KBase Performance Metric for FY17: Develop improved open access platforms for computational analysis of large genomic datasets.

Q2 Metric: Report on the new capabilities to perform reproducible genomics analyses on large datasets and share the results with other researchers.

1. BACKGROUND

The DOE Systems Biology Knowledgebase (KBase, http://kbase.us/) [1] is an open access software and data platform designed to make it easier for scientists to create, execute, collaborate on, and share sophisticated, reproducible analyses of their biological data in the context of public data and data other users have privately shared with them. KBase supports a growing and extensible set of applications (apps) for contig assembly, genome annotation, metabolic model reconstruction, flux balance analysis, expression analysis, and comparative genomics. New capabilities in KBase for bulk upload and execution make it possible to upload large datasets and quickly run them through sophisticated analysis workflows. These workflows can be saved and shared as KBase *Narratives*—dynamic, interactive documents that include all the data, analysis steps, parameters, visualizations, scripts, commentary, results, and conclusions of an experiment.

A growing number of KBase users have applied the system to address a range of scientific problems, including comparative genomics of plants, prediction of microbiome interactions, and deep metabolic modeling of environmental and engineered microbes. The workflows they have chosen to share publicly (see http://kbase.us/narrative-library for some examples) can be viewed, copied, and re-run, possibly with different parameters or new datasets. By enabling reproducible scientific analysis on large datasets and facilitating collaboration, KBase is helping to accelerate the pace of systems biology research.

2. KBASE NARRATIVE USER INTERFACE

The support for collaboration and reproducibility in KBase is centered around the Narrative Interface, which lies at the core of the KBase user experience. Built on the Jupyter Notebook [2], the Narrative Interface (Figure 1) supports both point-and-click and scripting access to system functionality in a “notebook” environment, enabling computational sophisticates and experimentalists to easily collaborate within the same platform and share their datasets and results. The Narrative Interface makes it easy for users to browse, select, configure and run analysis functionality in the form of “apps”. Users also can add their own code cells to incorporate custom analysis steps that are not available as KBase apps, or to run analyses in bulk over large datasets.

Fig. 1. *KBase Narrative. A Narrative is a shareable, reproducible computational experiment that can include data, analysis steps, results, visualizations and commentary.*

3. SUPPORT FOR SHARING AND COLLABORATION

Collaboration in KBase is supported by the ease with which all data in the system may be shared and copied among users. A user can share any Narrative that they own with other KBase users (or with the public). Importantly, when a user shares a Narrative, they also are sharing all the data objects loaded, used, or generated within the Narrative, complete with versioning and provenance. Users with read privileges for a Narrative can create their own copy, which they own and can edit. This enables users to quickly replicate and expand on any KBase Narrative shared with them. This approach to sharing facilitates reproducible interdisciplinary science by allowing researchers with different expertise to quickly and easily exchange data, results, methodologies, and workflows to address complex biological problems.

We demonstrate how KBase facilitates sharing, collaboration, and interdisciplinary research with a series of example Narratives (Fig. 2) featuring two hypothetical scientists: Alice, a wet-lab biologist with expertise in assembly, annotation, and comparative genomics, and Bob, a computational biologist with expertise in metabolic modeling. (If you would like to view and perhaps copy these Narratives, sign up for a KBase User Account at www.kbase.us.). In the first Narrative (Alice Narrative 1: Assembly and Annotation), Alice uploads raw reads from a new strain of *Shewanella* that she is analyzing. She uses KBase to assemble and annotate these reads, generating a new genome object. In a second Narrative (Alice Narrative 2: Comparative Genomics), Alice compares her new genome with other close strains of *Shewanella*. She finds growth phenotype data for *Shewanella oneidensis* MR-1, which is phylogenetically close to her strain. This inspires Alice to run a growth phenotype array on her own

strain, which she also uploads to her Narrative. Alice then compares both phenotype arrays and notices many differences that she cannot explain.

