JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Data Management Practices and Resources

Current Data Management Practices

Raw sequence data—Raw data produced by the JGI's sequencing platforms are base-called in real time and primary (image) data are discarded shortly after base-calling is complete. Sequence data are stored as FASTQ files containing the linear sequence and associated quality scores. All primary sequence data receive automated quality control screening to assess sequence quality and quantity, identify major contamination, and confirm that the data generated correspond to the expected organism/environment being sequenced. For most sequencing projects, raw sequence data is made available to registered users through the genome-specific portal web page and are submitted to NCBI's Sequence Read Archive (SRA) as soon as the embargo defined by the JGI Data Policy has ended.

Processed sequence data—For most user projects, the JGI conducts standard post-sequencing analyses that vary by project type (see Product Offerings). These analyses and associated metadata are available for download from project-specific data portals to the JGI data users after the embargo has ended. User-selected subsets of public data can also be downloaded from the JGI's comparative analysis platforms as described in Data Resources below. Beginning in fall 2013 users wishing to download data from the JGI portals will be required to register with the JGI.

Project metadata—Every user project has associated metadata that are important for interpretation of the data produced. These include, for example, authors of the data, PI contact information, detailed sample information and details of analyses performed. These data will be served with processed data from the JGI's data portals whenever possible, however metadata for older legacy projects may be limited.

Data archive and back-up—All data available for download by users is automatically archived onto a high-speed tape system located at the National Energy Research Scientific Computing facility (NERSC). A second redundant copy of these data is also stored on high-speed tape (but not on the same cassette!). Typical retrieval times range from minutes to hours depending primarily on file size.

Data Management Resources

The JGI provides some resources to users to facilitate analysis of genomic data produced outside of the JGI or to allow post-hoc analysis of JGI data. Because these resources are supported through public funds it is expected that such analyses and externally created data managed with these resources will be made public at the time of publication. The user bears sole responsibility for providing and maintaining public access to these data and analyses.

Phytozome —The JGI's comparative genomics platform for plants allows user-driven filtering of plant genome data by characteristics of interest that are then downloadable by individual users. Phytozome does not currently keep these analyses nor support their display for other users. Phytozome supports the display of user-provided genome tracks that are compatible with GBrowse, but does not generally display these tracks for other users. By mutual agreement between JGI and the data providers, Phytozome may display non-JGI genomes for comparative purposes or additional genome-anchored data for JGI genomes. The JGI will make these data public and downloadable following the embargo, but responsibility for depositing these data in public repositories resides with the data provider.

MycoCosm and PhycoCosm —The JGI's comparative genomics platforms for fungi and algae support user-driven filtering of comparative genomics data by characteristics of interest that are downloadable by individual users. These platforms support manual curation of genome annotations by registered users who have received specific training. These annotations are in turn served to the user community. MycoCosm and PhycoCosm may display genomes and annotations created elsewhere by mutual agreement between JGI and the owners of the external genome data. JGI will make these data public and downloadable following the embargo, but responsibility for depositing these data in public repositories resides with the data provider.

Integrated Microbial Genomes platform (IMG)—IMG supports user-driven filtering of comparative genomics data by characteristics of interest and maintains results for users. Results of comparative analyses are downloadable and may be shared with other users within the IMG system. Responsibility for providing public access to these data at the time of publication is the sole responsibility of the user. Isolate genomes, single cell genomes and metagenome datasets produced elsewhere may be uploaded for annotation and analysis within IMG. Data must be made public at the time of publication and this responsibility lies with the user.

Genomes OnLine Database (GOLD)—GOLD is an online database of worldwide genome projects and associated metadata that is maintained for the public by the JGI. The data are filterable and downloadable. Users may register their genome projects in GOLD, but must agree to make the registration and associated metadata publicly available at the time of

publication. Registration of non-JGI genome projects with NCBI or other public registries is the sole responsibility of the user.

The JGI's Policy on Microbial Genbank Submissions

Microbial genome sequencing projects become eligible for preparation for submission to Genbank when they become public in the Integrated Microbial Genomes system following the embargo. This includes sequences from both cultured organisms and single cells sequenced at the JGI only*.

All eligible projects are prepared as WGS submissions. Although the JGI has been submitting annotated genomes to Genbank in the past, moving forward we will only support WGS submissions and will serve annotations through IMG.

* Please note that genomes submitted by external users for annotation and comparative analysis in IMG are not automatically submitted to Genbank.