KBase: The DOE Systems Biology Knowledgebase

KBase PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase
kbase.us

An Overview

What is KBase?

KBase is an integrated software and data platform designed to meet the grand challenge of systems biology—predicting and designing biological function on a range of scales, from the biomolecular to the ecological. Users can perform large-scale analyses and combine multiple lines of evidence to model plant and microbial physiology and community dynamics (see Fig. 1).

Supported by the U.S. Department of Energy (DOE) Office of Science, KBase is the first large-scale bioinformatics system that enables users to upload their own data, access collaborator or public data, perform sophisticated analyses, build increasingly accurate models of dynamic cellular systems for microbes and plants, and publish their workflows and conclusions. With a mission beyond that of a simple database or workbench, KBase seeks to enable the iterative investigation of complex biological data by scientists encouraged and empowered by the system to collaborate and share results. As a true knowledgebase, KBase is being designed to quickly percolate these new insights across its body of knowledge, allowing researchers to more rapidly test new hypotheses against their own observations and models. In this integrated environment, results obtained from one biological system would readily inform others, thereby accelerating the pace of research.

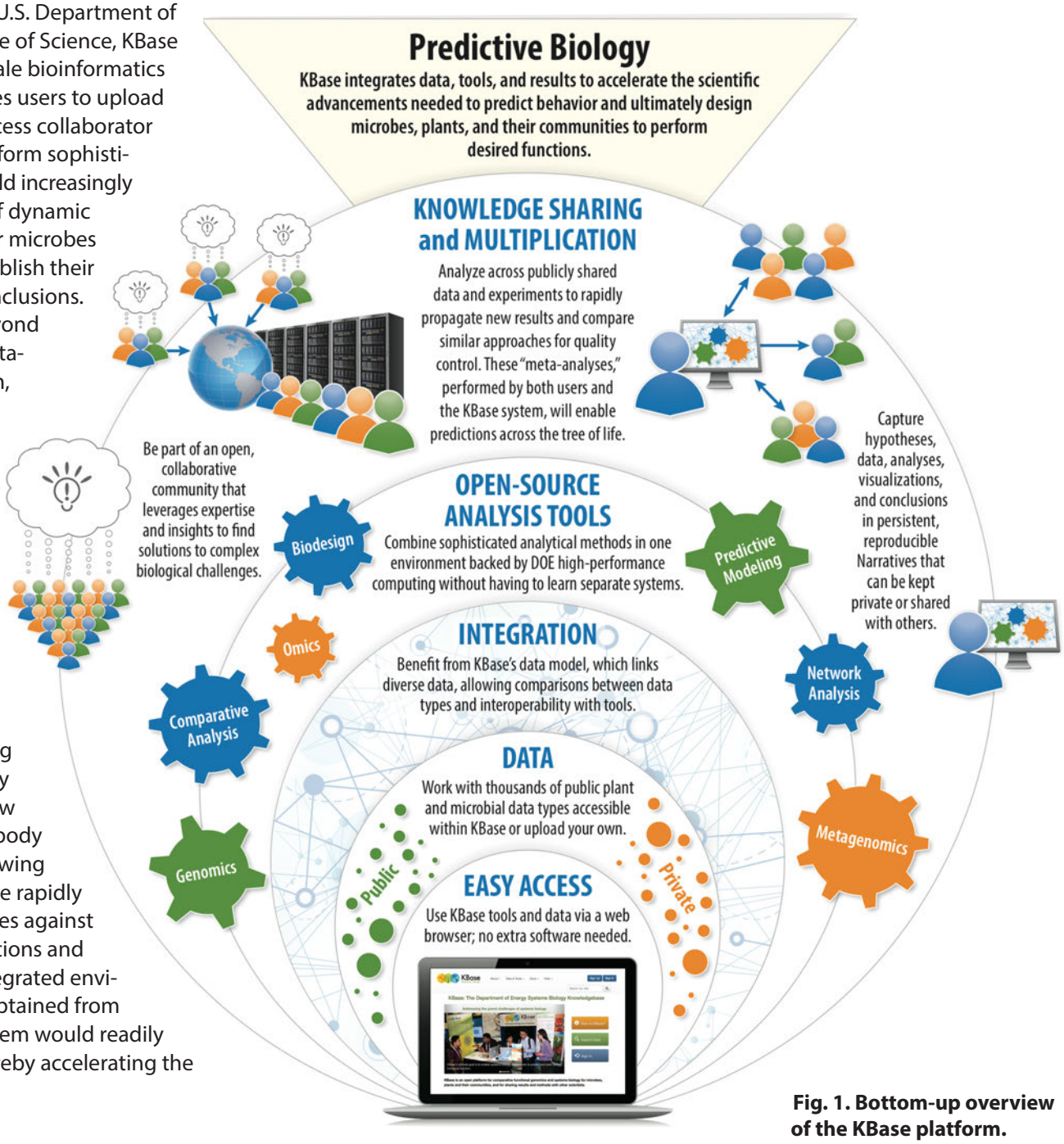


Fig. 1. Bottom-up overview of the KBase platform.

The Narrative Interface

In KBase, computational experiments are captured in dynamic, interactive documents called **Narratives** (see Fig. 2) that promote collaboration and reproducibility of scientific results. In addition to data and analysis steps, Narratives can include user images, notes, and commentary. They can be kept private, shared with colleagues and collaborators, or be made public for the benefit of the research community. Because the Narrative Interface is built on the Jupyter Notebook, users in the future will be able to write custom scripts in their Narratives.

