KBase: An Integrated Systems Biology Knowledgebase for Predictive Biological and Environmental Research

Adam P. Arkin[1], **Robert Cottingham***[3] (cottinghamrw@ornl.com), **Chris Henry**[2], Nomi Harris[1], Ben Allen[3], Jason Baumohl[1], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Paramvir Dehal[1], Meghan Drake[3], Janaka Edirisinghe[2], Jose Faria[2], Uma Ganapathy[4], Annette Greiner[1], Tian Gu[2], James Jeffryes[2], Marcin Joachimiak[1], Roy Kamimura[1], Keith Keller[1], Vivek Kumar[5], Sunita Kumari[5], Miriam Land[3], Sean McCorkle[4], Arman Mikaili[2], Dan Murphy-Olson[2], Arfath Pasha[4], Erik Pearson[1], Gavin Price[1], Priya Ranjan[3], William Riehl[1], Samuel Seaver[2], Alan Seleman[2], James Thomason[5], Doreen Ware[5], Shinjae Yoo[4], Qizhi Zhang[2], Diane Zheng[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 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.

The U.S. Department of Energy (DOE) has invested substantially in environmental and biological system science research to investigate the complex interplay between biological and abiotic processes that influence soil, water, and environmental dynamics of our biosphere. The community that has grown around these efforts has recognized the need to lower the barrier to accessing computational tools, data, and results, and to work collaboratively to accelerate the pace of their research. The DOE Systems Biology Knowledgebase (KBase, kbase.us) is a software platform designed to provide these needed capabilities.

KBase currently has over 160 analysis tools (see https://narrative.kbase.us/#appcatalog) that offer diverse scientific functionality for (meta)genome assembly, contig binning, genome annotation, sequence homology analysis, tree building, comparative genomics, metabolic modeling, community modeling, gap-filling, RNA-seq processing, and expression analysis (see Figure 1). Users can build and share sophisticated workflows by chaining together multiple apps–for example, one could predict species interactions from metagenomic data by assembling raw reads, binning assembled contigs by species, annotating genomes, aligning RNA-seq reads, and reconstructing and analyzing individual and community metabolic models.

Computational experiments in KBase are saved in the form of *Narratives*. A finished Narrative represents a complete record of everything the authors did to complete their analysis. This recording of a user's KBase activities within a sharable Narrative is a central pillar of KBase's support for reproducible transparent research, simplifying the re-purposing, re-application, and extension of scientific techniques.

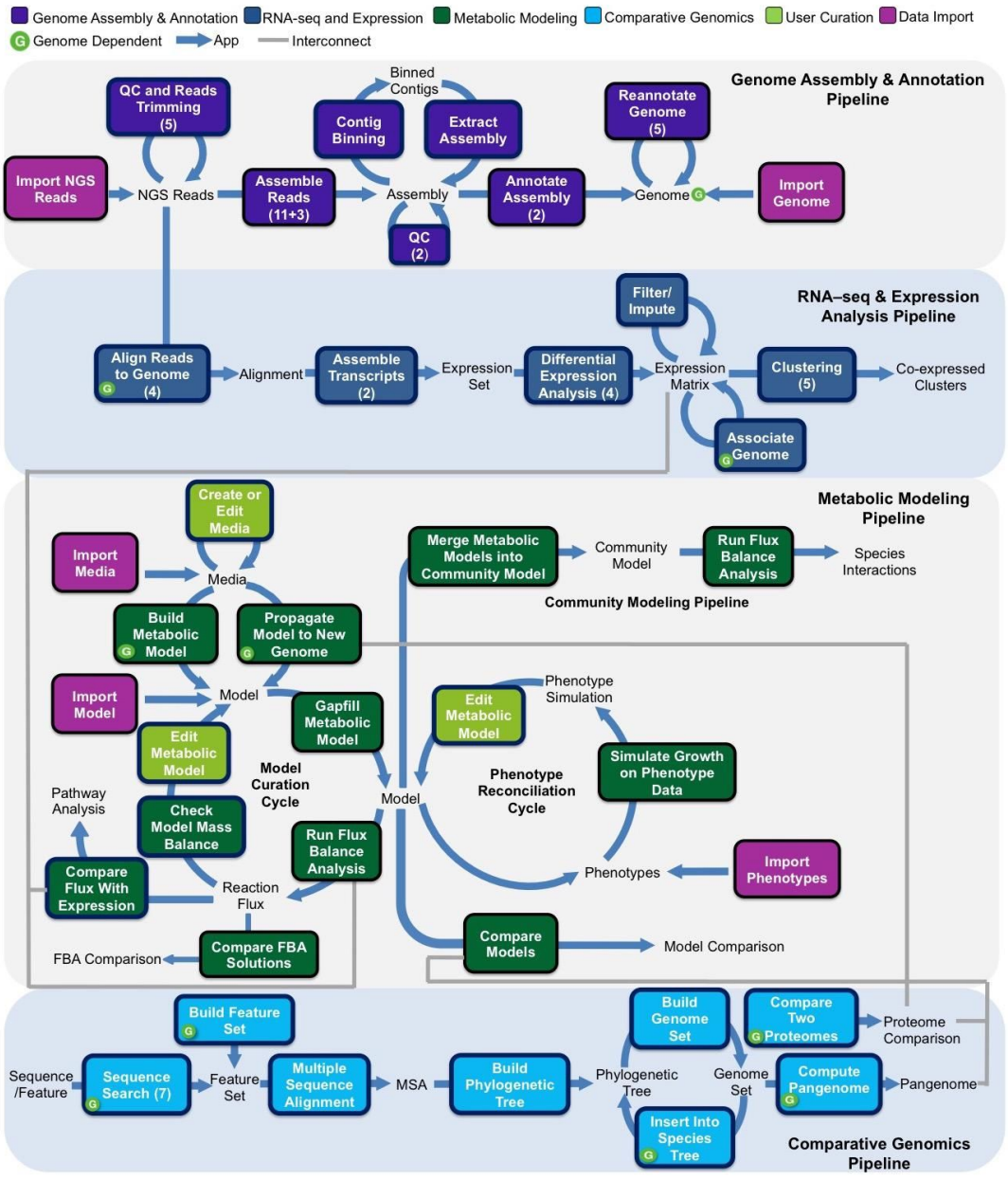

Figure 1. Outline of the major workflows and datatypes in KBase. See http://kbase.us/apps for more information.

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.