Title

KBase: The United States Department of Energy Systems Biology Knowledgebase

Journal

Nature Biotechnology, 36(7)

ISSN

1087-0156

Authors

Arkin, Adam P
Cottingham, Robert W
Henry, Christopher S
et al.

Publication Date

2018-08-01

DOI

10.1038/nbt.4163

Peer reviewed

KBase: The United States Department of Energy Systems Biology Knowledgebase

To the Editor:

Over the past two decades, the scale and complexity of genomics technologies and data have advanced from sequencing genomes of a few organisms to generating metagenomes, genome variation, gene expression, metabolites, and phenotype data for thousands of organisms and their communities. A major challenge in this data-rich age of biology is integrating heterogeneous and distributed data into predictive models of biological function, ranging from a single gene to entire organisms and their ecologies. The US Department of Energy (DOE, Washington, DC) has invested substantially in efforts to understand 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 recognizes the need for scientists of diverse backgrounds to have access to sophisticated computational tools that enable them to analyze complex and heterogeneous data sets and integrate their data and results effectively with the work of others. In this way, new data and conclusions could be rapidly propagated across existing, related analyses and easily discovered by the community for evaluation and comparison with previous results[1-3].

Here we present the DOE Systems Biology Knowledgebase (KBase, http://kbase.us), an open-source software and data platform that enables data sharing, integration and analysis of microbes, plants and their communities. KBase maintains an internal reference database that consolidates information from widely used external data repositories. This includes over 90,000 microbial genomes from RefSeq[4], over 50 plant genomes from Phytozome[5], over 300 Biolog media formulations[6], and >30,000 reactions and compounds from KEGG[7], BIGG[8], and

MetaCyc[9]. These public data are available for integration with user data where appropriate (e.g., genome comparison or building species trees). KBase links these diverse data types with a range of analytical functions within a web-based user interface. This extensive community resource facilitates large-scale analyses on scalable computing infrastructure and has the potential to accelerate scientific discovery, improve reproducibility and foster open collaboration.

Although similar integrative tools exist (**Supplementary Note 2**) no other open platform shares all KBase's features, which include the following: (i) comprehensive support for data provenance and analysis reproducibility; (ii) a flexible system for sharing data and workflows; (iii) an integrated database of genomes and biochemistry; (iv) a point-and-click interface that enables users to build, store, run, and share complex scientific analyses of fully integrated data; (v) built-in support for the use of custom code interleaved with point-and-click apps; and (vi) a software development kit that enables external developers to add applications to KBase. KBase has a suite of scientific applications that enable users to build and share sophisticated workflows. For example, a user can 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. KBase supports numerous branch points, alternative pipelines, alternative entry points, and internal curation loops that facilitate a wide range of scientific analyses, some of which are not available elsewhere (e.g., merging individual metabolic models into community models and using these to predict interspecies interactions). Although KBase was developed to support analysis of microbes, plants and their communities, it is potentially applicable to any area of science (aside from projects that require HIPAA compliance).

KBase's primary user interface, the Narrative Interface, provides a user experience distinct from other analysis platforms available today, although it shares some common features

with a few other systems (**Supplementary Note 2**). From this interface, which is built on the Jupyter[10, 11] platform, users can upload their private data, search and retrieve extensive public reference data, access data shared by others, share their data with others, select and run applications on their data, view and analyze the results from those applications and record their thoughts and interpretations along with the analysis steps. These activities take place within a point-and-click 'notebook' environment (Fig. 1). When a user begins a new computational experiment in KBase, they create a new "notebook" (referred to as a 'Narrative' in KBase) to hold this experiment. Every action performed by a user appears as a 'cell' in the Narrative. App cells show the chosen input parameters for the application and the results of the analysis. Markdown cells allow users to add formatted text and figures to a Narrative to describe the thought process behind the scientific workflow being crafted.

A finished Narrative is a precise record of everything the authors did to complete their analysis. Although Narratives are private by default, users may choose to make their Narrative public, or share it with other individual users. This recording of a user's KBase activities within a sharable Narrative is a central pillar of KBase's support for reproducible transparent research (**Supplementary Note 1**). Once a Narrative has been shared or made public, other users can copy the Narrative and rerun it on their own data, or modify it to suit their scientific needs. Thus, public Narratives serve as resources for the user community by capturing valuable data sets, associated computational analyses, and scientific context describing the rationale behind a scientific study in a form that is immediately reproducible and reusable. A growing number of public Narratives are available in KBase, some of which are showcased in the Narrative Library (kbase.us/narrative-library/).