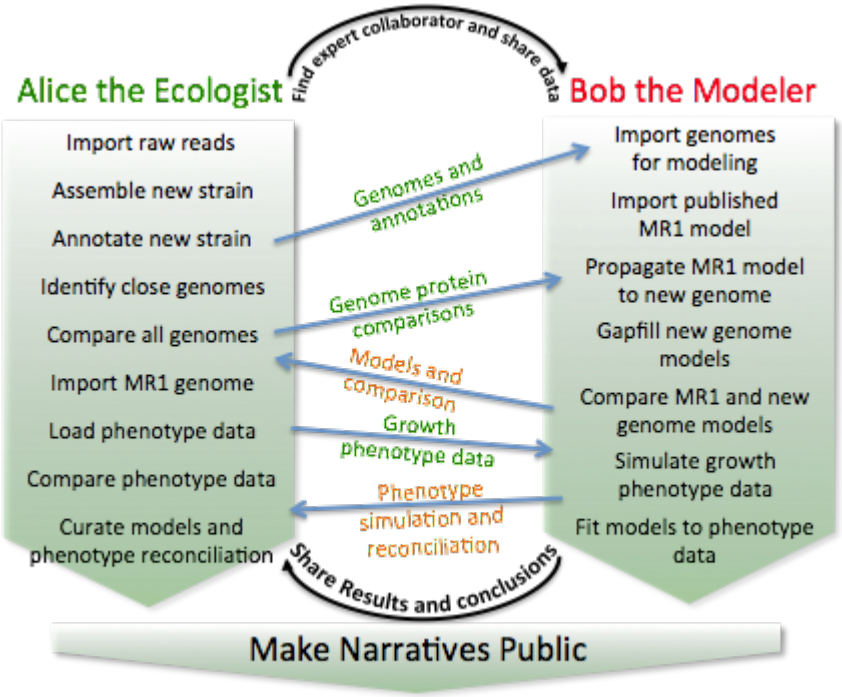

Fig. 2. *Example of collaboration in KBase. Two researchers collaborate using Narratives to reach more complete scientific conclusions than either could have achieved alone.*

At this point, Alice contacts Bob, who suggests that metabolic models can help her analyze the Biolog phenotype data. Alice shares her Narratives with Bob, who copies her genomes into a new third Narrative. In this third Narrative (Bob Narrative: Build Metabolic Models), Bob loads a published model of *S. oneidensis* MR-1, which he then propagates to Alice's genome. Bob compares the models, identifying some interesting metabolic differences. Then Bob creates a fourth Narrative (Bob and Alice Narrative 1: Phenotype Data Analysis), where he imports Alice's Biolog data and simulates the data with his *Shewanella* models. He optimizes his models to fit the Biolog data and shares the results with Alice.

Finally, they build a fifth Narrative (Bob and Alice Narrative 2: Phenotype Data Reconciliation) together in which Alice refines Bob's models by replacing gapfilled reactions with more biologically relevant selections, gaining a complete understanding of the differences between her strain and MR-1. All data used in this example are real: Alice's raw reads are from an existing genome, *Shewanella amazonensis* SB2B, and the growth phenotype data are from an existing experimental study. The computational experiment carried out in the five Narratives results in the development and validation of a new genome-scale metabolic model of *S. amazonensis* SB2B in KBase, as well as the improvement of the existing model for *S. oneidensis* MR-1.

This example demonstrates how KBase's user interface facilitates a seamless collaboration between scientists with different but complementary expertise who are able to accomplish more together than they could individually. By saving their work as Narratives, scientists who use KBase also enable other researchers to quickly reproduce and extend their work.

4. SUPPORT FOR REPRODUCIBLE RESEARCH

Although there is increasing support in the scientific community for the *idea* of reproducibility, most scientific work is not easy to reproduce. Research results are normally published in journals; these papers may include links to the raw input data and/or output data. However, reproducing the authors' results would be far from a simple process. To do this, you would need to obtain the software used by the researchers, in the right version--this is not always possible, particularly if the software was not open source. It would then be necessary to install, configure and run the software on the original datasets, with the same parameters used by the authors (which may or may not have been explicitly listed in the paper). Even after all that work, if your environment was different in some way from the original researchers', you would be unlikely to get the same results. If you did, you would still need to utilize additional tools to visualize them in an accessible way. Finally, much of the work you did in installing and configuring the software would itself not be reproducible--you couldn't just hand it off to another researcher at a different lab.

