1 **The DOE Systems Biology Knowledgebase (KBase)**

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

17 **Abstract**
18 The U.S. Department of Energy Systems Biology Knowledgebase (KBase) is an open-source
19 software and data platform designed to meet the grand challenge of systems biology—
20 predicting and designing biological function from the biomolecular (small scale) to the ecological
21 (large scale). KBase is available for anyone to use, and enables researchers to collaboratively
22 generate, test, compare, and share hypotheses about biological functions; perform large-scale
23 analyses on scalable computing infrastructure; and combine experimental evidence and
24 conclusions that lead to accurate models of plant and microbial physiology and community
25 dynamics. The KBase platform has (1) extensible analytical capabilities that currently include
26 genome assembly, annotation, ontology assignment, comparative genomics, transcriptomics,
27 and metabolic modeling; (2) a web-browser-based user interface that supports building, sharing,
28 and publishing reproducible and well-annotated analyses with integrated data; (3) access to
29 extensive computational resources; and (4) a software development kit allowing the community
30 to add functionality to the system.

31 **Introduction**
32 Over the past two decades, the scale and complexity of genomics technologies and data have
33 advanced from simple genomic sequences of only a few organisms to metagenomes and
34 genome variation, gene expression, metabolite, and phenotype data for thousands of organisms
35 and their communities. A major challenge in this data-rich age of biology is integrating
36 heterogeneous, distributed, and error-prone primary and derived data into predictive models of
37 biological function ranging from a single gene to entire organisms and their ecologies. To
38 develop models of these biological processes, organisms and their interactions, scientists of
39 diverse backgrounds need, at a minimum, the ability to use a variety of sophisticated
40 computational tools to analyze their own complex and heterogeneous data sets, and then

41 integrate their data and results effectively with the work of others. Such integration requires
42 discovering and accessing others' information, understanding its source and limitations, and
43 using tools to perform additional analyses upon it. Ideally, new data and conclusions would be
44 rapidly propagated across existing, related analyses and easily discovered by the community for
45 evaluation and comparison with previous results[1-3].

46 Nowhere are the barriers to discovery, characterization, and prediction more formidable than in
47 efforts to understand the complex interplay between biological and abiotic processes that
48 influence soil, water, and climate dynamics and impact the productivity of our biosphere. A
49 bewildering diversity of plants, microbes, animals, and their interactions needs to be discovered
50 and characterized to mechanistically understand ecological function and thereby facilitate
51 interventions to improve outcomes. The U.S. Department of Energy (DOE) has invested in
52 various large- and small-scale programs in climate and environmental science and biological
53 system science. These efforts have demonstrated the power of integrated science programs to
54 make progress on complex systems. However, the community that has grown around these
55 efforts has recognized the need to lower the barrier to accessing tools, data, and results, and to
56 work collaboratively to accelerate the pace of their research[4].

57 The DOE Systems Biology Knowledgebase (KBase, www.kbase.us) is a software platform
58 designed to provide these needed capabilities. Specifically, KBase seeks to make it easier for
59 scientists to create, execute, collaborate on, and share sophisticated reproducible analyses of
60 their own biological data in the context of public and other users' data. Results and conclusions
61 can be shared with individuals or published within KBase's integrated data model that will
62 increasingly support user-driven and automated meta-analysis.

63 While a number of recent computational environments address different aspects of this
64 challenge (see *Comparison with other Platforms*), they are generally decentralized, limiting the
65 extent to which data and workflows can be integrated, shared, and extended across the
66 scientific community. Moreover, there is minimal support for key capabilities such as more
67 iterative scientific analysis, in-depth collaboration, integration of new results in the context of
68 others' results, and automatic propagation of new findings that may inform research across
69 disciplines.

70 KBase users have already applied the system to address a range of scientific problems,
71 including comparative genomics of plants, prediction of microbiome interactions, and deep
72 metabolic modeling of environmental and engineered microbes. KBase currently supports a
73 growing and extensible set of applications or "apps" for genome assembly, annotation,
74 metabolic model reconstruction, flux balance analysis, expression analysis, and comparative
75 genomics. In addition to these tools, the KBase platform provides data integration and search,
76 along with easy access to shared user analyses of public plant and microbial reference data
77 from a number of external resources including National Center for Biotechnology Information
78 (NCBI) and the DOE Joint Genome Institute (see *KBase Data Model and Apps*). As the platform
79 matures and is adopted more widely, the data, analysis tools, and computational experiments
80 contributed by users are also expected to increase, leading to wider biological applications with
81 richer and more sophisticated support for functional prediction and comparison.

82   Based on a service-oriented design, KBase is built to run primarily on a cloud-computing
83   infrastructure, although high-performance computing (HPC) resources are only now being
84   integrated. The platform is completely open source, with all core infrastructure and service code
85   available on GitHub (https://github.com/kbase). The central instance of KBase that serves
86   analysis and modeling of plants, microbes, and their communities is maintained and run on
87   DOE enterprise computing resources. This resource is open and free for anyone to use.

88   **KBase Narratives and User Interface**
89   KBase's graphical user interface, the Narrative Interface, supports both point-and-click and
90   scripting access to system functionality in a "notebook" environment, enabling computational
91   sophisticates and experimentalists to easily collaborate within the same platform. Built on the
92   Jupyter Notebook[5], the interface allows researchers to design, carry out, record and share
93   computational experiments in the form of *Narratives*—dynamic, interactive documents that
94   include all the data, analysis steps, parameters, visualizations, scripts, commentary, results, and
95   conclusions of an experiment (Fig. 1).

96

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

99   Although private by default, users can choose to share their Narratives and data with select
100 collaborators or even publicly. Sharing among collaborators facilitates communication, reuse
101 and management of scientific projects. Public Narratives serve as critical resources for the

102 broader user community by capturing valuable data sets, associated computational analyses,
103 and enriched scientific context describing the rationale and design of original experiments, data
104 upload and organization, and the application of various analytical techniques, complete with the
105 selected parameters and interpretation of results. Narrative sharing and publication are key
106 capabilities enabling other scientists to quickly view, copy, replicate, and expand on work
107 performed within KBase. A growing number of public Narratives are available in KBase today,
108 several of which are described in detail in the "Science Performed within the KBase Platform"
109 section. A selection of these is also available in the KBase Narrative Library
110 (www.kbase.us/narrative-library).

111 Because Narratives are built upon the Jupyter Notebook framework, users can create and run
112 scripts within a Narrative using a "code cell." KBase is building a code cell application
113 programming interface (API) enabling users to run KBase apps programmatically from within the
114 system. Users also can leverage the flexibility of code cells to incorporate custom analysis steps
115 into their Narratives not yet available as KBase apps.

KBase Data Model and Apps

117 KBase provides a seamless amalgamation of (1) curated and periodically updated reference
118 data, (2) uploaded private and shared user data (and derived data products), and (3) a growing
119 compendium of apps covering a wide range of analytical capabilities. KBase's reference
120 database includes all public genome sequences from RefSeq[6] and Phytozome[7] and selected
121 plant genomes from Ensembl[8]. These genomes are maintained with their original gene IDs and
122 annotations, along with updated gene calls and annotations provided by the KBase annotation
123 pipeline. The reference collection also contains biochemistry data, including 27,692 biochemical
124 compounds, 34,705 reactions, and 529 growth media formulations, and integrates many
125 prominent ontologies such as GO[9], SEED Subsystems[10], PFam[11], and Interpro[12]. All reference
126 data are periodically updated from source databases and are available for integration with user
127 data where appropriate (e.g., when identifying isofunctional genes or building species trees).

