

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

Adam P. Arkin¹, Robert Cottingham^{*3} (cottinghamrw@ornl.com), Chris Henry², Nomi Harris¹, Ben Allen³, Jason Baumohl¹, Shane Canon¹, Stephen Chan¹, John-Marc Chandonia¹, Dylan Chivian¹, Paramvir Dehal¹, Meghan Drake³, Janaka Edirisinghe², Jose Faria², Uma Ganapathy⁴, Annette Greiner¹, Tian Gu², James Jeffryes², Marcin Joachimiak¹, Roy Kamimura¹, Keith Keller¹, Vivek Kumar⁵, Sunita Kumari⁵, Miriam Land³, Sean McCorkle⁴, Arman Mikaili², Dan Murphy-Olson², Arfath Pasha⁴, Erik Pearson¹, Gavin Price¹, Priya Ranjan³, William Riehl¹, Samuel Seaver², Alan Seleman², James Thomason⁵, Doreen Ware⁵, Shinjae Yoo⁴, Qizhi Zhang², Diane Zheng¹

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

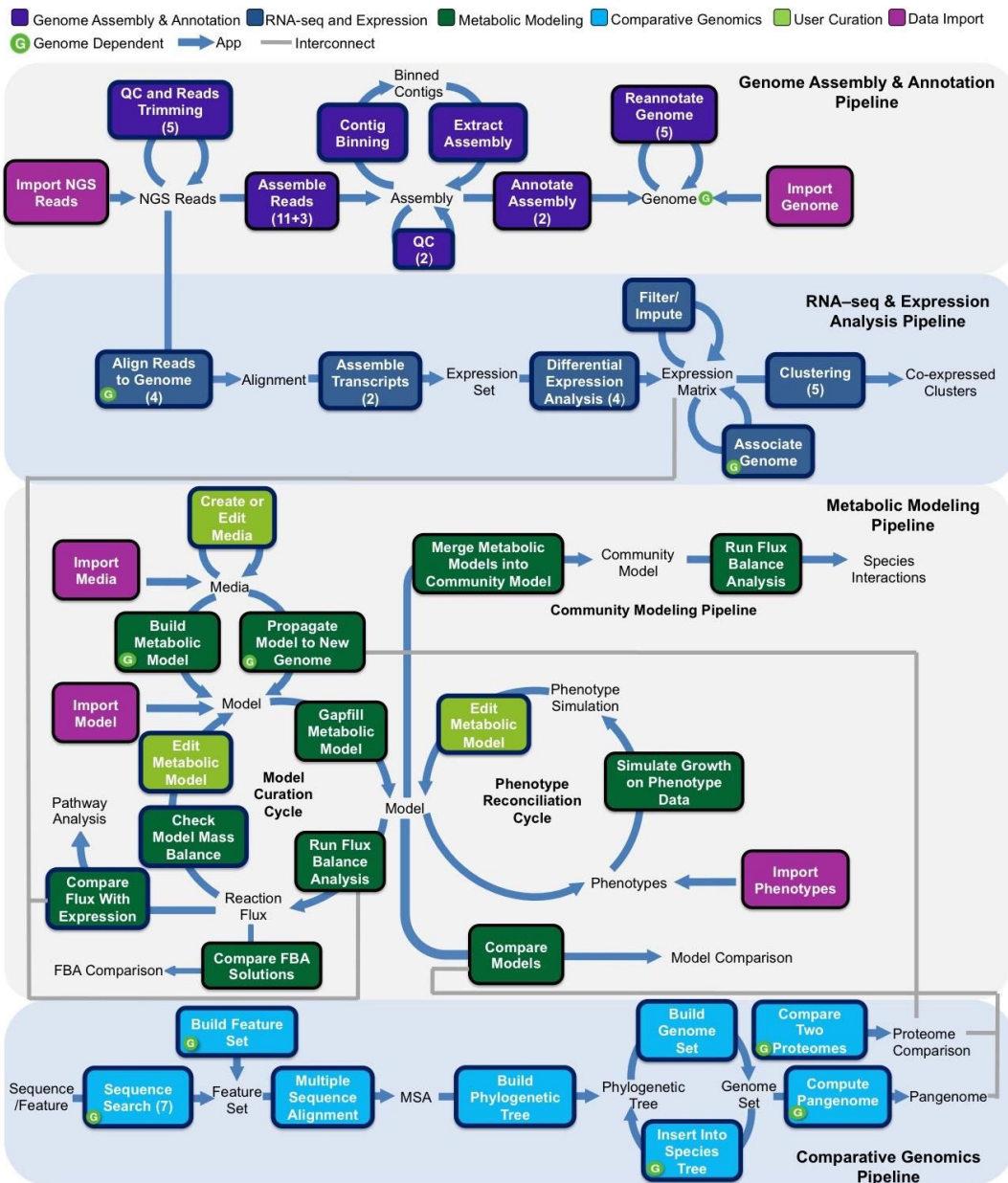


Figure 1. Outline of the major workflows and datatypes in KBase. See <http://kbase.us/apps> for more information.

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.