KBase Narratives, in contrast, are inherently reproducible "publications" that allow you to not only see what a researcher did, but to quickly and easily replicate their analyses and obtain the same results. Any Narrative that has been shared with you, even with view-only access, can be copied to your KBase account with the "Copy" button (Fig. 3). This automatically does a "deep" copy, which includes not just the results but also copies all the original data objects. You can then rerun the steps in the Narrative to reproduce the computational experiment, or try changing the parameters or input data to alter or perhaps even improve the results.

Fig. 3. The "Copy" button in the KBase interface generates a complete copy of a Narrative that includes all the analysis steps as well as the input and output data. A copied Narrative can be re-run to reproduce the entire computational experiment.

The KBase data and application model are also designed from the ground up to comprehensively support collaboration and reproducibility. KBase is built upon an object-oriented data model, which holds all user and reference data. All data objects in KBase, including Narratives themselves, have their own provenance and version information. Provenance data capture when, how, and by whom each object was created, including the apps or upload tools used and their associated parameters and files. All data is also versioned, so that even if a user overwrites an object, the original data is recoverable.

KBase apps are also versioned. All analysis software exposed within KBase was built and dynamically registered using the KBase Software Development Kit (SDK). The SDK ensures that all software, including any related system dependencies, is captured and versioned using Docker, which allows the individual programs and system dependencies of each Narrative app to be saved as an image and run securely and identically across physical resources within individual, isolated environments called containers. Docker images provide a basis for reproducible execution of apps in Narratives.

5. SUPPORT FOR LARGE-SCALE ANALYSIS

5.1 Support for bulk upload of data into KBase

KBase was designed with a scalable computational infrastructure to support large-scale analysis of biological data. Large-scale analysis involves large-scale data, in two senses: large *numbers* of data objects (for example, analyzing thousands of microbial genomes to compare them or assign them to phylogenetic trees, or computing over vast compendia of metabolic data) and also large *individual* data objects (such as reads generated by sequencing projects). Recent work on the KBase platform has enabled it to support analysis on a larger scale by providing better ways to upload large individual data files or large collections of data, as well as enabling bulk processing.

Uploading large numbers of files, such as sequencing read sets, one at time through the Data Panel in the Narrative Interface can be tedious and time consuming. To address this issue, KBase has expanded its data import functionality to support bulk upload of data, meaning that users can import multiple files simultaneously. Currently, KBase supports bulk upload of sequencing reads (Paired End, Single End, SRA) and genomes (GenBank files) through its Narrative "staging" interface (Fig. 4). Support for additional data types is in progress.