128 KBase's faceted search utility enables users to query reference data by text, DNA sequence, or
129 protein sequence. Genomes and genes identified through searches can be copied to a
130 Narrative for deeper analysis. Ultimately, the search utility will be extended to query the user
131 data in KBase, while still ensuring appropriate data privacy. This capability will facilitate the
132 process by which users search their own data and data shared with them by collaborators.

133 In addition to reference data, KBase stores a wide range of uploaded and derived data sets
134 shared by users and connected to analyses performed in Narratives (Fig. 2A). Currently
135 supported data objects include reads, contigs, genomes, metabolic models, growth media,
136 RNA-seq, expression, growth phenotype data, and flux balance analysis solutions. This set of
137 data objects is fully extensible, and the links among them are expanded as new data sources
138 and apps are added to the platform. Third-party developers cannot yet extend the KBase data
139 model, but providing this capability is an important near-term goal.

140 KBase currently has over 70 released apps in production offering diverse scientific functionality
141 for genome assembly, genome annotation[13], sequence homology analysis, tree building[14],

142 comparative genomics, metabolic modeling[15], community modeling[16], gap-filling[17,18], RNA-seq
143 processing[19], and expression analysis[20]. In addition there are dozens of beta (pre-release) apps
144 available to try. Apps are engineered to interoperate seamlessly to enable a range of scientific
145 workflows (Fig. 2B). Current apps, production and beta, with their associated documentation are
146 listed in the KBase App Catalog (https://narrative.kbase.us/#appcatalog). KBase enables third-
147 party developers to add their own apps for use by the broad scientific community (see KBase
148 Software Development Kit section).
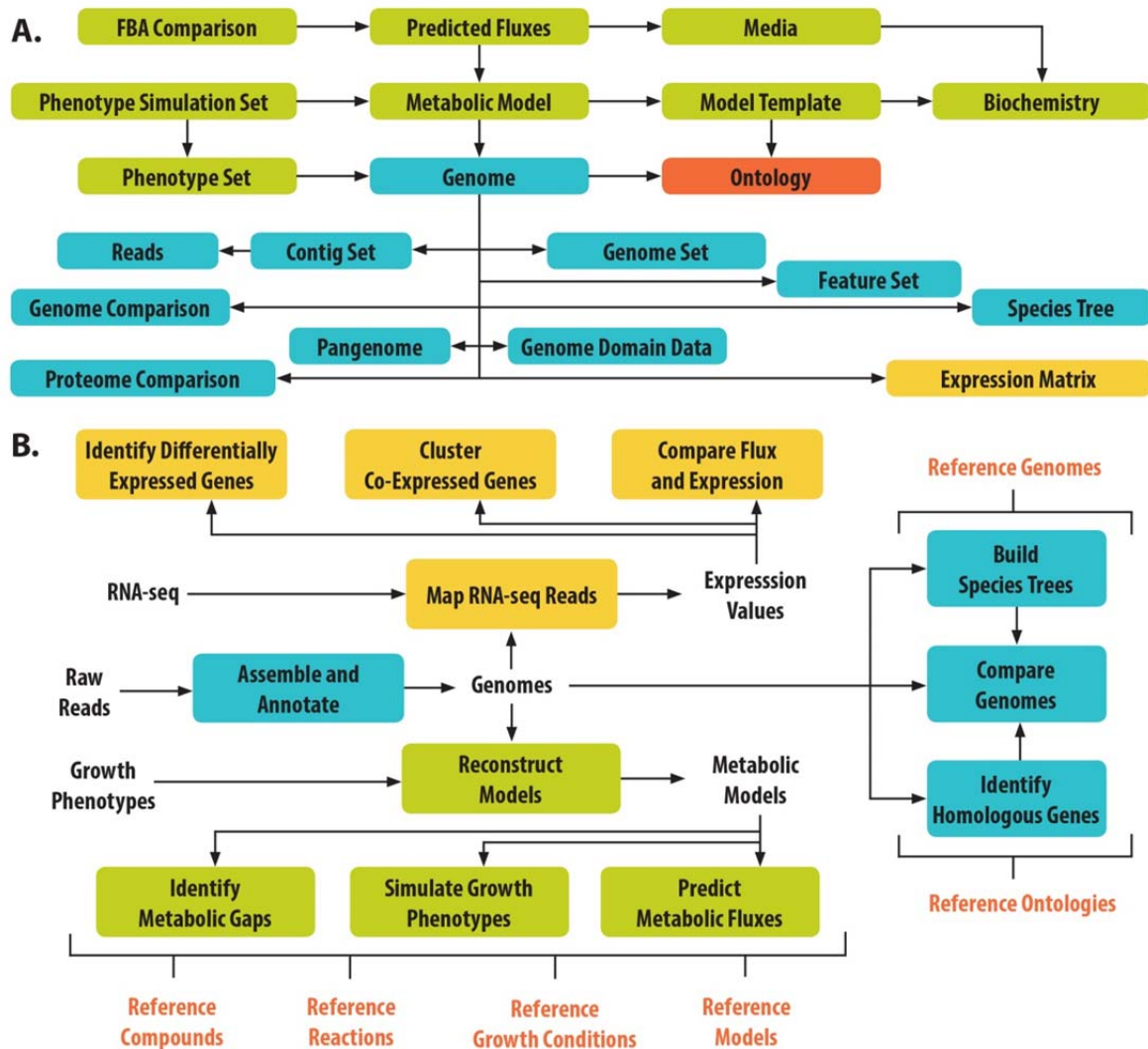
149
150 *Figure 2. Currently supported (A) data types and (B) workflows in the KBase platform. KBase integrates*
151 *data types and apps for analyzing and comparing genomes (blue), transcriptomes (orange), and*
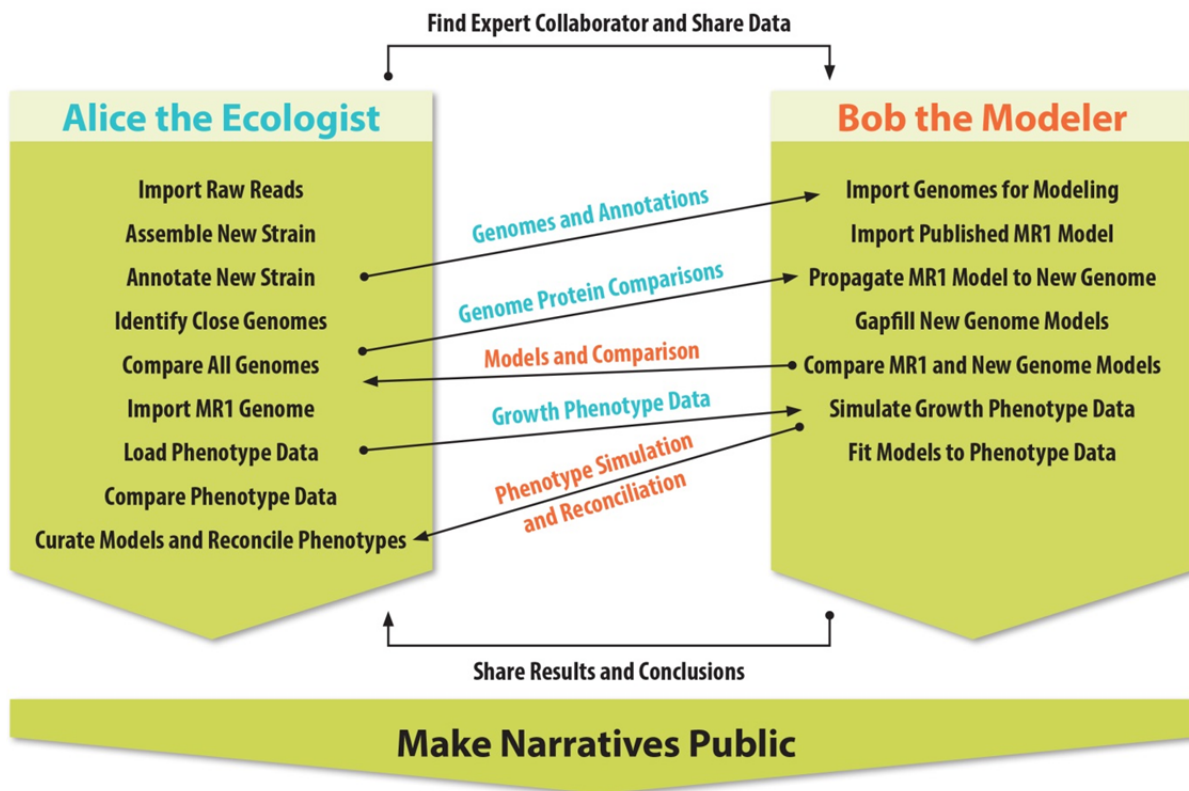152 *metabolic models (green).*