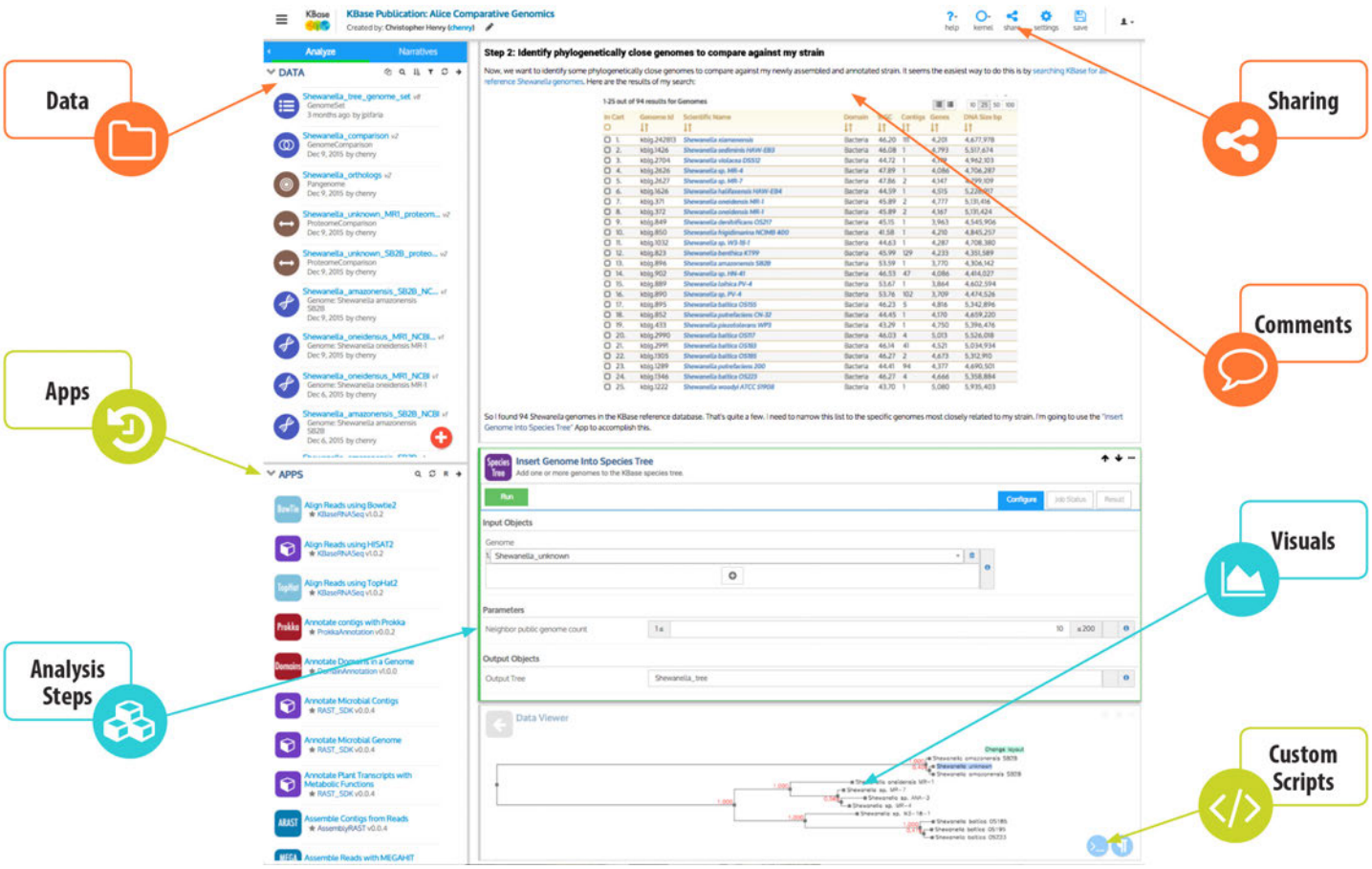


Fig. 2. Narrative Interface.

What Can You Do in KBase?

- Use sequence reads to quickly generate assembled and annotated genomes and metabolic models.
- Compare and contrast growth phenotypes under hundreds of conditions.
- Explore the comparative genomic organization, phylogeny, and gene content of organisms interactively.
- Identify the enriched species and functions among sets of metagenomes.

Reasons to Use KBase

- Enables collaboration and transparent sharing of results.
- Puts results in the context of knowledge in the field.
- Lowers the barrier for analyzing complex datasets.
- Gives users credit for their work and control over how it is shared.
- Integrates diverse biological data from many sources.
- Provides access to enterprise-class computing.

Open Source

Nearly all KBase software is publicly available through GitHub (www.github.com/kbase), where it can be reviewed and extended by the community.

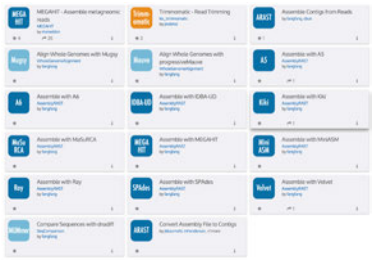
KBase Tools for Data Analysis

kbase.us/apps/

In KBase, data analysis is driven by a variety of apps that can be explored using the **App Catalog** (narrative.kbase.us/#appcatalog). The catalog provides advanced options for browsing apps; designating them as favorites; and filtering them based on analysis type, popularity, and more.

In the future, the App Catalog will contain not only the tools generated by KBase staff, but also those created and contributed by outside developers. Paving the way for these third-party contributions is KBase's new **software development kit (SDK)**, which provides a mechanism for users to add their own open-source, open-license apps to the system (kbase.us/developer). Below are summaries of available apps (some of which are in development).

Microbial Genome Assembly

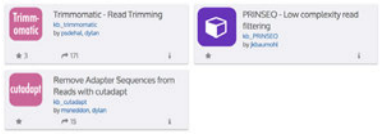


Choose various apps to assemble short, long, pair-end, and single-end reads into contig sets. Assemblers include A5, A6, Kiki, MaSuRCA, MiniASM, Ray, SPAdes, and Velvet.