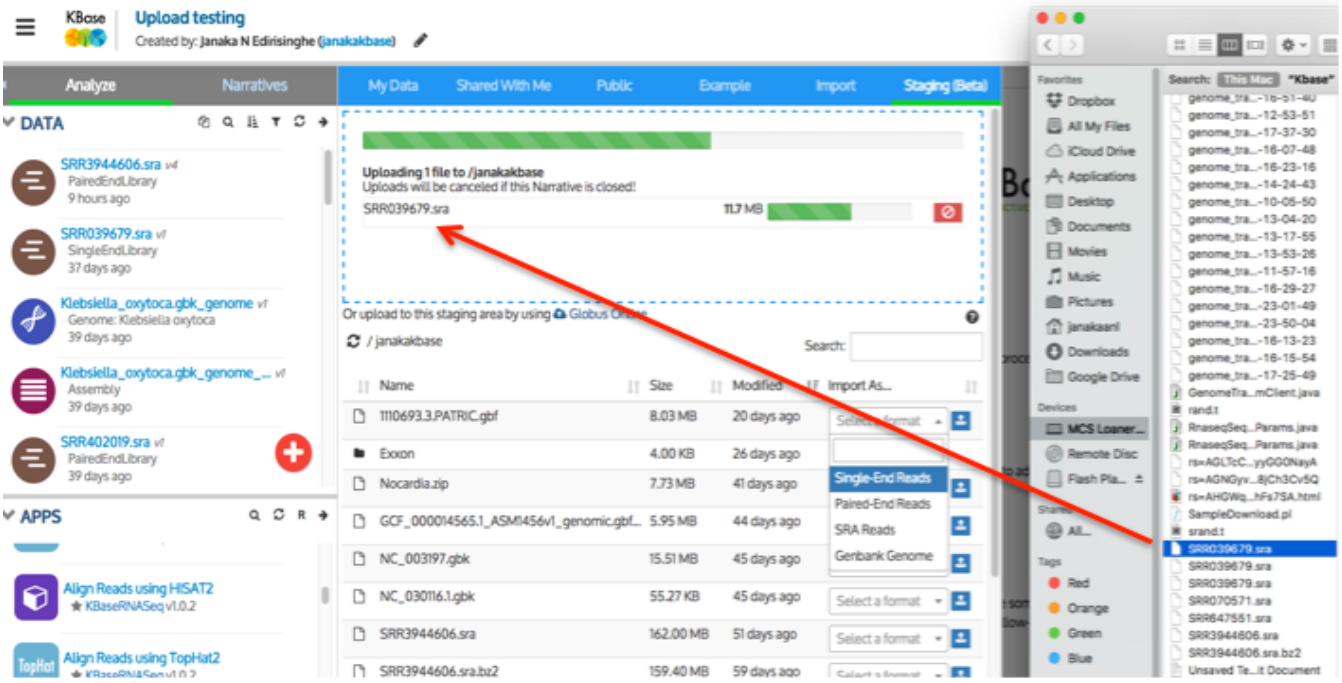

Fig. 4. The new "staging" uploader allows files to be dragged from the user's computer and dropped into their staging area, from which they can be imported into a Narrative in one of the supported formats (currently reads or genomes).

The "staging" upload involves a two-phase process (see Fig. 4):

1. Users drag and drop files from their local file space to a staging area in their KBase account
2. Data files in the staging area can be imported to a Narrative as a typed object for analysis. This is done by accessing the pulldown menu of types next to a file and choosing the appropriate type (for example, "SRA Reads") and then clicking the "import" icon to its right.

In addition, data stored in public FTP repositories can be directly uploaded into a Narrative using the new KBase web upload apps. This makes it more convenient to upload large data files to KBase, as it does not require the user to first download the data from an external source and save it on their computer.

5.2 Support for programmatic and bulk execution of KBase apps

The Narrative Interface in KBase is built on the Jupyter Notebook framework, which brings substantial advantages to the KBase platform. One of the most significant advantages is the availability of “code cells” in the Jupyter framework, which enable any user to write, execute, and share custom code from within KBase’s graphical user interface. This means that within a single Narrative, it is possible to intersperse chunks of custom code with drag-and-drop apps. Computational biologists who write code can thus use the same user interface as non-coding bench biologists, and both can work in the way they find most effective and even work together on collaborative Narratives.

While code cells have been a part of the KBase platform for a long time (with limited support), a recent advance in KBase has greatly amplified their power and utility. KBase now offers a simple programmatic interface to its app execution engine from within the code cells in the Narrative UI, enabling a user to write a code cell that programmatically executes any KBase app. Because code cells support programming paradigms such as loops, they can be easily applied to run apps in bulk. KBase now provides an intuitive support mechanism to help users who wish to programmatically run apps without having to overcome the kind of steep learning curve that the use of a programmatic API often imposes. Specifically, the dropdown menu in the upper right-hand corner of every KBase app now includes a “show code” menu item, which exposes a simple snippet of code showing how the KBase App API can be used to run the app programmatically with the specified parameters (see Figure 5). The user can easily copy this code snippet into a code cell, tweak the input parameters, and run the app without needing to learn the App API. It is then a simple task to add a loop to run the app an arbitrary number of times, greatly simplifying the bulk execution process.

