

Connecting JGI and KBase

Shane Canon*¹ (scanon@lbl.gov), Kjersten Fagnan¹, Harika Tandra¹, Chris Beecroft¹, **Adam P. Arkin¹, Chris Henry², Bob Cottingham³** and the KBase Team at the following institutions

¹Lawrence Berkeley National Laboratory, Berkeley, CA; ²Argonne National Laboratory, Argonne, IL; ³Oak Ridge National Laboratory, Oak Ridge, TN; ⁴Brookhaven National Laboratory, Upton, NY; ⁵ Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

<http://kbase.us>

Project Goals: The DOE Systems Biology Knowledgebase (KBase) is a free, open-source software and data platform that enables researchers to collaboratively generate, test, compare, and share hypotheses about biological functions; analyze their own data along with public and collaborator data; and combine experimental evidence and conclusions to model plant and microbial physiology and community dynamics. KBase's analytical capabilities currently include (meta)genome assembly, annotation, comparative genomics, transcriptomics, and metabolic modeling. Its web-based user interface supports building, sharing, and publishing reproducible, annotated analysis workflows with integrated data. Additionally, KBase has a software development kit that enables the community to add functionality to the system.

To maximize the utility of KBase, users need to be able to easily import the data they are interested in exploring, analyzing, and sharing. For many researchers in the DOE Biological Science community, this data is initially generated at the Joint Genome Institute (JGI), and the ability to easily find and transfer data from JGI to KBase lowers the barrier to data analysis in KBase. JGI and KBase have collaborated closely over the past year and a half to connect KBase and JGI in a more seamless manner for our users. In this poster we will describe this effort and how we have integrated these two systems. This enhanced integration enables a user to seamlessly search JGI data from within KBase and easily import data sets of interest into private workspaces where they can conduct further analysis in KBase. This greatly improves on previous methods available to users that required them to jump between multiple web systems and had limited support for different data types.

In order to streamline and improve the data import from JGI to KBase, the two teams have heavily redesigned how the systems interconnect and created new interfaces to facilitate search and data transfers. These enhancements heavily leverage the JGI Archive and Metadata Organizer (JAMO) system which is used to store JGI data and maintain critical metadata about the data. JAMO is targeted as being the primary data tracking system for JGI and already tracks over 6 PB of data and 5.5 million files. In cooperation with KBase, JGI developed a new RESTful interface that the KBase platform uses to search data in JAMO and request transfers. This service layers on top of an Elasticsearch database that creates indexes of the data the enable rapid searching. KBase has developed an enhanced user interface (Figure 1) that enables users to specify query terms and filter on key fields like data type, Project ID, or PI. Once a user has found data of interest, they can easily transfer it to their KBase staging area (Figure 2) for import using KBase's existing suite of data importers.