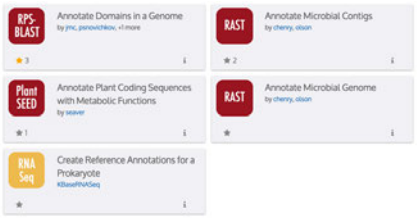
Use the *Assemble Contigs from Reads* app to compare several assemblies and design workflows with customized quality filtering, trimming, error correction, adapter removal, assembly, scaffolding, and post-processing.

Read Processing

Processing short reads enables generating higher quality *de novo* assemblies for more accurate downstream analysis. Perform adapter removal, quality filtering, and read trimming with these apps.



Annotation



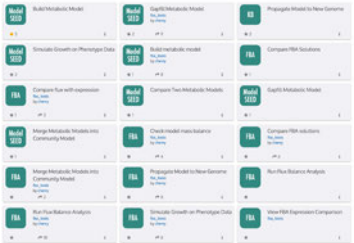
For microbes, functionally annotate contigs or genomes based on the RAST toolkit. For plants, assign metabolic functions derived from PlantSEED to cDNA or protein sequences. These annotations prepare the data for downstream KBase analyses, such as metabolic modeling.

Other apps assign protein domains to sequences from libraries such as COG and PFAM and create reference annotations in gene transfer format for a prokaryotic genome.

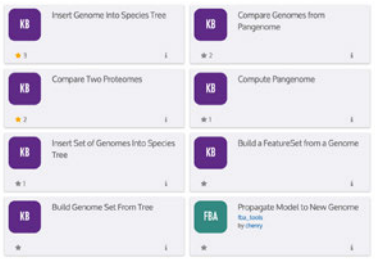
Metabolic Modeling

Build genome-scale metabolic models for subsequent detailed analysis of a plant or microbe's metabolic potential using apps for gapfilling, flux balance analysis (FBA), model comparison and editing, simulation of growth phenotypes, and more.

Translate a model to another organism, compare FBA solutions, edit or create media, and compare reaction fluxes with gene expression.



Comparative Genomics

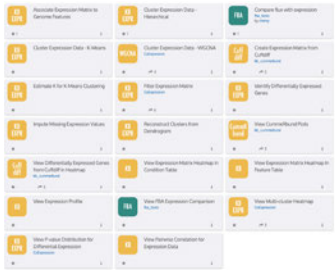


Using one or more genomes, build a phylogenetic species tree to examine closely related organisms. Extract a feature set from the tree for use in other analyses.

Additional apps will compare two proteomes, compute a pangenome for a genome set, and then compare isofunctional and homologous gene families for all the constituent genomes.

Expression

Quantify gene expression from RNA-Seq reads using a series of apps based on the Tuxedo workflow (e.g., Bowtie, TopHat, Cufflinks, Cuffmerge, Cuffdiff, and CummeRbund).



Use other tools to (1) impute missing expression values, (2) perform k-means or hierarchical clustering for observing and analyzing gene expression patterns, (3) associate an expression matrix with an annotated genome, (4) filter the matrix to show differentially expressed genes, and (4) compare average expression values for features in a matrix. Visualize expression data in sortable heatmaps, dendrograms, tables, and more.

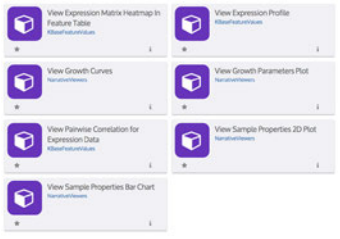
Communities

Assemble metagenomic reads from next-generation sequencing technologies using MEGAHIT or IDBA-UD apps.

Merge two or more metabolic models into a compartmentalized microbial community model capable of predicting flux profiles, trophic interactions, and nutrient consumption and production for the entire community.

...And More!

As KBase scientists and third-party developers integrate new tools into the system, our collection of apps continues to grow!