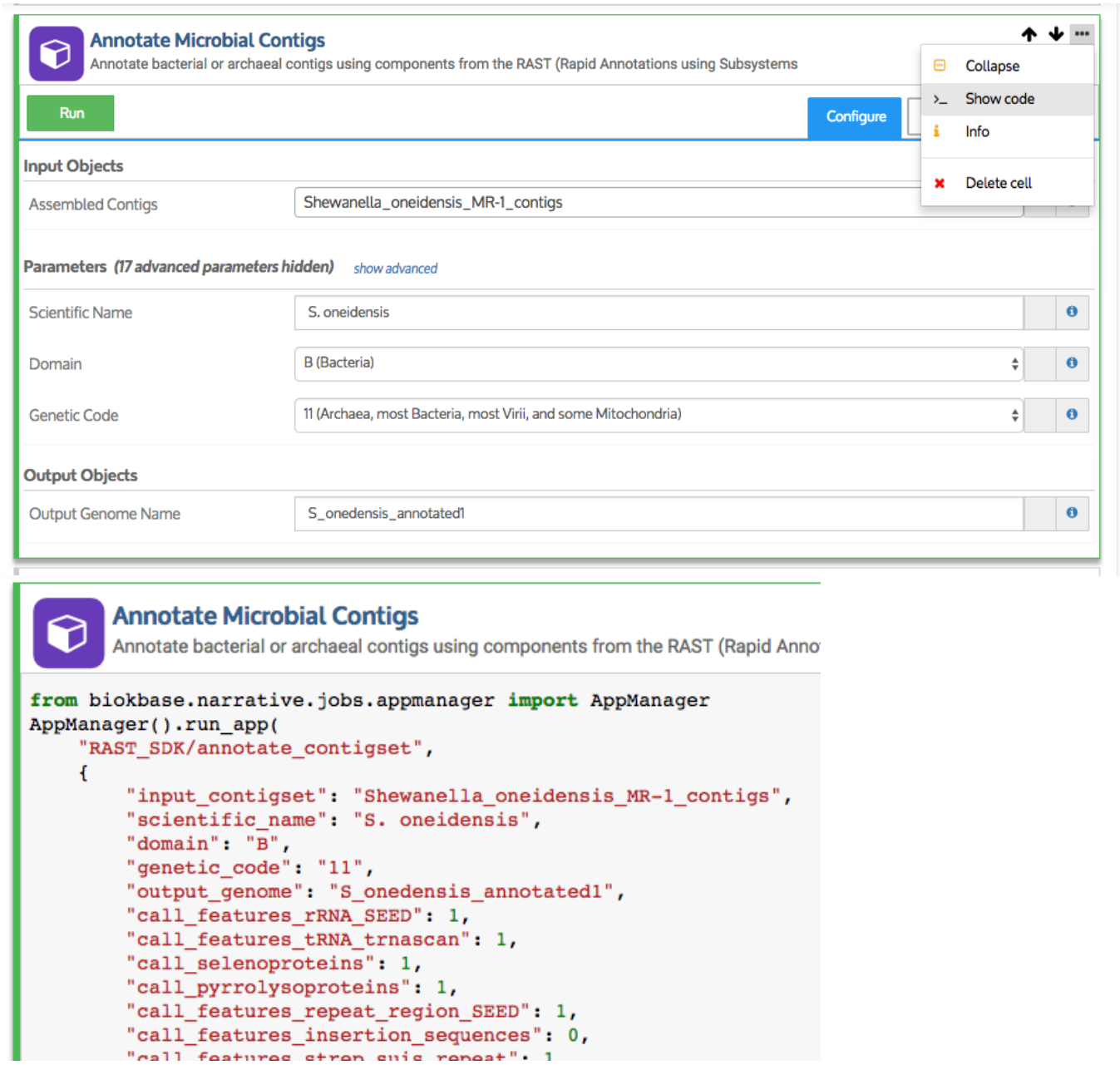

Figure 5: Apps in KBase can be added and run via the graphical user interface; the "Show code" menu item exposes the call to the KBase API that executes the app on the specified parameters. This code can then be copied into a code cell to run the app programmatically, which makes it simple to loop over multiple input objects to execute the analysis in bulk.