153 **Support for Reproducible, Interdisciplinary, Collaborative Science**
154 The KBase platform is designed from the ground up to comprehensively support reproducible,
155 interdisciplinary, and collaborative science. KBase's object-based data store, for example, holds

156 all user and reference data, and all objects, including Narratives themselves, have their own
157 provenance and version information. Provenance data capture when, how, and by whom each
158 object was created, including the apps or upload tools used and their associated parameters
159 and files. All apps and data are also versioned, so that even if a user overwrites an object or an
160 app is updated, the original information is recoverable. These provenance and reversibility
161 features ensure that all science performed within KBase is fully specified and reproducible,
162 averting the time-consuming effort often required to understand how published computational
163 studies or even simple data analyses were carried out.

164 Collaboration in KBase is supported by the ease with which all data in the system may be
165 shared and copied among users. A user may share any Narrative that they own (or have
166 administrative access to) with other KBase users simply by searching for their names or email
167 addresses. Importantly, when a user shares a Narrative, they also are sharing all the data
168 objects loaded, used, or generated within the Narrative, complete with versioning and
169 provenance. This feature allows users to share every aspect of the science that they have done
170 in KBase. Three levels of sharing are supported: (1) *read* sharing lets others see and copy, but
171 not edit the Narrative and its associated data; (2) *write* sharing expands these privileges to
172 include editing, enabling users to work together on a single Narrative; and (3) *admin* sharing
173 allows others to edit the Narrative and also share it with third parties. Moreover, any Narrative
174 can be made public by the user who owns it, which essentially gives all KBase users read
175 access to that Narrative and its underlying data. Users with read privileges for a Narrative can
176 create their own copy, which they own and can edit. This copying feature enables users to
177 quickly replicate and expand on any KBase Narrative shared with them. This approach to
178 sharing facilitates interdisciplinary science by allowing researchers with different expertise to
179 quickly and easily exchange data, results, methodologies, and workflows used to solve complex
180 biological problems.

181 We demonstrate how KBase facilitates sharing, collaboration, and interdisciplinary research with
182 a series of example Narratives (Fig. 3 and Table 1) from two scientists: Alice, a wet-lab biologist
183 with expertise in assembly, annotation, and comparative genomics, and Bob, a computational
184 biologist with expertise in metabolic modeling. (Before proceeding to view these Narratives, sign
185 up for a KBase User Account at www.kbase.us) In the first Narrative (Table 1, Narrative 1), Alice
186 uploads raw reads from a new strain of *Shewanella* that she is analyzing. She uses KBase to
187 assemble and annotate these reads, generating a new genome object in KBase. In a second
188 Narrative (Table 1, Narrative 2), Alice compares her new genome with other close strains of
189 *Shewanella*. She finds growth phenotype data for *Shewanella oneidensis* MR-1, which is
190 phylogenetically close to her strain[21]. This inspires Alice to run a growth phenotype array on her
191 own strain, which she also uploads to KBase. Alice then compares both phenotype arrays and
192 notices many differences that she cannot explain.

Find Expert Collaborator and Share Data

Alice the Ecologist

- Import Raw Reads
- Assemble New Strain
- Annotate New Strain
- Identify Close Genomes
- Compare All Genomes
- Import MR1 Genome
- Load Phenotype Data
- Compare Phenotype Data
- Curate Models and Reconcile Phenotypes

Bob the Modeler

- Import Genomes for Modeling
- Import Published MR1 Model
- Propagate MR1 Model to New Genome
- Gapfill New Genome Models
- Compare MR1 and New Genome Models
- Simulate Growth Phenotype Data
- Fit Models to Phenotype Data

Genomes and Annotations

Genome Protein Comparisons

Models and Comparison

Growth Phenotype Data

Phenotype Simulation and Reconciliation

Share Results and Conclusions

Make Narratives Public

193

194   ***Figure 3.*** *Example of collaboration in KBase. Two researchers collaborate using Narratives to reach more*
195   *complete scientific conclusions than either could have achieved alone.*

196

197   At this point, Alice contacts Bob, who suggests that metabolic models can help her analyze the
198   Biolog phenotype data. Alice shares her Narratives with Bob, who copies her genomes into a
199   new third Narrative. In this third Narrative (Table 1, Narrative 3), Bob loads a published model of
200   *S. oneidensis* MR-1[22], which he then propagates to Alice's genome. Bob compares the models,
201   identifying some interesting metabolic differences. Then Bob creates a fourth Narrative (Table 1,
202   Narrative 4), where he imports Alice's Biolog data and simulates the data with his *Shewanella*
203   models. He optimizes his models to fit the Biolog data and shares the results with Alice.

204   Finally, they build a fifth Narrative (Table 1, Narrative 5) together in which Alice refines Bob's
205   models by replacing gap-filled reactions with more biologically relevant selections, gaining a
206   complete understanding of the differences between her strain and MR-1. All data used in this
207   example are real: Alice's raw reads are from an existing genome, *Shewanella amazonensis*
208   SB2B[23,24], and the growth phenotype data are from an existing experimental study[25]. The
209   computational experiment carried out in the five Narratives results in the development and
210   validation of a new genome-scale metabolic model of *S. amazonensis* SB2B in KBase, as well
211   as the improvement of the existing model for *S. oneidensis* MR-1.

