KBase: The Systems Biology Knowledgebase for Predictive Biological and Environmental Research in an Integrated Data Platform

Adam P. Arkin[1], **Robert Cottingham**[3], **Chris Henry**[2], Benjamin Allen*[3] (allenbh@ornl.gov), Jason Baumohl[1], Jay Bolton[1], Shane Canon[1], Stephen Chan[1], John-Marc Chandonia[1], Dylan Chivian[1], Zachary Crockett[3], Paramvir Dehal[1], Meghan Drake[3], Janaka N. Edirisinghe[2], José P. Faria[2], Annette Greiner[1], Tianhao Gu[2], James Jeffryes[2], Marcin P. Joachimiak[1], Sean Jungbluth[1], Roy Kamimura[1], Keith Keller[1], Vivek Kumar[5], Sunita Kumari[5], Miriam Land[3], Sebastian Le Bras[1], Zhenyuan Lu[5], Akiyo Marukawa[1], Sean McCorkle[4], Cheyenne Nelson[1], Dan Murphy-Olson[2], Erik Pearson[1], Gavin Price[1], Priya Ranjan[3], William Riehl[1], Boris Sadkhin[2], Samuel Seaver[2], Alan Seleman[2], Gwyenth Terry[1], James Thomason[5], Doreen Ware[5], Pamela Weisenhorn[2], Elisha Wood-Charlson[1], Shinjae Yoo[4], Qizhi Zhang[2]

[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 Department of Energy Systems Biology Knowledgebase (KBase) is a knowledge creation and discovery environment designed for both biologists and bioinformaticians. KBase integrates a large variety of data and analysis tools, from DOE and other public services, into an easy-to-use platform that leverages scalable computing infrastructure to perform sophisticated systems biology analyses. KBase is a publicly available and developer extensible platform that enables scientists to analyze their own data within the context of public data and share their findings across the system.

The U.S. Department of Energy (DOE) supports biological and environmental research to investigate the complex interactions within biological systems and the processes that shape soil, water, and ecological dynamics of our biosphere, and to harness these processes for sustainable production of energy and materials. KBase is an open-source data-science platform funded by DOE to enable sharing, integration, and analysis of many types of data associated with microbes, plants, and their communities using scalable computing infrastructure. This extensive public resource is designed to facilitate large-scale analyses of biological and environmental systems while accelerating scientific discovery, improve reproducibility, and foster open collaboration.

KBase offers a suite of scientific applications to enable users to build sophisticated analytical workflows, share their findings, and organize their projects. Over 200 apps in KBase offer diverse scientific functionality across the realms of comparative genomics, microbial communities, metabolic modeling, and transcriptomics. Several tools and services in KBase have been co-developed with the DOE Joint Genome Institute. Users can build and share sophisticated workflows through a combination of chaining together multiple analysis tools, writing scripts for automation, and using batch processing, all within notebook-style *Narrative*s that contain the employed data and tools. Projects, laboratories, and even whole institutions can organize their users and associated Narratives into a shared *Organization* with multiple permission levels and

management features. Developers can build, test, register, and deploy new or existing software as KBase apps using the Software Development Kit, thereby extending the platform's scientific capabilities.

Newly added services include several tools collaboratively developed by DOE SFA programs and KBase staff, including NWchem, GOTTCHA2, VIRSorter, and vConTACT2. KBase is unique in offering these diverse and integrated capabilities to a growing community of several thousand users. A central premise of KBase is to maximize the impact of data shared and developed between all users on system. This is the premise of KBase's emerging Knowledge Engine technology, which draws inferences between user data and public repositories, so scientists can better understand their work in the context of public knowledge. As data propagates across the system, it can be continually updated with new information like predictions of gene function, allowing the user to see how these changes scale from genes to ecologies and better predict outcomes.

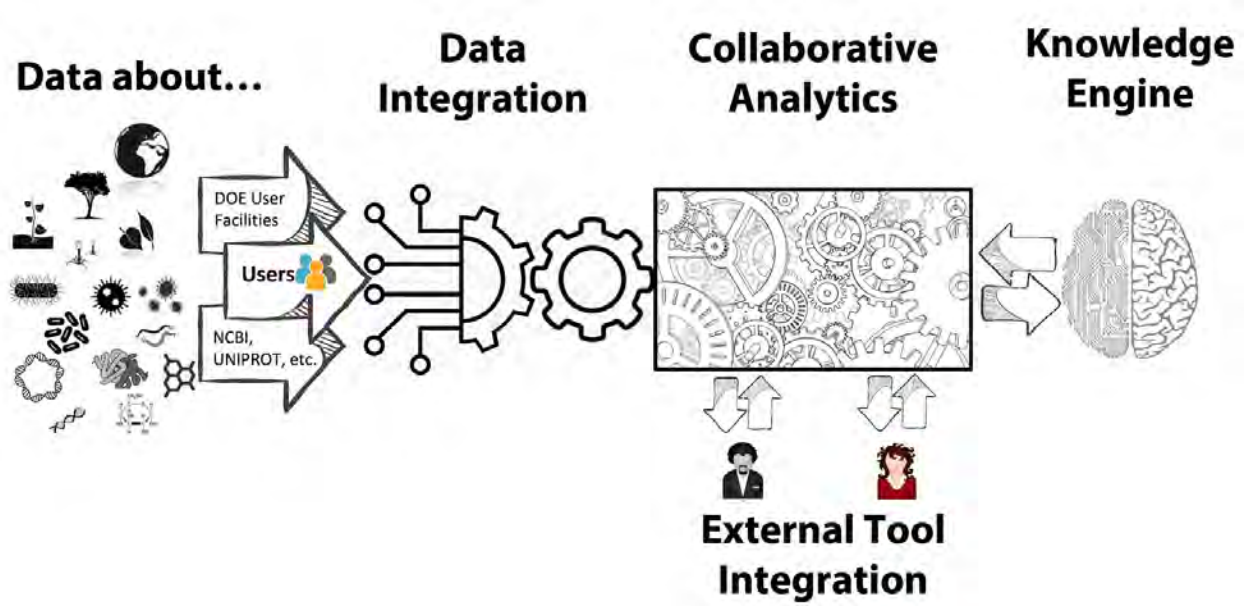

Figure 1. Integration of data and tools into KBase.

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.