The data model in KBase is fundamental to supporting reproducibility and collaboration. KBase is built upon an object-oriented data model where each object instance is automatically

versioned and linked to provenance information describing how it was generated. Each data object is also associated with the specific Narrative in which it was uploaded or generated. When a Narrative is shared or copied, all its input and output data is shared or copied with it. Currently supported data types include reads, contigs, genomes, metabolic models, growth media, RNA-seq, expression, growth phenotype data, and flux balance analysis solutions. This set of types can be extended to support new apps and functionality.

Many existing systems (**Supplementary Note 2**) provide similar support for object-level sharing and provenance, but these systems operate on raw files only, without integration into a common data model. In KBase, objects are not simple files—they are explicitly defined and validated data structures, within which associated objects are linked to one another. For example, a metabolic model is linked to its associated genome, which is linked to its associated taxonomy. This data model enhances interoperability by requiring apps to operate on a common data representation. Furthermore, it enhances awareness of interdependence so users could be notified when an object on which an analysis is based has been updated and it will ultimately enable data discovery and meta-analysis across the KBase platform.

Presently, KBase has over 160 apps (narrative.kbase.us/#appcatalog) offering 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 Supplementary Note 2 for references). Apps interoperate seamlessly to enable a range of scientific workflows (**Fig. 2**). For reproducibility, all apps in KBase are containerized in versioned Docker modules, enabling a user to run any version at any time.

In addition to running apps, users can create and run blocks of code within a Narrative using “code cells.” KBase has an application programming interface (API) that allows users to call any KBase app programmatically from within these code cells. This enables users to, for example, run large-scale studies in KBase (e.g., building thousands of models at once) by using loops within a code cell (**Supplementary Note 1**). Users can also leverage the flexibility of code cells to add custom analysis steps that are not yet available as KBase apps.

Although there are other systems that allow users to create workflows consisting of a series of analysis tool runs and code blocks, the app functionality in KBase differs from these systems in several ways (**Supplementary Note 2**). Currently, KBase’s capabilities for community model reconstruction, plant model reconstruction, community model gapfilling, and expression data model integration are unique to the KBase platform.

KBase was designed to be an extensible community resource. This extensibility is supported by the KBase Software Development Kit (SDK), which is a set of command-line tools and a web interface that enable any developer to build, test, register, and deploy new or existing software as KBase apps, thereby extending the platform's scientific capabilities. All software contributed to the central KBase software repository must adhere to a standard open-source license (opensource.org/licenses). Information about the app developer is maintained in the documentation for that app so credit can be given to the contributor. Data provenance, job management, usage logging, and app versioning are handled automatically by the platform, allowing developers to wrap new scientific tools quickly with minimal KBase-specific training. Other existing platforms offer similar support for third-party development (**Supplementary Note 2**), but KBase’s data model provides the additional benefit of improving interoperability of third-party applications by imposing a single data format and specification on all data types

consumed or produced by each app. More information about the KBase SDK is available in the supplement and at https://github.com/kbase/kb_sdk/blob/master/README.md.

Many users have already discovered and applied KBase to meet their scientific needs. As of September 2017, over 3000 users have KBase accounts, and users have created over 5000 Narratives. These Narratives contain a total of over 250,000 data objects, or an average of 96 data objects and five apps per Narrative. Science done within KBase has been published in over 30 peer-reviewed publications (**Supplementary Note 1**; http://kbase.us/publications), including reconstruction of >8000 models of core metabolism across the microbial tree of life[12]; reconstruction of semi-curated metabolic models for 773 gut microbes[13]; predicting trophic interactions within a microbial community[14]; and reconstruction of regulons from expression data[15].

Much of the research performed within KBase has been publicly shared as Narratives that any user can view, copy, and re-run. Through these public Narratives, scientists can rapidly follow the examples set by their peers to apply similar approaches to new data and scenarios. Thus KBase goes beyond supporting reproducible science to enable rapid re-purposing, re-application, and extension of scientific techniques. As more users apply the system to address their scientific questions, and share their resulting Narratives, KBase will have a continually growing body of experiments, results and scientific use cases that can be adapted and extended by other researchers.

KBase's integration of data and tools and the ease of creating and running large-scale analysis workflows have the potential to empower scientists in a broad range of application areas for systems biology, including environmental analysis, biosystems design, and human

health. KBase's sharing capabilities amplify this potential by enabling scientists with differing expertise to easily work together and leverage each other's work.

Future development of KBase will build upon the concept of KBase as a knowledgebase. The social aspects of the platform will be enhanced, enabling scientists to discover colleagues with complementary talents. New 'data-discovery' features will allow the platform to suggest datasets and Narratives that may be of interest to a particular user based on interconnections found in the data in KBase. These features will ultimately evolve into 'knowledge-discovery' features, enabling KBase to propose new hypotheses by making connections across the system.

