

The Department of Energy's (DOE) Joint Genome Institute (JGI) is one of the major publicly funded high throughput sequencing centers. The current capacity of the Production Genomics Facility (PGF) in Walnut Creek, California is approximately three billion bases per month and this year will generate up to 52 million lanes. JGI sequencing projects are initiated through several programs (<http://www.jgi.doe.gov/programs/index.html>). The three main programs for peer review of genome project proposals are the Community Sequencing Program (CSP), the DOE Microbial Program and the Laboratory Science Program (LSP). This year, the JGI processed a collection of DOE mission relevant sequencing projects ranging from prokaryotes to eukaryotes as well as several microbial communities. Data is released publicly on the JGI website and deposited in Genbank for all projects. An array of quality control measures and metrics has been put in place to evaluate projects prior to large scale sequencing to ensure efficient and timely completion of projects. This poster will present a menu of current JGI sequencing projects and describe the process of how projects are scheduled through the sequencing line, tracked and assessed for quality.

Introduction

The DOE Joint Genome Institute (JGI) is a "virtual institute" that integrates the genomic capabilities of six partner institutions: Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, Stanford University, and Pacific Northwest National Laboratory. In January 1999, high-throughput sequencing began at the Production Genomics Facility (PGF) in Walnut Creek, California, which also is home to the informatics and research and development groups.

After completing the sequencing of the Human Genome portion (Chromosomes 5,16 and 19), JGI has shifted its focus to the non-human components of the biosphere, particularly those relevant to the science mission of the Department of Energy. The menu of completed projects includes wide variety of microbes and microbial communities as well as many important eukaryotic model systems such as puffer fish (*Fugu rubripes*) and sea squirt (*Ciona intestinalis*). JGI has also sequenced a frog (*Xenopus tropicalis*), a green algae (*Chlamydomonas reinhardtii*), a diatom (*Thalassiosira pseudonana*), a white rot fungus (*Phanerochaete chrysosporium*), filamentous fungus (*Trichoderma reesei*), poplar tree (*Populus trichocarpa*) as well as a number of plant pathogens.

There are three major DOE-directed sequencing programs that utilize the high-throughput sequencing of the JGI: Community Sequencing Program (CSP), the DOE Office of Biological and Environmental Research (BER) Microbial Sequencing Program and the Laboratory Science Program (LSP). Each program targets a specific scientific community shown in Table 1.

Table 1.

Program	Targeted Community
Community Sequencing Program (CSP)	Prokaryotes, Eukaryotes, Microbial Communities
DOE Office of Biological and Environmental Research (BER) Microbial Sequencing Program	Microbial Communities
Laboratory Science Program (LSP)	Plant Pathogens, Plant Genomes

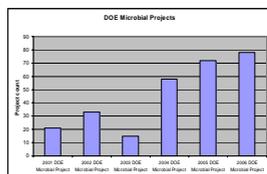
Discussion

Overview of the BER Microbial Sequencing Program

The BER Microbial Sequencing program was founded in 2001. Its focus is to provide DNA sequence infrastructure to address issues relevant to DOE missions of environmental remediation, carbon sequestration, and alternative energy production. Program candidates are microbes, microbial consortia, and organisms that are 250 Mb or smaller in size that are sequenced to 6 to 8x coverage. A subset of those organisms is identified for finishing.

As of January 2006 JGI has sequenced over 100 different microbial genomes to draft quality and 60 of those have been finished. We are currently working on 100 additional microbial projects. Figure 1 shows DOE Microbial projects broken down by year. A list of all of the Microbial sequencing projects and their status can be found at <http://microbialgenome.org/organisms.shtml>.

Figure 1.

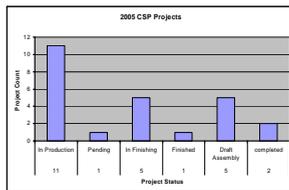


Overview of the Community Sequencing Program

The Community Sequencing Program (CSP) was created in 2005. The goal was to provide the scientific community access to high-throughput sequencing capability at the DOE's Joint Genome Institute (JGI). Sequencing projects are chosen based on scientific merit and judged through independent peer review. The CSP consists of two programs: a small-genome program for shotgun sequencing of genomes smaller than 250 Mb and other sequencing projects with a total request of less than 1 Gb. A large-genome program for shotgun sequencing of genomes larger than 250 Mb. Proposals to the large-genome program must address relevance to the DOE missions of environmental remediation, carbon sequestration, and alternative energy production. Some of the larger projects for 2006 include Sorghum spp, *Arabidopsis lyrta*, *Capsoella rubella* and *Mimulus guttatus*. A list of all of the current sequencing projects and their status can be found at <http://www.jgi.doe.gov/sequencing/cspseplans.html>.

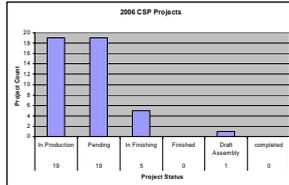
CSP 2005- program consisted of eleven microbial genomes, six basal organisms, four higher plants and animals, and four ESTs and targeted sequencing projects. Figure 2 shows current project status.

Figure 2.



CSP 2006- program consists of four large and nineteen small eukaryotic projects, and twenty-one bacteria and archaea species. Figure 3 shows current project status.

Figure 3.



Overview of the Laboratory Science Program

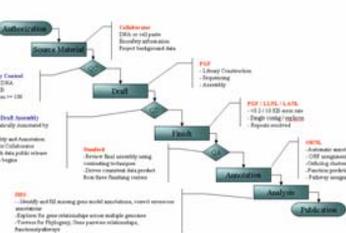
The Laboratory Science Program (LSP) is a new program that was created in FY2005, and sequencing under this program will begin in the first quarter of FY2006. It provides the DOE national laboratories with broad access to high-throughput DNA sequencing to support their biology programs. The LSP will be allocated approximately six billion bases of raw sequence in FY2006.