5.3 Support for bulk download of output from KBase analysis

KBase's capabilities for bulk upload of data and bulk execution of apps allow users to easily run large-scale analyses. After uploading a large amount of data and running large-scale analyses, users may want to download the results in bulk as well. The new "Bulk Download Modeling Objects" App allows users to download (as a zip file) selected modeling objects or all modeling objects from a Narrative (Fig. 6). This new functionality completes the cycle of large-scale analysis in KBase, from bulk upload

through app execution to bulk download of the analysis results. Although bulk download currently only supports modeling objects, this capability will soon be extended to all object types.

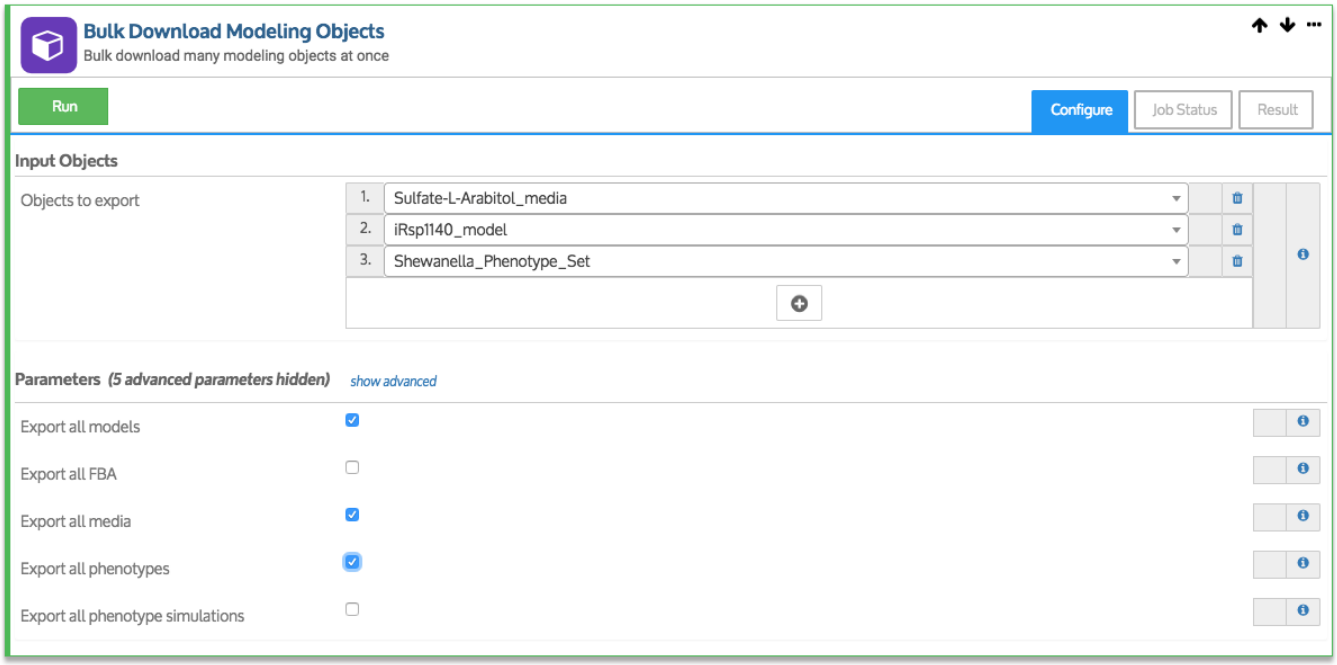

Fig 6. Users can download in bulk any or all of the modeling objects generated by their KBase analyses, including models, FBA simulation outputs, media formulations, phenotype datasets and phenotype simulation outputs.

6. EXAMPLES OF LARGE-SCALE COLLABORATIVE SCIENCE ACCOMPLISHED IN KBASE

KBase's support for bulk upload, analysis, and collaboration is enabling a growing number of scientific research teams to perform large-scale collaborative projects that have yielded published results. A few of these are described in this section.