Adam P Arkin[1,2], Robert W Cottingham[3], Christopher S Henry[4], Nomi L Harris[2], Rick L Stevens[5,6], Sergei Maslov[7,24], Paramvir Dehal[2], Doreen Ware[8], Fernando Perez[9-11,25], Shane Canon[12], Michael W Sneddon[2], Matthew L Henderson[2], William J Riehl[2], Dan Murphy-Olson[4], Stephen Y Chan[2], Roy T Kamimura[2], Sunita Kumari[8], Meghan M Drake[3], Thomas S Brettin[6], Elizabeth M Glass[4], Dylan Chivian[2], Dan Gunter[9], David J Weston[3], Benjamin H Allen[3], Jason Baumohl[2], Aaron A Best[13], Ben Bowen[2], Steven E Brenner[14], Christopher C Bun[4], John-Marc Chandonia[2], Jer-Ming Chia[8], Ric Colasanti[4], Neal Conrad[6], James J Davis[6], Brian H Davison[3], Matthew DeJongh[15], Scott Devoid[4], Emily Dietrich[6], Inna Dubchak[2], Janaka N Edirisinghe[4,16], Gang Fang[17,26], José P Faria[4], Paul M Frybarger[4], Wolfgang Gerlach[4], Mark Gerstein[17], Annette Greiner[12], James Gurtowski[8], Holly L Haun[3], Fei He[7,27], Rashmi Jain[18,19], Marcin P Joachimiak[2], Kevin P Keegan[4], Shinnosuke Kondo[15], Vivek Kumar[8], Miriam L Land[3], Folker Meyer[4], Marissa Mills[3], Pavel S Novichkov[2], Taeyun Oh[18,19,28], Gary J Olsen[20], Robert Olson[4], Bruce Parrello[4], Shiran Pasternak[8], Erik Pearson[2], Sarah S Poon[9], Gavin A Price[2], Srividya Ramakrishnan[8,29], Priya Ranjan[3,21], Pamela C Ronald[18,19], Michael C Schatz[8,29], Samuel M D Seaver[4], Maulik Shukla[6], Roman A Sutormin[2], Mustafa H Syed[3,30], James Thomason[8], Nathan L Tintle[22,31], Daifeng Wang[17,32], Fangfang Xia[6], Hyunseung Yoo[6], Shinjae Yoo[23], Dantong Yu[23,33]

Affiliations

[1]Department of Bioengineering, University of California, Berkeley, California, USA.

[2]Environmental Genomics and Systems Biology Division, E. O. Lawrence Berkeley National Laboratory , Berkeley, California, USA.

[3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.

[4]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA.

[5]Computer Science Department and Computation Institute, University of Chicago, Chicago, Illinois, USA.

[6]Computing, Environment, and Life Sciences Directorate, Argonne National Laboratory, Argonne, Illinois, USA.

[7]Biology Department, Brookhaven National Laboratory, Upton, New York, USA.

[8]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA.

[9]Computational Research Division, E. O. Lawrence Berkeley National Laboratory, Berkeley, California, USA.

[10]Berkeley Institute for Data Science, University of California, Berkeley, California, USA.

[11]Department of Statistics, University of California, Berkeley, California, USA.

[12]National Energy Research Scientific Computing Center, E. O. Lawrence Berkeley National Laboratory, Berkeley, California, USA.

[13]Department of Biology, Hope College, Holland, Michigan, USA.

[14]Department of Plant and Microbial Biology, University of California, Berkeley, California, USA.

[15]Department of Computer Science, Hope College, Holland, Michigan, USA.

[16]Computation Institute, University of Chicago, Chicago, Illinois, USA.

[17]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA.

[18]Department of Plant Pathology and Genome Center, University of California, Davis, Davis California, USA.

[19]Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA.

[20]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

[21]Department of Plant Sciences, University of Tennessee, Knoxville, Tennessee, USA.

[22]Department of Mathematics, Hope College, Holland, Michigan, USA.

[23]Computer Science and Math, Computer Science Initiative, Brookhaven National Laboratory, Upton, New York, USA.

Present Addresses

[24]Department of Bioengineering and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA.

[25]Department of Statistics, University of California, Berkeley, California, USA.

[26]New York University Shanghai Campus, Pudong, Shanghai, China.

[27]Department of Plant Pathology, Kansas State University, Manhattan, Kansas, USA.

[28]Insilicogen. Inc., Giheung-gu, Yongin-si, Gyeonggi-do, Korea.

[29]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA.

[30]Memorial Sloan Kettering Cancer Center, New York, New York, USA.

[31]Dordt College, Sioux Center, Iowa, USA.

[32]Department of Biomedical Informatics, Stony Brook University, Stony Brook, New York, USA.

[33]Martin Tuchman School of Management, New Jersey Institute of Technology, Newark, New Jersey, USA.