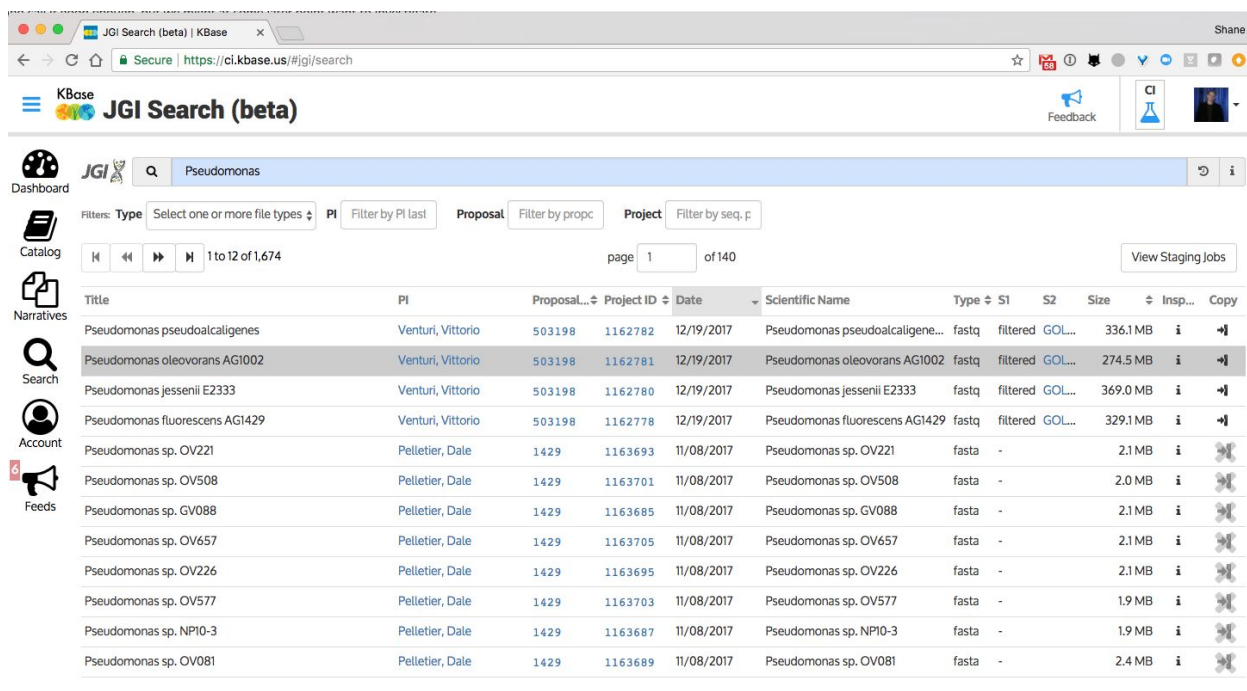


Figure 1. Improved interface for searching JGI data from inside KBase.

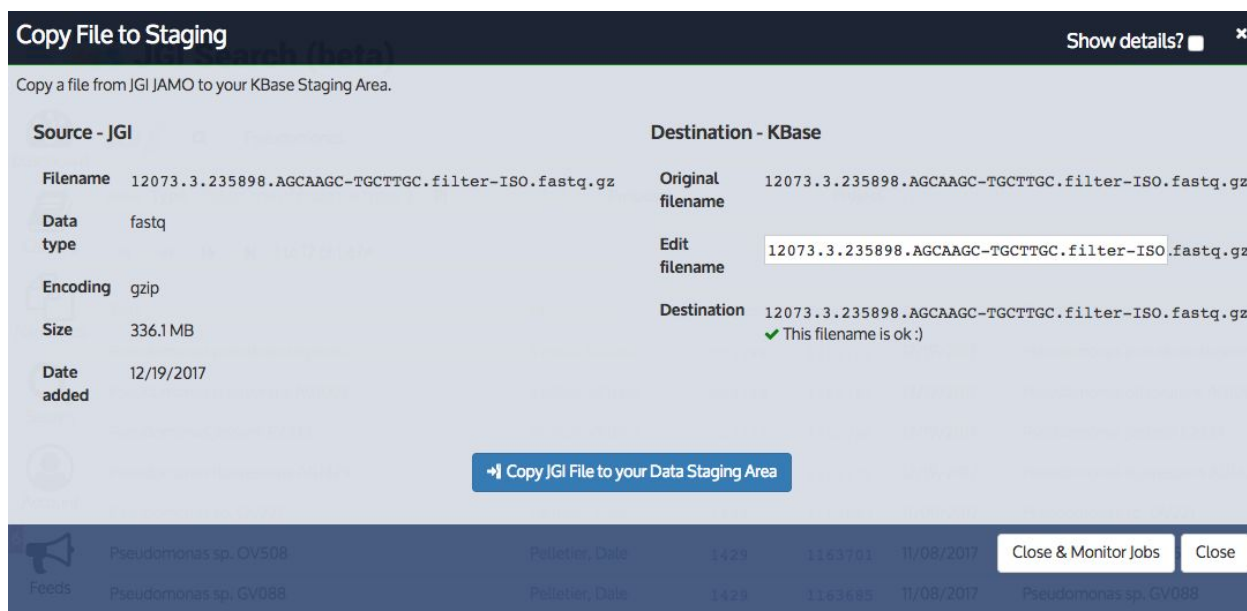


Figure 2. Staging interface that transfers JGI data to KBase for importing.

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

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.