KBase: Significant Improvements to the DOE Systems Biology Knowledgebase in 2021

Adam P. Arkin[1], **Robert Cottingham**[3], **Chris Henry**[2*], Benjamin Allen[3], Jason Baumohl[1], Kathleen Beilsmith[2], David Dakota Blair[4], Jay Bolton[1], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Zachary Crockett[3], Paramvir Dehal[1], 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.

Over the last year, substantial new improvements were integrated into the KBase platform, with the goal of empowering upload, integration, and analysis of large-scale multi-omics datasets progressing toward a mechanistic understanding of environmental microbiomes. Given their complexity and diversity, the study of environmental microbiomes may require the analysis of hundreds of samples, often with multi-omics data (e.g., metagenome, amplicon, metabolomics, etc.) collected for each sample. This creates multiple significant challenges in supporting the analysis of such data on a platform like KBase.

First, hundreds of disparate data files must be loaded and integrated into the KBase data-model, a task that was previously quite tedious as most upload tools in KBase operated on one file at a time. A new bulk-upload pipeline has been established, enabling users to load hundreds of objects at once. These tools are now being applied by the Genome Resolved Open Watersheds (GROW) CSP and Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA) SFA to load hundreds of metagenomes and thousands of MAGs.

A second challenge involves the association of distinct data types collected from common experimental samples to: (1) allow for the organization of all data related to a complex experiment and (2) organize the data in such a way as to enable sophisticated cross-analysis of various data types. Toward this end, a Samples management system was integrated into KBase, enabling users to load sample metadata into KBase and associate sample IDs with data objects

on system (e.g., linking together the metagenome, metabolome, and MAGs derived from a common sample to facilitate integrated analysis). Tools are now being developed that exploit the relationship amongst data created by these sample associations to permit complex statistical analysis of data. We developed our Samples metadata, templates, and ID management systems in close collaborations with our partners at EMSL, JGI, NMDC, and ESS-DIVE.

A third challenge is handling all the datatypes required for multi-omics microbiome analysis, and to meet this challenge, this year KBase added deeper support and analysis pipelines for metabolomic and amplicon data. Now users can upload these new datatypes and integrate them with their broader microbiome data as well as reference data generally available on KBase (e.g., MGnify). Metabolomics and expression data can be visualized on rich metabolic maps, including new genome-wide maps for plants and fungi. Amplicon data can be visualized, compared, used to predict microbiome functional potential, and analyzed in the context of environmental data related to samples.

A fourth challenge is the difficulty associated with annotating protein functions within microbiomes. To improve support for this activity, and in collaboration with our User Working Groups, new tools have been added to KBase in the area of genome annotation. Users can now upload and compare many alternative annotations using tools developed by the LLNL Biofuels SFA; they can build models from those annotations and compare performance of competing annotations on predicting phenotypes like Biolog growth profiles; and they can apply new annotation algorithms like DRAM (integrated into KBase by the Wrighton lab) (1) and the Exascale Analysis tool (integrated into KBase by the Jacobson lab). Tools for metabolic reconstruction are improving dramatically for plants, fungi, and microbes, including new support for visualizing and painting data on metabolic reconstructions. All of this culminates in KBase offering a sophisticated set of tools to load, store, create, compare, test, and visualize annotations.

While much still remains to be done, this new functionality has significantly lowered the barrier for KBase users to develop a deeper, integrated, mechanistic understanding of complex environmental microbiome data based on rich and diverse experimental datasets.

References

1. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Research. 48:16. 8883–8900. https://doi.org/10.1093/nar/gkaa621.

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.