Microbiome Genome Extraction, Phylogenomics, and Metabolic Modeling of Species Interactions in KBase

Dylan Chivian*[1] (DCChivian@lbl.gov), José P. Faria*[2] (JPLFaria@anl.gov), Janaka N. Edirisinghe*[2] (JanakaE@anl.gov), Paramvir S. Dehal*[1] (PSDehal@lbl.gov), Richard S. Canon[1], Elisha Wood-Charlson[1], **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 DOE Systems Biology Knowledgebase (KBase) is a free, open-source software and data platform that enables researchers to collaboratively generate, test, compare, and share hypotheses about biological functions; analyze their own data along with public and collaborator data; and combine experimental evidence and conclusions to model plant and microbial physiology and community dynamics. KBase's analytical capabilities currently include (meta)genome assembly, annotation, comparative genomics, phylogenomics, transcriptomics, and metabolic modeling. Its web-based user interface supports building, sharing, and publishing reproducible, annotated analysis workflows with integrated data. Additionally, KBase has a software development kit that enables the community to add functionality to 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 with CheckM. Comparison of results and quality assessment of the performance of tools and parameterization against data of various characteristics (e.g. low-complexity, high-complexity) by benchmarking at each stage of this process are offered. 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.

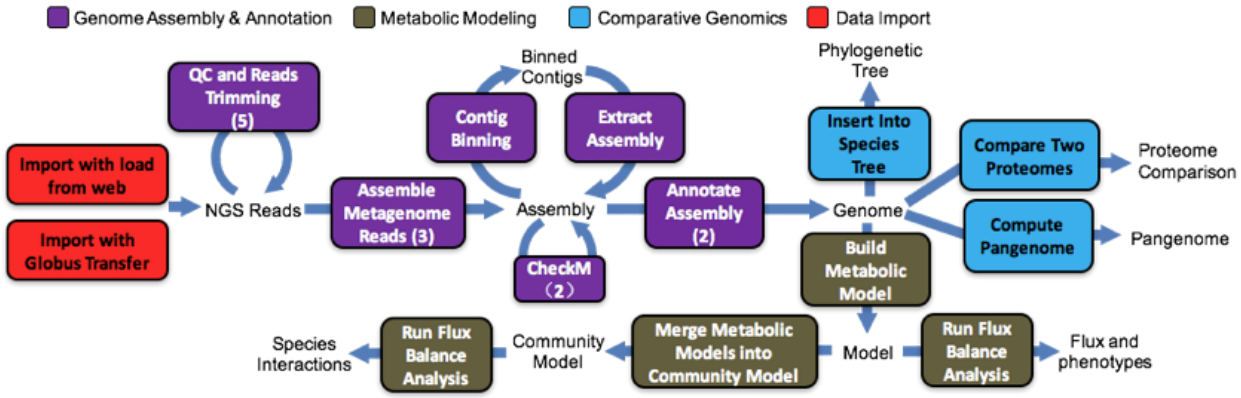

Figure 1. Example approach for use of KBase's microbiome genome extraction tool suite.

We have greatly expanded microbiome analysis in KBase. 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 (Figure 1), 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. Additionally, metabolic modeling and RNA-seq alignment can be performed. After generating metabolic models from the genomes assembled from a metagenome, individual metabolic models can be combined into a community metabolic model, which can be applied with the Flux Balance Analysis app to predict trophic interactions between species. Users have applied these tools to study: (i) interactions between plants and microbes in soil; (ii) why some microbes form stable communities; (iii) how a microbial community cooperates to produce a specific product; and (iv) how a community of heterotrophic species can feed on byproducts from an autotroph to grow autotrophically.

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 (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.