Comparative Phylogenomics in KBase

Dylan Chivian*[1] (DCChivian@lbl.gov), Priya Ranjan*[3], **Adam P. Arkin[1]**, **Bob Cottingham[3]**, **Chris Henry[2]** and the KBase Team at the following institutions

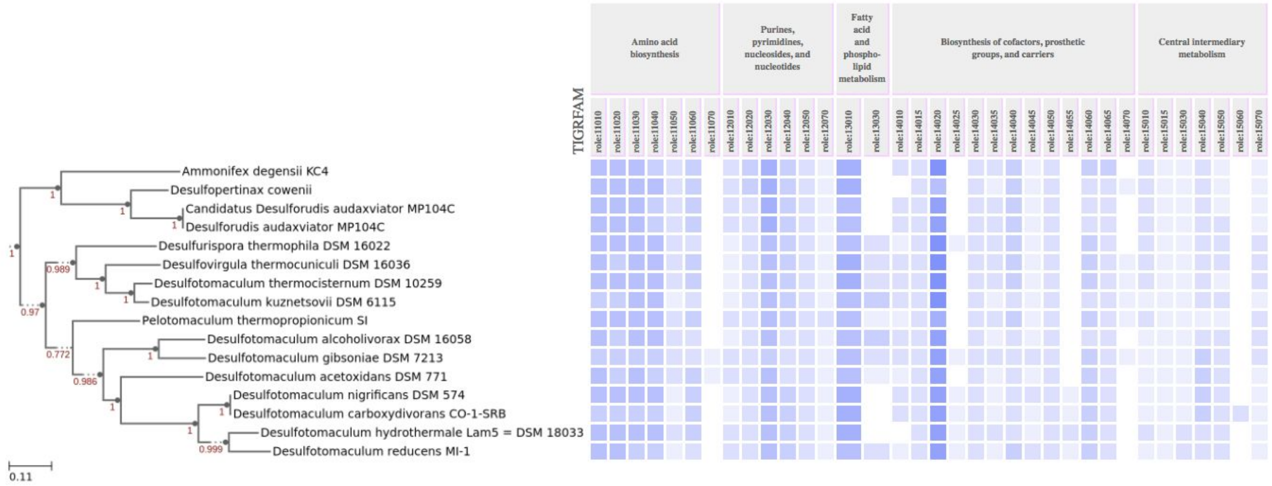
[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Argonne National Laboratory, Argonne, IL; [3]Oak Ridge National Laboratory, Oak Ridge, TN; [4]Brookhaven National Laboratory, Upton, NY; [5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

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, 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.

Understanding the functional complements of genomes is essential to placing them in the correct ecological and evolutionary context. Interactions between species, their roles within environments, and the genetic underpinnings of individual and community phenotype greatly benefit from holistic views of the genomes of related lineages available using comparative phylogenomic methodologies. To this end, KBase is developing a suite of tools for phylogenetic and pangenomic analyses that allow the user to analyze their uploaded or KBase-assembled genomes, including microbial reference genomes from NCBI's RefSeq database.

A foundational KBase App for such work is "Build Species Tree", which places a genome or genomes of interest into a Species Tree. Instead of using a 16S tree, KBase takes advantage of the universal protein-coding phylogenetic marker genes to build more reliable species trees. After identifying related species in RefSeq, the user can then tailor a GenomeSet that combines the user's genomes with the desired reference genomes for subsequent analyses. Such work includes "hands-on" analyses such as searching the genomes for gene matches with BLAST or HMMER, building and refining multiple sequence alignments with MUSCLE and Gblocks, and building gene trees with FastTree2.



KEY

Amino acid biosynthesis	TIGR_role:11010	Aromatic amino acid family
	TIGR_role:11020	Aspartate family
	TIGR_role:11030	Glutamate family
	TIGR_role:11040	Pyruvate family
	TIGR_role:11050	Serine family
	TIGR_role:11060	Histidine family
	TIGR_role:11070	Other

Figure 1. Functional Profiling with Gene Families

KBase has apps for computing pangenomes for a set of Genomes, such as OrthoMCL, as well as annotation with canonical protein-coding domain family content, such as COG, Pfam, and TigrFAMS. The results of these can be fed to downstream apps, including “Phylogenetic Pangenome Accumulation”, gene-family-based “Functional Profiling” (Figure 1), and “Pangenome Circle Analysis”, which shows visualizations of core, intermediate and singleton gene sets. These enable the user to focus in on functions, genes, and those genomes them. Additionally, KBase is developing the ability for a user to scan their genomes with custom gene family Hidden Markov Models, either created by the user or pre-developed ones, such as the dbCAN models for the CAZy database, or the antibiotic resistance gene family models of ResFam. We expect continued growth in the capabilities of the Comparative Phylogenomics apps suite as the community of KBase developers incorporates their own favorite tools as KBase apps using KBase’s Software Development Kit.

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.