212 **Table 1**. Example workflows demonstrating collaboration using KBase. (Sign up for a KBase
213 User Account at www.kbase.us before clicking the links to view these Narratives. This table with
214 active links is also available online at www.kbase.us/kbase-paper.)
215

Narrative #	Title and URL
1	Alice Narrative 1: Assembly and Annotation (https://narrative.kbase.us/narrative/ws.18152.obj.1)
2	Alice Narrative 2: Comparative Genomics (https://narrative.kbase.us/narrative/ws.18153.obj.1)
3	Bob Narrative: Build Metabolic Models (https://narrative.kbase.us/narrative/ws.18155.obj.1)
4	Bob and Alice Narrative 1: Phenotype Data Analysis (https://narrative.kbase.us/narrative/ws.18156.obj.1)
5	Bob and Alice Narrative 2: Phenotype Data Reconciliation (https://narrative.kbase.us/narrative/ws.18157.obj.1)

216
217 This example demonstrates the novel, collaborative science that can be done in KBase using
218 the extensive portfolio of tools and data currently in the system. KBase provides some new
219 analysis tools (e.g., community modeling[16] and growth phenotype gap-filling) that are not yet
220 available in any other platform. KBase also makes accessible a number of third-party, open-
221 source tools with the goal of making them easier to use and more powerful when integrated with
222 existing tools in the broader KBase ecosystem. Equally important in this example is how
223 KBase's user interface facilitates a seamless collaboration between scientists with different but
224 complementary expertise who are able to accomplish more together than they could
225 individually. Finally, it is important to note how KBase reference data (such as genomes, media
226 conditions, and biochemistry) can be integrated into users' analyses, enhancing their work and
227 offering new perspectives.

228 **Science Performed Within the KBase Platform**
229 KBase launched the first version of the Narrative user interface in February 2015. Since then,
230 more than 1200 users have signed up for KBase accounts, and over 750 have created at least
231 one Narrative. Excluding KBase staff, users have built 2657 Narratives, 294 of which have been
232 shared with at least one other user and 37 have been made public. These Narratives contain
233 255,148 data objects, or an average of 96 data objects and five apps per Narrative. In these
234 Narratives, users have applied KBase to address a wide range of scientific questions. Here we
235 highlight seven peer-reviewed publications that cite publicly available Narratives in KBase,
236 where the bulk of the analysis described in the publication was performed. These and other
237 science Narratives can be found in KBase's Narrative Library, www.kbase.us/narrative-library.

238

239 **Table 2**. Example Narratives Demonstrating the Use of KBase (sign up for a KBase User
240 Account at www.kbase.us before clicking the links to view these Narratives.)
241

Narrative #	Title and URL
1	The PlantSEED Resource in KBase (https://narrative.kbase.us/narrative/ws.15250.obj.1)
2	Microbial Comparative Genomics: Reconstruction and Comparison of Core Metabolism Across Microbial Life (https://narrative.kbase.us/narrative/ws.15253.obj.1)
3	Community Modeling Protocol: Multi-Species Model Reconstruction and Analysis (https://narrative.kbase.us/narrative/ws.10824.obj.1)
4	BP1 Meio Community Metabolic Modeling (https://narrative.kbase.us/narrative/ws.13838.obj.1)
5	Electrosynthetic Microbiome: Electrosynthesis Community Model (https://narrative.kbase.us/narrative/ws.15248.obj.1)
6	Comparative analysis of phylogenetically close genomes and their associated growth phenotypes (https://narrative.kbase.us/narrative/ws.8773.obj.1)
7	Computing and Applying Atomic Regulons to Understand Gene Expression and Regulation (https://narrative.kbase.us/narrative/ws.14533.obj.1)

242
243 Major applications for KBase include reconstructing, comparing, and analyzing metabolic
244 models for diverse genomes. In 2015, the PlantSEED metabolic model reconstruction pipeline
245 was integrated into KBase and used to build genome-scale metabolic models of 10 diverse
246 reference plant genomes, which were subsequently compared (Table 2, Narrative 1)[26]. More
247 recently, KBase implemented a core model reconstruction pipeline that was applied to more
248 than 8000 microbial genomes (Table 2, Narrative 2)[27]. The resulting core models were
249 compared within a phylogenetic context, identifying how the pathways comprising core
250 metabolism are clustered across the microbial tree of life.

251 KBase also has a sophisticated, unique pipeline for microbiome modeling and analysis, as
252 demonstrated by three recent publications. One is a book chapter
253 (www.kbase.us/community-modeling/) that explores various microbiome modeling paradigms,
254 predicting potential interactions between the gut microbes *Bacteroides thetaiotaomicron* and
255 *Faecalibacterium prausnitzii* as a case study (Table 2, Narrative 3)[16]. Another analysis
256 examines interactions between an autotrophic carbon-fixing cyanobacteria,
257 *Thermosynechococcus elongatus* BP-1, and the heterotrophic gram-positive species,
258 *Meiothermus ruber* Strain A. Model predictions from this analysis were validated by a
259 combination of published growth-condition data and comparison of model-predicted fluxes with
260 metatranscriptome-based expression data (Table 2, Narrative 4)[28]. Finally, metagenomic and
261 metatranscriptomic data from an electrosynthetic microbiome were assembled into genomes,

262 models, and flux predictions for three dominant species in the microbiome (Table 2, Narrative
263 5)[29]. To our knowledge, no other platform is capable of performing the combination of
264 community modeling analyses and expression data comparisons demonstrated in these
265 featured Narratives.

266 KBase also has powerful comparative genomics tools, including newly developed apps for
267 analyzing phenotype data. These tools are highlighted in a recent publication in which two
268 phylogenetically close genomes are systematically compared, identifying how changes in gene
269 content result in changes in growth phenotypes[30] (Table 2, Narrative 6)[31]. In this analysis, an
270 existing metabolic model[32] is propagated to another strain of the same species. The models are
271 then applied to understand the differences between Biolog data gathered for both species,
272 including identifying growth conditions where additional experimental validation is needed. The
273 models are also applied to predict essential and nonessential metabolic genes, with validation
274 performed via comparison with a recently published TN-seq dataset[33].

275 In another recent publication leveraging KBase's range of expression tools, expression data are
276 loaded for five genomes: *Escherichia coli*, *S. oneidensis*, *Pseudomonas aeruginosa*, *Thermus
277 thermophilus*, and *Staphylococcus aureus* (Table 2, Narrative 7)[20]. KBase apps are used to
278 identify clusters of co-expressed genes in these genomes, and the KBase Software
279 Development Kit (SDK) is applied to add a new app for computing these clusters. This new app
280 is compared to the others, and co-expressed clusters are also compared across all five
281 genomes.

282 These peer-reviewed studies demonstrate KBase's wide range of capabilities and the power of
283 Narratives in disseminating complex computational analyses that are well documented, easily
284 reproducible, and extensible.