Data in KBase

KBase provides a single comprehensive resource that enables users to analyze their own data as well as a wide range of public bioinformatics data. KBase supports a variety of data types including:

- Sequence reads and assemblies
- Annotations
- Genomes and genome features
- Metabolic models
- Media
- Biochemistry
- Metabolic pathways
- Pangenomes
- Species trees
- Taxonomic and functional profiles

Users can import their own reads, genomes, plant transcripts, media, flux balance analysis models, phenotype sets, and expression matrices for analysis. KBase also enables users to transfer microbial genome reads from the DOE Joint Genome Institute (JGI) to their KBase accounts with the push of a button. Efforts are under way to support additional data types from both public sources and users.

A distinguishing feature of KBase is its data model, which integrates diverse biological datasets and represents them as meaningfully rich types that describe relationships among data. This integration enables comparison across domains and interoperability with both standard and next-generation tools. With the data model, KBase also can add value to new information, treating it as additional evidence that enriches the understanding of previously acquired data.

Public Data Sources in KBase
kbase.us/data-policy-and-sources/

Search KBase Reference Data
narrative.kbase.us/functional-site/#/search/
(KBase login not required)

Getting Started

To begin exploring and using KBase, go to kbase.us and click "Get started" (see Fig. 3). From here, you can register for a user account, access a Quick Start guide, and learn more about features such as Narratives, apps, and KBase data.

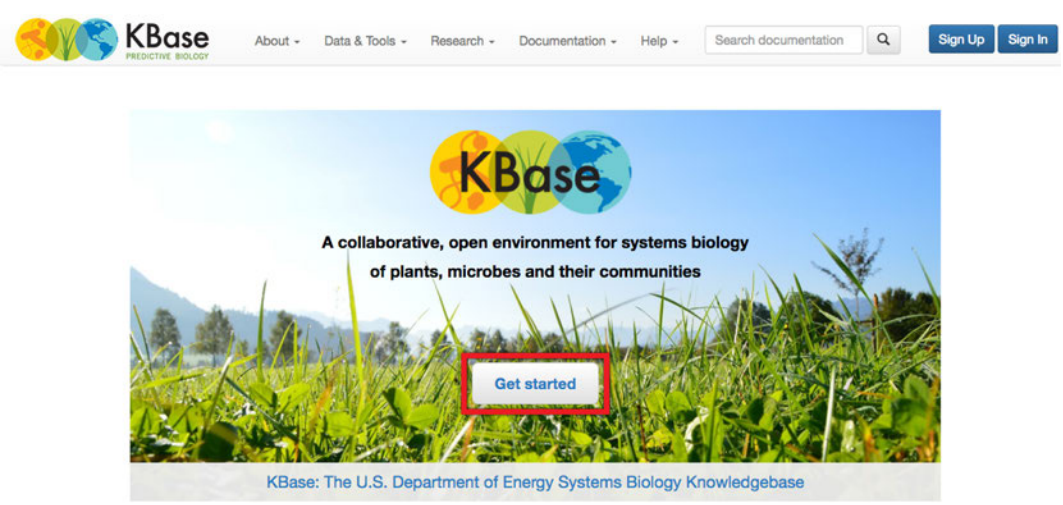


Fig. 3. Getting started from kbase.us.

Guides and Tutorials

Narrative Interface Quick Start:
kbase.us/narrative-quick-start/

Narrative Interface User Guide:
kbase.us/narrative-guide/

Data Search:
kbase.us/data-search-guide/

Data Upload and Download Guide: kbase.us/data-upload-download-guide/

Transfer JGI Data to KBase:
kbase.us/transfer-jgi-data/

Conferences, User Meetings, and Workshops

Visit the KBase booth at the Plant and Animal Genome, American Society for Microbiology, and American Society of Plant Biologists conferences.

Learn more at KBase user and developer workshops and the annual KBase user meeting, held jointly with the DOE Joint Genome Institute.

Calendar of Events:
kbase.us/events/

Social Media

KBase Blog:
kbase.us

Twitter:
twitter.com/DOEKBase/

Questions and Support

For more information or assistance, visit:
kbase.us/contact-us/

U.S. DEPARTMENT OF ENERGY
Office of Science

February 2017