6.1 Large-scale reconstruction of 773 microbes in a natural microbiome

Natural microbiomes are enormously complex, typically being comprised of thousands of distinct species that form dynamic webs of interactions. Such natural microbiomes are central to many processes in nature that are of great interest to the Department of Energy, including: (i) biodegradation of contaminants in natural systems; (ii) carbon cycling in soil and aquatic systems; and (iii) general microbiome response to perturbations in ecosystems caused by climate change. For these reasons, natural microbiomes have been the focus of extensive study, and yet, these systems form one of the ultimate large-scale challenge in science. One of the ultimate goals in microbiome research is to sequence, assemble, annotate and model in detail every species comprising a natural microbiome, enabling a deeper mechanistic understanding of how and why these species interact, and how they will as a community respond to perturbations. KBase is ideally suited as an environment in which to perform this type of study on a microbiome system, as it already encompasses the needed tools in assembly, annotation, and modeling, and the new capabilities in KBase that support large-scale

analysis means that these tools can be applied at the scale needed to properly address complex natural microbiomes.

KBase's ability to greatly facilitate large-scale analysis of microbiomes was demonstrated in the recent groundbreaking work to construct 773 genome-scale metabolic models for isolates from a complex gut microbiome, as recently published in Nature Biotechnology [3]. The authors developed a new resource called AGORA (Assembly of Gut Organisms through Reconstruction and Analysis), which provides a large collection of refined metabolic models to broadly catalyze research efforts to understand the human gut microbiome. In this work, the authors used KBase to construct initial models for the majority of the microbes in their studies. KBase was important to this work because it permitted the reconstruction of models at a scale that is very difficult, tedious, and time-consuming to achieve with other tools. This is particularly true now, with the new bulk analysis capabilities available in the KBase platform.

All the AGORA models are available for download from the AGORA website. To make the AGORA models easily accessible via KBase, we created a public Narrative that includes all 773 models.

6.2 Reconstruction of core metabolism across the microbial tree of life

Of course, natural microbiome analysis is far from being the only area of biology where large-scale analysis is essential. The analysis of individual genome isolates has also grown into a large-scale challenge due to the rapid proliferation of new reference genomes. GenBank has swelled to include over 90,000 reference microbial genomes, and it continues to grow by thousands of new genomes every month. Studies that encompass all reference genomes are tremendously valuable for the insights they can reveal in understanding the core principles that drive genome evolution, but performing such studies requires computational tools that can scale with the rapid growth in available data. KBase is making this type of large-scale analysis of reference genomes more tractable.

In recent work by Edirisinghe et al [4], KBase was applied to build core metabolic models (CMMs) representing central metabolism of bacteria for over 8000 genomes that span the prokaryotic tree of life. The authors used CMMs to determine: (i) accurate ATP yields based on different growth/environmental conditions (ii) ETC variations and respiration types (iii) ability to produce fermentation products (iv) presence and absence of classical biochemical pathways in central metabolism and (v) ability to produce key metabolic pathway intermediates in central metabolism which are precursors of essential biomass components of the cell.

This large-scale analysis was possible because KBase has the capacity to import and host large amounts of genomic data and quickly generate 8000 core metabolic models based on the imported genomes. To derive scientific conclusions in a large-scale analysis such as this, it is essential to have access to all model data in a single iteration, which KBase enabled.

The complete CMM analysis workflow, using *E. coli* as an example to demonstrate bulk metabolic model construction using code cells, is captured in a public Narrative which can be viewed and copied by any KBase user. The models can be downloaded in bulk as a compressed file using the "Bulk Download Modeling Objects" app.

6.3 Collaboration between Experimental Biologists and Computational Modelers

The rapid growth in the scale of biological datasets is not the sole challenge facing this field. Biology datasets are also rapidly growing in complexity and heterogeneity, as scientists apply an increasingly

diverse combination of technologies (including sequencing, metabolomics, transcriptomics, microfluidics) to improve our understanding of the systems we study. These diverse technologies and datasets increasingly demand more interdisciplinary approaches to our science, meaning collaboration and data-sharing are vital. Our computational tools must facilitate this collaboration and data-sharing, and once again, this is an area where KBase stands out as a platform. Two recent publications highlight how KBase facilitates collaboration and data sharing among scientists.