285 **KBase Software Development Kit**
286 Computational approaches to analyzing biological data are heterogeneous and can evolve
287 rapidly. Accordingly, the KBase platform must be able to integrate new software tools from
288 diverse sources with varying computational requirements while still maintaining a consistent
289 data model with data provenance and version history. This integration is accomplished using the
290 KBase SDK, a set of command-line tools and a web interface enabling any developer or
291 advanced user to build, test, register, and deploy new or existing software as KBase apps. To
292 facilitate transparent, reproducible, and open science, all software contributed to the central
293 KBase software repository must adhere to a standard open-source license
294 (https://opensource.org/licenses). Information about the app developer is maintained in the user
295 documentation for that app so that credit can be properly given to the contributor. Data
296 provenance, job management, usage logging, and app versioning are handled automatically by
297 the platform, allowing developers to contribute new scientific tools quickly with minimal KBase-
298 specific training.

299 Scientific computation in KBase is managed with a distributed Docker-based[34] execution model.
300 Docker allows the individual programs and system dependencies of each Narrative app to be
301 saved as an image and run securely and identically across physical resources within individual,

302   isolated environments called containers. The SDK enables developers to build and configure
303   images and containers in a standard way, making them directly accessible to Narrative apps.
304   The Docker execution model is powerful because contributors can test code on their own
305   machine exactly as it runs on KBase production systems, and containers can be deployed and
306   executed as soon as they are registered without any other manual intervention or system
307   deployments. Docker images also provide a basis for reproducible execution of apps in
308   Narratives.

309   KBase apps built with the SDK include a wide and growing collection of systems biology tools.
310   The SDK has already been applied by KBase developers and external users to implement all of
311   the tools presently available in KBase, including all tools mentioned in the "Alice and Bob" use
312   case example and the science applications described above.

Comparison with Other Platforms

313

314   KBase builds on many ideas from previous systems and infrastructures designed to support
315   large-scale bioinformatics analysis and model building. Many KBase developers have also
316   worked on other systems including the SEED[35], MicrobesOnline[36], RAST[37], PATRIC[38], MG-
317   RAST[39], ModelSEED[15], Gramene[40], iPlant (now CyVerse)[41], RegTransBase[42], RegPrecise[43],
318   and others.

319   KBase differs from existing systems in several ways. While platforms such as Galaxy[44],
320   Taverna[45], CyVerse, XSEDE[46], myExperiment[47], and GenePattern[48] permit a user to develop
321   and run complex bioinformatics workflows, they currently lack tools for workflow annotation or
322   predictive modeling. In contrast, systems such as COBRA Toolbox[49], Pathway Tools[50], and
323   RAVEN Toolbox[51] support predictive modeling of metabolism but lack support for genome
324   assembly, annotation, and comparison. More general computational platforms like Kepler[52],
325   Pegasus[53], Globus[54], and Jupyter[5] extensively support data and workflow management but lack
326   integration with bioinformatics analysis tools. Most of these tools do not allow formal integration
327   of user data with organized reference data or sharing models for data and analyses.

328   A key distinction between KBase and such systems is KBase's extensible core data model,
329   whose aim is to provide consistent access to and integration of reference data sets and user-
330   contributed data across all applications, user interfaces, and services such as annotation from
331   well-curated ontologies. These features mean that KBase users do not have to adjust data files
332   or formats once data are loaded into the system and that both user and reference data can be
333   analyzed using the same tools.

334   KBase serves many of the emerging requirements for open, verifiable science supported wholly
335   or in part by systems such as Galaxy, CyVerse, and Synapse[55]. These capabilities include
336   promoting reproducibility and transparency by persistently linking data and the analyses used to
337   produce them. KBase also improves user accessibility to sophisticated computing by providing
338   an environment populated with tools and data that users can apply on a high-end computing
339   infrastructure without additional setup or maintenance of any software or hardware. Granular

340 sharing with designated individuals or the community at large is supported as well, with
341 increasing capabilities for collaborative analysis and editing of data, experiments, and results.

342 The sharing, collaborative editing, and emerging user-discovery systems in KBase are the first
343 steps toward a social networking and scientific project management system enabling users to
344 find other researchers working in their areas, form teams, organize data and analytical projects,
345 and communicate their results effectively.

346 While some systems such as Synapse and Galaxy do provide provenance information on how
347 particular data files were produced, KBase supports a rich provenance system that integrates
348 such information with both the KBase data model and social network. This integration enables
349 more effective discovery of newly integrated reference data and the relevant work products of
350 other users, paving the way toward a system continually enriched by user contributions and
351 interactions.

Methods

System Architecture

354 KBase has a scalable service-oriented architecture (Fig. 4) in which core components
355 communicate through web APIs. The Narrative user interface is the central hub where users
356 submit computational tasks and manage data. KBase's "Execution Engine" uses a distributed
357 execution model that queues and routes tasks to the appropriate physical compute resources of
358 the internal cloud or HPC facilities. Scientific analysis software is exposed through Narrative
359 apps built and dynamically registered with KBase using the SDK. The software made accessible
360 by apps and any system dependencies is captured and versioned using Docker[34]. The App
361 Catalog maintains the dynamic repository of apps and Docker images and serves this
362 information to the user interface and Execution Engine on demand. Reference and user data
363 are stored identically in a system composed of structured data objects, flat files, and structured
364 references. The references define relationships among types of data objects, enabling a
365 traversable, integrated data model. A set of data services maintains versioning, provenance,
366 permissions, and searching.

