KBase: Leveraging Amplicon Analysis Tools to Generate Testable Hypotheses From Complex Natural Communities

Pamela Weisenhorn*[2](pweisenhorn@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], Janaka N. Edirisinghe[2], 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], Elisha Wood-Charlson[1], Shinjae Yoo[4], Qizhi Zhang[2]**,** **Robert Cottingham[3]**, **Chris Henry[2], 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.

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.

The U.S. Department of Energy (DOE) supports biological and environmental research to investigate the complex interactions within biological systems and the processes that shape soil, water, and ecological dynamics of our biosphere, and to harness these processes for sustainable production of energy and materials. KBase enables researchers to advance our fundamental knowledge of complex biological and environmental systems by providing the computing infrastructure necessary to integrate and analyze heterogenous data types and share their findings with the broader community. By simplifying data integration and exploration, KBase enables researchers to identify patterns in their data and provides workflows to move beyond patterns to predictive understanding of processes.

For many researchers, amplicon datasets (e.g. 16S targeted metagenomes) often provide the first insights into the dynamics and functioning of complex natural or synthetic microbial communities. Recently implemented tools in KBase will allow researchers to use these amplicon datasets to examine: patterns in taxon abundance at different taxonomic levels; differences in the overall composition of communities in response to experimental treatments or environmental conditions; and microbe-microbe interactions in relation to their environment. While many of these approaches can be implemented independently, the computational infrastructure of KBase allows researchers to use publicly available data and a breadth of bioinformatic tools to quickly move beyond mere identification of patterns and begin to explore the potential underlying mechanisms.

Using data collected as part of the Argonne Wetland Hydrobiogeochemistry project, we demonstrate how the amplicon tools in KBase can be used to identify patterns in microbial community composition and dynamics in response to environmental heterogeneity. We then demonstrate how KBase's diverse and integrated capabilities allow researchers to maximize the impact of their data and accelerate scientific discovery through deeper exploration of these patterns. Specifically, we demonstrate how connection between amplicon data in KBase and application of KBase's metabolic modeling, genome comparison, and auxotrophy prediction tools can generate testable hypotheses regarding the mechanisms underlying predicted microbe-microbe interactions. These predictions can then be used for the efficient design of focused experiments or field campaigns.

The ability to easily explore patterns and predictions both within and across projects will continue to be advanced through implementation of metadata standards for environmental samples and taxonomic abundance matrices (currently under development by DOE Environmental System Science's Data Infrastructure for a Virtual Ecosystem project). We discuss how these standards will be used to facilitate import of data from the ESS-DIVE archive and the role of such cross-platform standards in advancing our fundamental understanding of complex biological and environmental systems.

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. Data were collected by the Argonne Wetland Hydrobiogeochemistry SFA which is funded by the Subsurface Biogeochemical Research Program, under Contract DE-AC02-06CH11357.