First, in recent work by Henry et al., scientists from Argonne National Laboratory (ANL) and Pacific Northwest National Laboratory (PNNL) collaborated extensively to study interactions between an autotroph and heterotroph in a simple microbial community. In this collaboration, scientists from PNNL uploaded extensive raw data into the KBase platform, including annotated genomes, media formulations and transcriptomic data. This data was shared with scientists at ANL, who constructed models for the genomes, assembled a community metabolic model, simulated flux with the model, and compared flux with the expression data loaded by PNNL. The PNNL and ANL scientists then reviewed the KBase Narratives with all analyses together, refining and enhancing the analyses, and then writing a paper based on the work [5]. The final Narrative from this work is available at http://www.kbase.us/predict-interspecies-interactions/

In other recent work, scientists from Northwestern University and ANL collaborated to construct and validate a new genome-scale metabolic model of *Klebsiella pneumonia KPPR1* [6]. In this collaboration, experimental biologists at Northwestern worked with the ANL team to load the *Klebsiella KPPR1* genome and Biolog phenotype array data into KBase. The ANL team constructed a model of this data, refined the model, and simulated the Biolog experiment, reconciling the KPPR1 model to improve agreement with the experimental Biolog data. The Northwestern team reviewed the Narrative produced by the ANL team, identifying specific conditions of interest in the Biolog analysis and performing additional targeted growth culture experiments. The outcome of these new experiments was used to refine the Biolog data and the model in KBase. Ultimately, in this example, the KBase Narrative served as an ideal platform to share and iterate on computational analyses with experimental collaborators.

7. CONCLUSION

From its conception, KBase was designed to be a scalable computational platform for performing large-scale systems biology analyses and sharing the workflows reproducibly. Recent additions to KBase have made it possible to easily upload large datasets such as sets of sequencing reads and analyze them in bulk. Bulk analysis in KBase leverages the powerful and flexible code cells that enable programmatic calls to the analysis tools to be interleaved with user-friendly apps added via the graphical user interface. KBase's support of both programmatic and graphical access to functionality, rare among bioinformatics systems, allows both programming and non-programming scientists to work collaboratively and combine their expertise to accelerate their research and generate reproducible Narratives that can be built upon by future scientists. KBase is increasingly being used by both experimental and computational biologists to perform and share large-scale research studies, some of which were highlighted in this report.

Q2 Report Authors: Chris Henry, Nomi Harris, José Pedro Lopes Faria, Janaka Edirisinghe
KBase PIs: Adam Arkin, Robert Cottingham, Chris Henry

REFERENCES

1. Arkin, A.P. et al. The DOE Systems Biology Knowledgebase (KBase).
bioRxiv preprint first posted online Dec. 22, 2016; doi: http://dx.doi.org/10.1101/096354.

2. Pérez F, Granger BE. IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering, vol. 9, no. 3, pp. 21-29, May/June 2007, doi:10.1109/MCSE.2007.53. URL: http://ipython.org

3. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature Biotechnology. 2016; doi:10.1038/nbt.3703

4. Edirisinghe JN, Weisenhorn P, Conrad N, Xia F, Overbeek R, Stevens RL, Henry CS. Modeling central metabolism and energy biosynthesis across microbial life. BMC Genomics. 2016;17. doi:10.1186/s12864-016-2887-8

5. Henry CS, Bernstein HC, Weisenhorn P, Taylor RC, Lee J-Y, Zucker J, Sung H-S. Microbial Community Metabolic Modeling: A Community Data-Driven Network Reconstruction. Journal of Cellular Physiology. 2016;231: 2339–2345. doi:10.1002/jcp.25428

6. Henry CS, Rotman E, Lathem WW, Tyo KEJ, Hauser AR,d, Mandel MJ. "Generation and validation of the iKp1289 metabolic model for Klebsiella pneumoniae KPPR1." Journal of Infectious Diseases. 2017;215: S37-S43. doi: https://doi.org/10.1093/infdis/jiw465