367 The physical infrastructure hosting the KBase software platform operates at Argonne National
368 Laboratory and Lawrence Berkeley National Laboratory using an OpenStack managed cloud in
369 combination with large-scale computers at DOE's National Energy Research Scientific
370 Computing Center (NERSC) and Argonne Leadership Computing Facility (ALCF). Multiple
371 deployments of platform services across sites provide some fault tolerance and failover
372 capabilities.
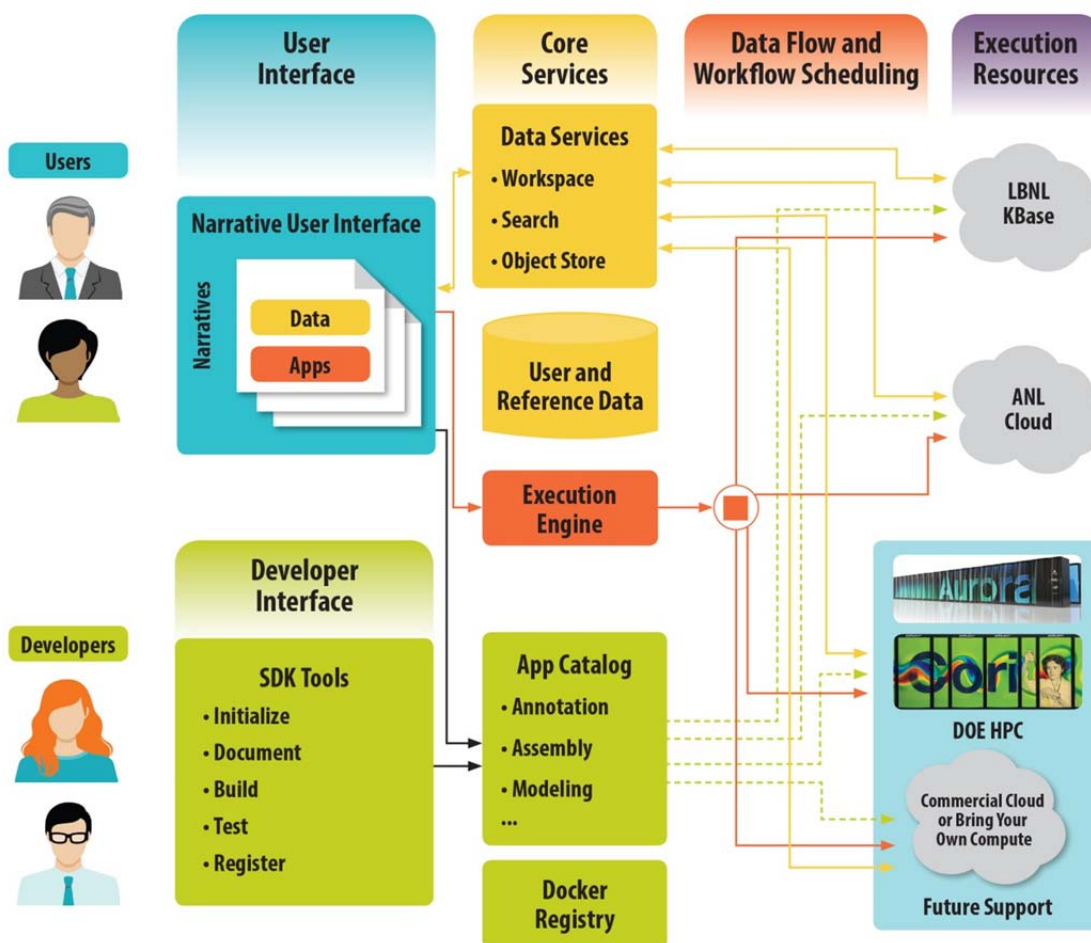
373
374 **Figure 4.** *KBase high-level architecture. KBase is based loosely on a service-oriented architecture that*
375 *bundles related functionality into a set of independently scalable services that are managed to provide*
376 *responsive interaction via the Narrative Interface. [Supercomputer images of Aurora courtesy Argonne National*
377 *Laboratory (under a Creative Commons license) and Cori courtesy the DOE National Energy Research Scientific*
378 *Computing Center (NERSC) run by Lawrence Berkeley National Laboratory.]*

Code Availability

380 KBase software, available at https://github.com/kbase, is open source and freely distributed
381 under the MIT License.

Discussion

383 KBase has made some significant strides in leveraging the opportunities of this new data-rich
384 era of biology. Important developments include (1) a detailed data model with support for
385 provenance and versioning; (2) a rich and growing body of powerful analytical and modeling
386 tools; (3) a collaborative user interface that seamlessly integrates data, analyses, and
387 commentary to capture deep, reproducible scientific analyses; (4) an SDK enabling third-party

388 developers to extend the system with new tools; and (5) a robust, scalable, and extensible
389 underlying computational infrastructure.

390 Since its first public production release in February 2015, the KBase platform has been used
391 extensively by hundreds of users who have built thousands of Narratives. Among these
392 analyses is work described in over 10 peer-reviewed publications covering a wide range of
393 topics. KBase is the only platform where users can immediately apply many of the analysis tools
394 used in these publications in a turnkey, compute-, and data-ready environment using a
395 graphical interface.

396 Extensive user testing is being done to improve the user experience offered by KBase, as well
397 as the range of apps. User feedback indicates that the user interface, with its graphical access
398 to data and apps, enables researchers to quickly learn how to run sophisticated multi-step
399 analyses, find collaborators, and share results. The current KBase release also contains
400 prototype functionality allowing researchers to more fully leverage the power of the Jupyter
401 Notebook by using use code cells to programmatically access KBase and build custom
402 analyses. As KBase improves support for its programming interface and lowers the barrier for
403 third-party incorporation of apps and data types, we anticipate rapid growth in platform
404 functionality and adoption by the community.

405 A key area of improvement in the short term is extending the number and variety of apps and
406 associated data types contributed by KBase personnel and external developers. Development
407 will focus on expanding metagenomics capabilities, improving support for eukaryotic genomic
408 analyses, and providing more-comprehensive comparative genomics tools. Upcoming
409 developments specifically will include improving the input and processing of bulk data sets,
410 easing the process for defining new data types and their relationships to other information in the
411 system, enabling large-scale execution on cloud and HPC resources, and adding frameworks
412 that simplify creation of visualizations.

413 KBase also plans to extend the social platform and user interface to allow (1) more flexible
414 discovery of users; (2) formation of "projects" organizing people, data, and Narratives that can
415 control their group privacy and sharing options more effectively; and (3) a formal process for
416 publishing Narratives.

Acknowledgements

References

427

428 1   Prlić, A. & Procter, J. B. Ten Simple Rules for the Open Development of Scientific
429   Software. *PLOS Computational Biology* **8**, e1002802, doi:10.1371/journal.pcbi.1002802
430   (2012).
431 2   Millman, K. J., and Fernando Pérez. in *Implementing reproducible research*   (ed
432   Friedrich Leisch Victoria Stodden, and Roger D. Peng) 149-183 (CRC Press, 2014).
433 3   Stodden, V. *et al.* Enhancing reproducibility for computational methods. *Science* **354**,
434   1240-1241, doi:10.1126/science.aah6168 (2016).
435 4   DOE Systems Biology Knowledgebase Implementation Plan,
436   http://genomicscience.energy.gov/compbio/kbase_plan/index.shtml (2010).
437 5   Perez, F. & Granger, B. E. IPython: A system for interactive scientific computing.
438   *Comput Sci Eng* **9**, 21-29, doi:10.1109/MCSE.2007.53 (2007).
439 6   O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status,
440   taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745,
441   doi:10.1093/nar/gkv1189 (2016).
442 7   Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics.
443   *Nucleic Acids Res* **40**, D1178-1186, doi:10.1093/nar/gkr944 (2012).
444 8   Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res* **30**, 38-41
445   (2002).
446 9   Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene
447   Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
448 10   Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the
449   project to annotate 1000 genomes. *Nucleic Acids Res* **33**, 5691-5702,
450   doi:10.1093/nar/gki866 (2005).
451 11   Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-230,
452   doi:10.1093/nar/gkt1223 (2014).
453 12   Mulder, N. J. *et al.* New developments in the InterPro database. *Nucleic Acids Res* **35**,
454   D224-228 (2007).
455 13   Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology.
456   *BMC Genomics* **9**, 75 (2008).
457 14   Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
458   trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490
459   (2010).
460 15   Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-
461   scale metabolic models. *Nat Biotechnol* **28**, 977-982, doi:10.1038/nbt.1672 (2010).
462 16   Faria, J. P. *et al.* in *Hydrocarbon and Lipid Microbiology Protocols*   1-27 (Humana
463   Press, 2016).
464 17   Latendresse, M. Efficiently gap-filling reaction networks. *BMC Bioinformatics* **15**, 225,
465   doi:10.1186/1471-2105-15-225 (2014).
466 18   Dreyfuss, J. M. *et al.* Reconstruction and validation of a genome-scale metabolic model
467   for the filamentous fungus Neurospora crassa using FARM. *PLoS Comput Biol* **9**,
468   e1003126, doi:10.1371/journal.pcbi.1003126 (2013).
469 19   Ghosh, S. & Chan, C. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks.
470   *Methods Mol Biol* **1374**, 339-361, doi:10.1007/978-1-4939-3167-5_18 (2016).
471 20   Faria, J. P. *et al.* Computing and Applying Atomic Regulons to Understand Gene
472   Expression and Regulation. *Frontiers in Microbiology* **7**, doi:10.3389/fmicb.2016.01819
473   (2016).

