

Microbiome Genome Extraction and Metabolic Modeling of Species Interactions in KBase

Dylan Chivian*¹ (DCChivian@lbl.gov), José P. Faria*² (JPLFaria@anl.gov), **Adam P. Arkin**¹, **Bob Cottingham**³, **Chris Henry**², and the KBase Team at the following institutions

¹Lawrence Berkeley National Laboratory, Berkeley, CA; ²Argonne National Laboratory, Argonne, IL; ³Oak Ridge National Laboratory, Oak Ridge, TN; ⁴Brookhaven 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, 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 has powerful tools for metabolic modeling and comparative phylogenomics of microbial genomes that can be used for developing mechanistic understanding of functional interactions between species in microbial ecosystems. Essential to this process is obtaining high-quality genomes to annotate, either via cultivation or genome extraction from metagenome assembly. KBase is developing 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, community member interaction covariation with SparCC, 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 genome completeness and contamination assessment with CheckM. The apps that have been incorporated to date represent a single offering in each of these analysis steps; we expect additional apps will be incorporated into KBase that offer alternate algorithms to allow users to obtain variation in their results, such as alternate binning tools.

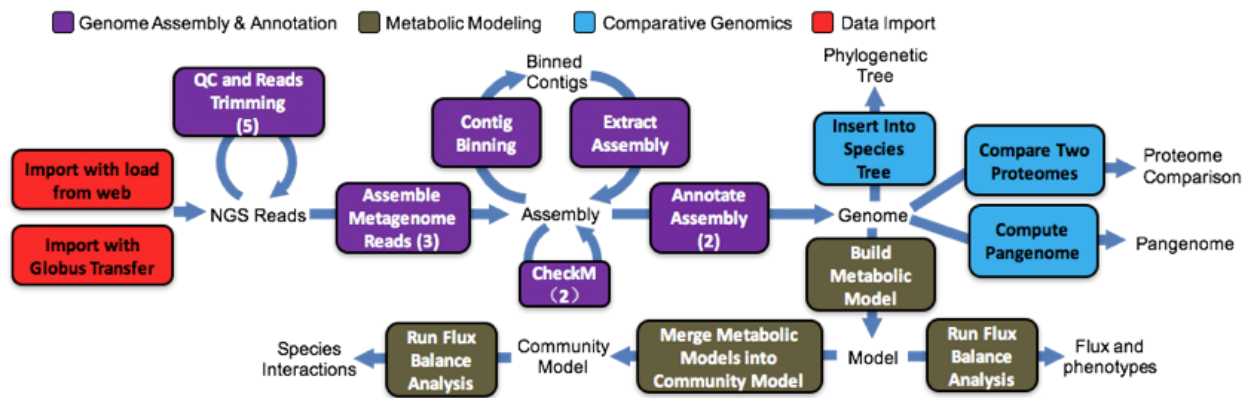


Figure 1. KBase's microbiome analysis pipeline.

KBase's microbiome analysis pipeline (see Fig. 1) begins with the upload of reads files. Next, the user can apply one of the metagenome assembly apps and bin the assembled contigs so that individual genomes can be extracted from the bins. Once the high-quality individual genomes are extracted, they 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 KBase reference genomes, metabolic modeling, and RNA-seq alignment. 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.

KBase's current microbiome analysis pipeline is just a starting point. It works best on microbial communities with limited diversity, and presently can only be applied to shotgun metagenomic data. In the future, we plan to support analysis of 16S amplicon data, as well as enabling further direct taxonomic and functional profiling of shotgun metagenomic reads.

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.