Corresponding author

Correspondence and requests for materials should be addressed to: APArkin@lbl.gov.

Code Availability

The KBase code, available at github.com/kbase, is open source and freely distributed under the MIT License. The web-accessible KBase system (narrative.kbase.us) is run on DOE computing infrastructure and is freely available for anyone to use. KBase adheres to the FAIR (Findable,

Accessible, Interoperable, Re-usable) data principles endorsed by many funding agencies and scientific organizations[16].

Data Availability

All data generated or analyzed during this study are included in this published article and Supplementary Note 1 as links to the original work, or in the associated KBase Narratives linked here. An earlier version of this paper was published as a preprint[17].

Acknowledgements

This work is supported by the Office of Biological and Environmental Research's Genomic Science program within the U.S. Department of Energy Office of Science, under award numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.

Editors Note: This manuscript has been peer reviewed.

Competing financial interests

FP declares competing financial interest related to his work for Plot.ly and research funding from Microsoft, Google, and Anaconda Inc.

SEB receives funding and has a research collaboration with Tata Consultancy Service that is unrelated to the KBase project.

All other authors declare no competing financial interests.

Author Contributions

APA, RWC, CSH, RLS, SM, PD, DW and FP developed the concept and vision.

APA, CSH, RLS, SC, MWS, MLH, WJR, DMO, SYC, TSB, DC, DG, JB, AAB, BPB, SEB, CCB, JMC, JC, RC, NC, JJD, MDJ, SD, AG, FH, MPJ, KPK, FM, PSN, RO, EP, SP, GAP, SR, PR, SMDS, MS, RAS, MHS, JT, FX, HY, SJY and DY designed and developed the system.

RWC, NLH, RTK, SK, MMD, EMG, DC, DJW, BHA, BHD, ED, ID, JNE, GF, JPF, PMF, WG, MG, JG, RJ, SNK, VK, MLL, MM, TYO, GJO, BP, SSP, PCR, MCS, NLT and DFW developed, documented and conducted testing and validation.

APA, CSH and NLH drafted the manuscript.

NLH, HLH, BHA, MMD, MPJ, AAB, JMC, DC, RO, BHD, NLT, SM, PCR, MDJ and VK revised the manuscript and provided important intellectual content.

JB, MPJ, JMC, VK, JNE, JPF, SMDS provided content for the supplemental material.

APA, RWC CSH and NLH reviewed and approved the final version to be published.

References

1. Stodden, V. et al. Enhancing reproducibility for computational methods. *Science* **354**, 1240-1241 (2016).
2. Prlic, A. & Procter, J.B. Ten simple rules for the open development of scientific software. *PLoS Comput Biol* **8**, e1002802 (2012).
3. Millman, K.J. & Pérez, F. in Developing Open-Source Scientific Practice. *Implementing reproducible research*. (eds. F.L.V. Stodden & R.D. Peng) 149-183 (CRC Press, Boca Raton, FL; 2014).
4. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65 (2007).
5. Goodstein, D.M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**, D1178-1186 (2012).
6. Bochner, B.R. Global phenotypic characterization of bacteria. *Fems Microbiology Reviews* **33**, 191-205 (2009).
7. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**, 27-30 (2000).
8. Schellenberger, J., Park, J.O., Conrad, T.M. & Palsson, B.O. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213.
9. Caspi, R. et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research* **34**, D511-D516 (2006).

10. Perez, F. & Granger, B.E. IPython: A system for interactive scientific computing. *Computing in Science & Engineering* **9**, 21-29 (2007).
11. Kluyver, T. et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87-90 (2016).
12. Edirisinghe, J.N. et al. Modeling central metabolism and energy biosynthesis across microbial life. *BMC Genomics* **17**, 568 (2016).
13. Magnusdottir, S. et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* **35**, 81-89 (2017).
14. Henry, C.S. et al. Microbial community metabolic modeling: A community data-driven network reconstruction. *J Cell Physiol* (2016).
15. Faria, J.P. et al. Computing and Applying Atomic Regulons to Understand Gene Expression and Regulation. *Front Microbiol* **7**, 1819 (2016).
16. Wilkinson, M.D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
17. Arkin, A.P. et al. The DOE Systems Biology Knowledgebase (KBase). *bioRxiv* **preprint first posted online Dec. 22, 2016** (2016).

Figure Legends

Figure 1. KBase Narratives. A Narrative is an interactive, dynamic, and persistent document created by users that promotes open, reproducible, and collaborative science.

Figure 2. Major workflows and data types in KBase. The unboxed labels represent data types, while each colored box represents a single app. The box colors signify the category of functionality, and the numbers in parentheses indicate the number of alternative apps that implement each function. Apps that require a genome data type as input are marked with a green ‘G’ icon. For more information see http://kbase.us/apps/.