To predict these new pathways and compounds, we use a cheminformatics pipeline in order to expand the existing biochemistry based on enzymatic and spontaneous chemical rules, then reconcile against unmapped metabolomics data (Fig. 1C-D). With a continuously interconnected

metabolic network of known and predicted reactions, it is feasible to activate all pathways that are implicated based on omics data. While this strategy can generate multiple probable pathway routes, we would be able to filter out high confidence pathways based on thermodynamic feasibility of the predicted reactions (Fig. 1F) and also by eliminating the pathways that lack metabolomic, transcriptomic, or proteomic evidence (Fig 1. D, F, E). By having a set of high confidence novel reactions, we would be able to map potential gene candidates via chemical and or structural based gene finding approaches (reaction finger-printing) (Fig 1C, D2, D3) coupled with genes that are expressed according to transcriptomic and/or proteomic data. Finally, a subset of these predicted genes can be validated experimentally (Fig 1J).

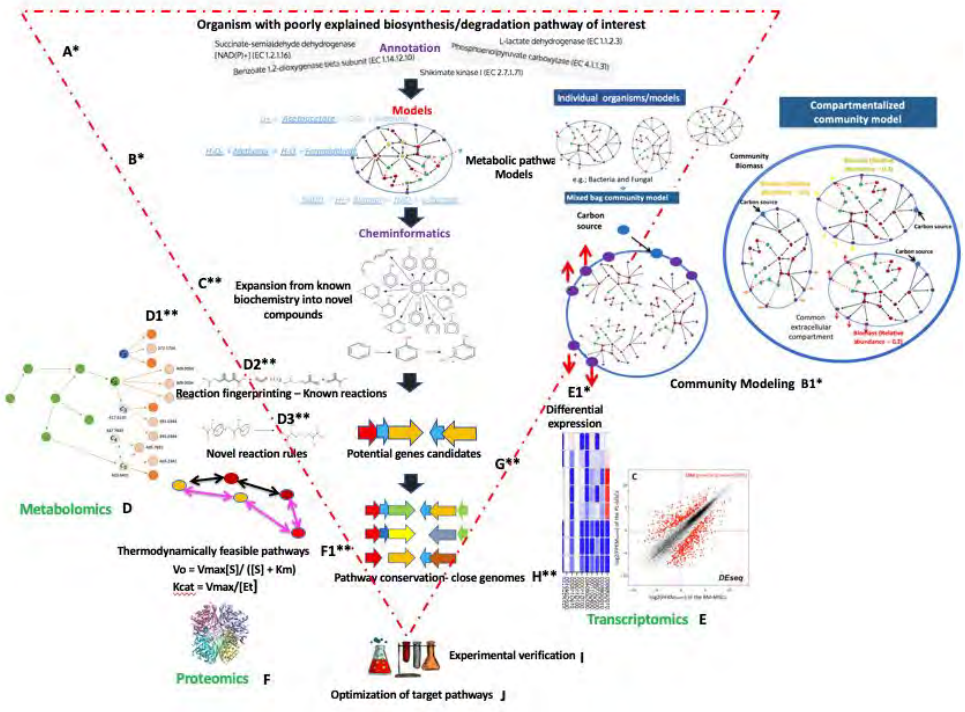

Figure 1. The figure shows the pathway discovery pipeline that has being implemented in DOE KBase and the integration of omics data at certain levels. The red triangle indicates the process of narrowing down into few target genes from thousands of genes in a genome for experimental validation. Steps labeled with a single asterisk (*) indicate that the specific component of the pipeline has already been implemented in KBase while the labels with a double asterisk (**) indicate that the functionality is still being developed. **B1**: **Community Modeling** - Construction of community models from individual models. There are two types of models that can be built in understanding the microbial communities. (i) Mixed-bag models (on left): Able to assess the overall metabolic capability of a community (ii) Compartmentalized community models (on right) able to assess the contribution of each member in community

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.