KBase: Microbiome and Phylogenomics Capabilities

Dylan Chivian*[1] (DCChivian@lbl.gov), **Adam P. Arkin[1]**, **Robert Cottingham[3]**, **Chris Henry[2]**, Benjamin Allen[3], Jason Baumohl[1], Jay Bolton[1], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[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], Pamela Weisenhorn[2], Elisha Wood-Charlson[1], 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.

KBase was designed to enable systems biology analysis of communities of microbes and/or plants. KBase is extensible and currently includes powerful tools for metabolic modeling, comparative and phylogenomics of microbial genomes that can be used for developing mechanistic understanding of functional interactions between species in microbial ecosystems. Essential to gaining new insight is obtaining high-quality genomes to annotate, either via cultivation or genome extraction, from metagenome assembly. KBase has incorporated and added to a suite of microbiome analysis apps meant to be used in concert, including sequence QA/QC tools such as Trimmomatic and FastQC, taxonomic structure profiling of shotgun metagenome sequence with Kaiju, custom KBase apps for generating sample-specific *in silico* reads for downstream benchmarking, several metagenome assembly algorithms including MEGAHIT, IDBA-UD, and metaSPAdes, custom KBase apps for comparing metagenome assemblies, grouping assembled genome fragments (contigs) into putative genomes (bins) with MaxBin2 and other binners, and genome completeness and contamination assessment and filtering with CheckM. Tools for fractionation of unassembled reads and unbinned contigs to permit taxonomic and functional assessment of unbinned portions of samples are offered, and can also be used to reassemble individual bins of interest. Additionally, we've recently released tools and services that allow users to search rapidly (seconds to minutes) all reference genome databases, metagenomes and published metagenome-assembled genomes (MAGs) using their

reads, assemblies or MAGs. This is implemented using a MinHash like sketching process that works well for identifying matches above ~90% identity.

We have greatly expanded microbiome analysis in KBase. In addition to support for amplicon-based analyses (please see poster KBase: Leveraging Amplicon Analysis Tools to Generate Testable Hypotheses From Complex Natural Communities), it is now possible to incorporate and use tools that enable users to get from shotgun reads through to MAGs to phylogenomics and metabolic modeling. As an example from our initial set of tools, a user can upload or find data from collaborators or the public and apply one of the metagenome assembly apps and bin the assembled contigs so that individual genomes can be extracted from the bins. Once individual MAGs are extracted, the highest quality fraction can be piped into a wide range of downstream analysis apps in KBase, including genome annotation, phylogenetic placement and genome content comparison with respect to one another, KBase reference genomes, and other public genome and MAG collections. Unbinned metagenome assemblies can also have gene annotation for analysis. For high-quality MAGs, metabolic modeling and RNA-seq alignment can be performed (please see poster KBase: Omics driven discovery of novel functional capabilities in biological systems). Pangenome calculations among related genomes can be combined with phylogenetic and functional analysis to capture evolutionary histories of gene families and allow researchers to investigate functional repertoires and niche roles of microbial lineages.

In addition to efforts by KBase developers to expand the functionality of our Microbiome tool suite, community developers have been adding tools that they use and have developed, including members of the DOE Joint Genome Institute (BBMap, MetaBAT2, RQCFileter, JGI Metagenome Assembly Pipeline), the ENIGMA SFA, the LLNL Soils SFA (vConTACT2, VirSorter), and the LANL Bacterial:Fungal Interactions SFA (GOTTCHA2). All Apps in KBase are openly available for users to apply with their own data.

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.