474   21   Deutschbauer, A. *et al.* Evidence-based annotation of gene function in Shewanella
475   oneidensis MR-1 using genome-wide fitness profiling across 121 conditions. *PLoS*
476   *Genet* **7**, e1002385, doi:10.1371/journal.pgen.1002385 (2011).
477   22   Ong, W. K. *et al.* Comparisons of Shewanella strains based on genome annotations,
478   modeling, and experiments. *BMC Syst Biol* **8**, 31, doi:10.1186/1752-0509-8-31 (2014).
479   23   Copeland, A. *et al.* Complete sequence of Shewanella amazonensis SB2B,
480   *EMBL/GenBank/DDBJ databases* https://www.ncbi.nlm.nih.gov/genome/1223 (2006).
481   24   Venkateswaran, K., Dollhopf, M. E., Aller, R., Stackebrandt, E. & Nealson, K. H.
482   Shewanella amazonensis sp. nov., a novel metal-reducing facultative anaerobe from
483   Amazonian shelf muds. *Int J Syst Bacteriol* **48**, 965-972, doi:10.1099/00207713-48-3-
484   965 (1998).
485   25   Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by
486   sequencing randomly bar-coded transposons. *MBio* **6**, e00306-00315,
487   doi:10.1128/mBio.00306-15 (2015).
488   26   Seaver, S. M. *et al.* High-throughput comparison, functional annotation, and metabolic
489   modeling of plant genomes using the PlantSEED resource. *Proceedings of the National*
490   *Academy of Sciences of the United States of America* **111**, 9645-9650,
491   doi:10.1073/pnas.1401329111 (2014).
492   27   Edirisinghe, J. N. *et al.* Modeling central metabolism and energy biosynthesis across
493   microbial life. *BMC Genomics* **17**, 568, doi:10.1186/s12864-016-2887-8 (2016).
494   28   Henry, C. S. *et al.* Microbial Community Metabolic Modeling: A Community Data-Driven
495   Network Reconstruction. *J Cell Physiol* **231**, 2339-2345, doi:10.1002/jcp.25428 (2016).
496   29   Marshall, C. *et al.* Electron transfer and carbon metabolism in an electrosynthetic
497   microbial community, *Preprint at* http://biorxiv.org/content/early/2016/07/07/059410
498   (2016).
499   30   Broberg, C. A., Wu, W., Cavalcoli, J. D., Miller, V. L. & Bachman, M. A. Complete
500   Genome Sequence of Klebsiella pneumoniae Strain ATCC 43816 KPPR1, a Rifampin-
501   Resistant Mutant Commonly Used in Animal, Genetic, and Molecular Biology Studies.
502   *Genome Announc* **2**, doi:10.1128/genomeA.00924-14 (2014).
503   31   Henry, C. S. *et al.* Generation and validation of the iKp1289 metabolic model for
504   Klebsiella pneumoniae KPPR1. *Journal of Infectious Disease* **In press** (2016).
505   32   Liao, Y. C. *et al.* An experimentally validated genome-scale metabolic reconstruction of
506   Klebsiella pneumoniae MGH 78578, iYL1228. *J Bacteriol* **193**, 1710-1717,
507   doi:10.1128/JB.01218-10 (2011).
508   33   Bachman, M. A. *et al.* Genome-Wide Identification of Klebsiella pneumoniae Fitness
509   Genes during Lung Infection. *MBio* **6**, e00775, doi:10.1128/mBio.00775-15 (2015).
510   34   Anderson, C. Docker. *IEEE Software* **32**, 102-105 (2015).
511   35   Overbeek, R., Disz, T. & Stevens, R. The SEED: A peer-to-peer environment for
512   genome annotation. *Commun ACM* **47**, 46-51, doi:Doi 10.1145/1029496.1029525
513   (2004).
514   36   Dehal, P. S. *et al.* MicrobesOnline: an integrated portal for comparative and functional
515   genomics. *Nucleic Acids Res* **38**, D396-400, doi:10.1093/nar/gkp919 (2010).
516   37   Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST
517   algorithm for building custom annotation pipelines and annotating batches of genomes.
518   *Sci Rep* **5**, 8365, doi:10.1038/srep08365 (2015).
519   38   Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis
520   resource. *Nucleic Acids Res* **42**, D581-591, doi:10.1093/nar/gkt1099 (2014).
521   39   Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic
522   phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* **9**, 386,
523   doi:10.1186/1471-2105-9-386 (2008).

524 40   Ware, D. *et al.* Gramene: a resource for comparative grass genomics. *Nucleic Acids Res*
525   **30**, 103-105 (2002).
526 41   Goff, S. A. *et al.* The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Frontiers*
527   *in plant science* **2**, 34, doi:10.3389/fpls.2011.00034 (2011).
528 42   Cipriano, M. J. *et al.* RegTransBase--a database of regulatory sequences and
529   interactions based on literature: a resource for investigating transcriptional regulation in
530   prokaryotes. *BMC Genomics* **14**, 213, doi:10.1186/1471-2164-14-213 (2013).
531 43   Novichkov, P. S. *et al.* RegPrecise 3.0--a resource for genome-scale exploration of
532   transcriptional regulation in bacteria. *BMC Genomics* **14**, 745, doi:10.1186/1471-2164-
533   14-745 (2013).
534 44   Goecks, J., Nekrutenko, A., Taylor, J. & Galaxy, T. Galaxy: a comprehensive approach
535   for supporting accessible, reproducible, and transparent computational research in the
536   life sciences. *Genome Biol* **11**, R86, doi:10.1186/gb-2010-11-8-r86 (2010).
537 45   Oinn, T. *et al.* Taverna: a tool for the composition and enactment of bioinformatics
538   workflows. *Bioinformatics* **20**, 3045-3054, doi:10.1093/bioinformatics/bth361 (2004).
539 46   Towns, J. *et al.* XSEDE: Accelerating Scientific Discovery. *Comput Sci Eng* **16**, 62-74
540   (2014).
541 47   Goble, C. A. *et al.* myExperiment: a repository and social network for the sharing of
542   bioinformatics workflows. *Nucleic Acids Res* **38**, W677-682, doi:10.1093/nar/gkq429
543   (2010).
544 48   Reich, M. *et al.* GenePattern 2.0. *Nat Genet* **38**, 500-501, doi:10.1038/ng0506-500
545   (2006).
546 49   Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-
547   based models: the COBRA Toolbox v2.0. *Nat Protoc* **6**, 1290-1307,
548   doi:10.1038/nprot.2011.308 (2011).
549 50   Karp, P. D. *et al.* The EcoCyc and MetaCyc databases. *Nucleic Acids Res* **28**, 56-59
550   (2000).
551 51   Agren, R. *et al.* The RAVEN toolbox and its use for generating a genome-scale
552   metabolic model for Penicillium chrysogenum. *PLoS Comput Biol* **9**, e1002980,
553   doi:10.1371/journal.pcbi.1002980 (2013).
554 52   Altintas, I. *et al.* Kepler: An extensible system for design and execution of scientific
555   workflows. *16th International Conference on Scientific and Statistical Database*
556   *Management, Proceedings*, 423-424 (2004).
557 53   E, D. Pegasus: A framework for mapping complex scientific workflows onto distributed
558   systems. *Sci Programming-Neth* **13**, 219-237 (2005).
559 54   Ananthakrishnan, R., Chard, K., Foster, I. & Tuecke, S. Globus Platform-as-a-Service for
560   Collaborative Science Applications. *Concurr Comp-Pract E* **27**, 290-305,
561   doi:10.1002/cpe.3262 (2015).
562 55   Derry, J. M. J. *et al.* Developing predictive molecular maps of human disease through
563   community-based modeling. *Nature Genetics* **44**, 127-130, doi:10.1038/ng.1089 (2012).
564
565

566 Author Information

567 Affiliations

568 Department of Bioengineering, University of California, Berkeley, California, USA.
569 Adam P. Arkin
570

571 Environmental Genomics and Systems Biology Division, E. O. Lawrence Berkeley National
572   Laboratory, Berkeley, California, USA.
573 Adam P Arkin, Paramvir Dehal, Fernando Perez, Nomi L Harris, Shane Canon, Michael W
574   Sneddon, Matthew L Henderson, William J Riehl, Dan Gunter, Stephen Chan, Roy T
575   Kamimura, Dylan Chivian, Jason Baumohl, Ben Bowen, John-Marc Chandonia, Inna
576   Dubchak, Marcin P Joachimiak, Pavel Novichkov, Erik Pearson, Sarah S Poon, Gavin A
577   Price, Roman A Sutormin
578

579 Computer Science Department and Computation Institute, University of Chicago, Chicago,
580   Illinois, USA.
581 Rick L Stevens
582

583 Computing, Environment, and Life Sciences Directorate, Argonne National Laboratory,
584   Argonne, Illinois, USA.
585 Rick L Stevens, Thomas S Brettin, James J Davis, Emily Dietrich, Maulik Shukla, Fangfang Xia,
586   Hyunseung Yoo
587

588 Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.
589 Robert W Cottingham, David J Weston, Brian H Davison, Benjamin H Allen, Meghan M Drake,
590   Holly L Haun, Miriam L Land, Marissa Mills, Priya Ranjan, Mustafa H Syed
591

592 Biology Department, Brookhaven National Laboratory, Upton, New York, USA.
593 Sergei Maslov, Fei He, Shinjae Yoo
594

595 Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois,
596   USA.
597 Christopher S Henry, Dan Murphy-Olson, Folker Meyer, Elizabeth M Glass, Christopher C Bun,
598   Ric Colasanti, Neal Conrad, Scott Devoid, Janaka N Edirisinghe, José P Faria, Paul M
599   Frybarger, Wolfgang Gerlach, Kevin P Keegan, Gary J Olsen, Bob Olson, Bruce
600   Parrello, Samuel M D Seaver
601

602 Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA.
603 Doreen Ware, Sunita Kumari, Jer-Ming Chia, James Gurtowski, Vivek Kumar, Shiran Pasternak,
604   Srividya Ramakrishnan, Michael C Schatz, James Thomason
605

606 Berkeley Institute for Data Science, University of California, Berkeley, California, USA.
607 Fernando Perez
608
609 Department of Biology, Hope College, Holland, Michigan, USA.
610 Aaron A Best
611
612 Department of Plant and Microbial Biology, University of California, Berkeley, California, USA.
613 Steven E Brenner
614
615 Department of Computer Science, Hope College, Holland, Michigan, USA.
616 Matthew DeJongh, Shinnosuke Kondo
617
618 Computation Institute, University of Chicago, Chicago, Illinois, USA.
619 Janaka N Edirisinghe
620
621 Program in Computational Biology and Bioinformatics, Yale University, New Haven,
622   Connecticut, USA.
623 Gang Fang, Mark Gerstein, Daifeng Wang
624
625 Department of Plant Pathology and Genome Center, University of California-Davis, Davis
626   California, USA.
627 Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA.
628 Rashmi Jain, Taeyun Oh, Pamela C Ronald
629
630 Department of Microbiology, University of Illinois at Urbana-Chamgain, Urbana, Illinois, USA.
631 Gary J Olsen
632
633 Department of Plant Sciences, University of Tennessee, Knoxville, Tennessee, USA.
634 Priya Ranjan
635
636 Department of Mathematics, Hope College, Holland, Michigan, USA.
637 Nathan L Tintle
638

639 **Present Addresses**

640 Department of Bioengineering and Carl R. Woese Institute for Genomic Biology, University of
641   Illinois at Urbana-Champaign, Urbana, Illinois, USA.
642 Sergei Maslov
643
644 New York University Shanghai Campus, Pudong, Shanghai, China.
645 Gang Fang
646

647   Department of Plant Pathology, Kansas State University, Manhattan, Kansas, USA.

648   Fei He

649

650   Center for Data Intensive Science, University of Chicago, Chicago, Illinois, USA.

651   Kevin P Keegan

652

653   Insilicogen. Inc., Giheung-gu, Yongin-si, Gyeonggi-do, Korea.

654   Taeyun Oh

655

656   Memorial Sloan Kettering Cancer Center, New York, New York, USA.

657   Mustafa H Syed

658

659   Dordt College, Sioux Center, Iowa, USA.

660   Nathan L Tintle

661

662   Department of Biomedical Informatics, Stony Brook University, Stony Brook, New York, USA.

663   Daifeng Wang

664

665 **Contributions**

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

667

668   APA, RLS, SC, MWS, MLH, WJR, DG, DMO, SYC, TSB, FM, DC, JB, AAB, BB, SEB, CCB, JMC,
669   JC, RC, NC, JJD, MDJ, SD, FH, MPJ, KPK, BO, SP, EP, GAP, SR, PR, SMDS, MS, RAS, MHS, JT, FX,
670   HSY and SY designed and developed the system.

671

672   RWC, NLH, RK, DC, DJW, EMG, BHD, SK, BHA, ED, MMD, ID, JNE, GF, JPF, PMF, WG, MG, JG, RJ,
673   SNK, VK, MLL, MDM, PN, OTY, GJO, BDP, SSP, PCR, MCS, NLT and DFW developed, documented
674   and conducted testing and validation.

675

676   APA and CSH drafted the manuscript.

677

678   NLH, HLH, BHA, MDM, MPJ, AAB, JMC, DC, BO, BHD, NLT, SM, PCR and MDJ revised the
679   manuscript and provided important intellectual content.

680

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

682 **Competing financial interests**

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

687 All other authors declare no competing financial interests.

688 **Corresponding author**

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