The program will be composed of two general types of sequencing projects: (a) large-scale, cross-national-laboratory, strategic projects that target select DOE missions and (b) small-scale sequencing that meets the needs of individual investigators at the DOE national laboratories. The large-scale projects must address one of the two selected focal areas: "genomes to energy" and "molecular response to low-dose damage". For more information see website <http://www.jgi.doe.gov/programs/LSP/index.html>.

Project Workflow

Once project is approved through one of the three programs, it is ready to enter the process. DNA sample will go through many different process steps: library construction, sequencing, quality assurance, assembly, finishing, annotation and analysis. Various quality control measures have been implemented throughout the process in order to ensure utmost quality of the DNA sample before large scale sequencing begins. Figure 4 shows different process steps and their subsequent QC steps for a microbial project.

Figure 4.



Scheduling and Tracking of Projects

Projects are scheduled for sequencing upon their approval. Each project receives a unique project ID and set of specifications that are entered into the data base. As projects move forward through the process, they are documented via LIMS at each process step. Current status of each project is tracked using excel spreadsheet. The spreadsheet outlines specific information for each project in order to track it efficiently. For example, species, strain information, date sample received, date sample plated, picked, QC pass/fail date, total number of plates, sequencing date and most current project status. See Figure 5 for more detail.

Figure 5.



Prioritization of Projects

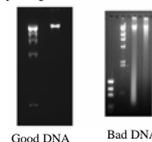
We have implemented a process for project prioritization. There are several guidelines and considerations that JGI uses to determine the order in which projects enter the sequencing line. Those are project readiness, DOE mission relevance, and rank order determined by peer review.

Quality Control Steps

Initial DNA QC

In order to construct a good DNA library, JGI requires high-molecular-weight genomic DNA (HMW gDNA) of a specific concentration and quantity. The bulk of the gDNA prep must be larger than the 23-Kb lambda Hind III size standard on an agarose gel. The specifications of HMW, concentration, and quantity of DNA are equally important for generating successful subclone libraries for sequencing. Figure 6 shows an example of good and bad DNA run on an agarose gel.

Figure 6.



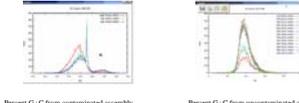
10 plate QC

Before large scale sequencing begins, 10 plates are initially sent for sequencing per project. This is a large enough data set to examine the library insert size distribution (Figure 7) and generate a GC plot to check for contamination (Figure 8). If both conditions are satisfied, project can enter the large scale sequencing effort. If a library insert size distribution is not adequate, new library will be attempted. If contamination is present, a collaborator is notified and new DNA sample will be sent to JGI.

Figure 7.



Figure 8.



Sequence Assessment and Analysis

Once 10 plate QC is passed the large scale sequencing of a project can begin. Consensus sequences are assembled by Phrap or Jazz assembler, depending on the size of the project, to examine the quality of the assembly. Source DNA is checked for contamination once again. Since every organism has a unique percentage of its genome made up of Gs and Cs (G+C content), contamination can be identified by plotting this distribution. Suspicious G+C plots are then verified by performing a BLAST search. This program is used to compare JGI sequences to the known sequences of other organisms. Hits to closely related organisms validate the source DNA, whereas hits to distantly related organisms may represent contamination. Projects that are found to have contamination are taken out of the process. Collaborator is informed and a new, more purified sample will be sent to JGI. Once the draft assembly has passed quality assessment, the sequence is submitted to NCBI's GenBank for public use.

For larger eukaryotic projects (>30Mb), there is one more QC step at 2x sequencing depth. At this step sequencing results together with depth coverage determine how much more sequencing needs to be done to complete the project.

JGI Web site information

The JGI makes high-quality genome sequencing data freely available to the greater scientific community through its web portal. For eukaryotic projects go to <http://genome.jgi-psf.org>, and for microbial projects go to http://genome.jgi-psf.org/mic_home.html. From the web portal site you can obtain details about our past, current and upcoming projects or go directly to the individual genome sites http://genome.jgi-psf.org/uk_curl.html. All of the individual sites include direct access to download sequence files, BLAST, search, view and ability to navigate the genomic annotations.

The Integrated Microbial Genomes (IMG) system site provides a framework for comparative analysis of the genomes sequenced by the Joint Genome Institute. Its goal is to facilitate the visualization and exploration of genomes from a functional and evolutionary perspective. The user interface for IMG 1.3 was released on Dec. 1st 2005. <http://img.jgi.doe.gov/cgi-bin/pub/main.cgi>

Sequencing Statistics: you can view total number of lanes and Q20 bases generated updated hourly <http://www.jgi.doe.gov/sequencing/statistics.html>.

Sequencing plan and progress can be viewed at <http://www.jgi.doe.gov/sequencing/seqplans.html>.

Sequencing protocols can be accessed at <http://www.jgi.doe.gov/sequencing/protocols/>.

Conclusion

JGI has gone through a major transition from sequencing human genome using BAC by BAC approach to sequencing many different genomes using whole genome shotgun approach. The three major scientific programs allow a wide variety of projects to enter the sequencing pipeline and address DOE mission as well as provide the scientific community with high throughput DNA sequencing capability. All projects are entered into the data base and tracked and scheduled through the process. Scheduling and tracking of projects ensures meeting established sequencing timelines and that no project is left behind. Quality control measures implemented at different steps ensure sample quality and prevent contaminants from being present in the final product. JGI will continue to provide integrated high-throughput sequencing and computational analysis to enable genomic-scale/systems-based scientific approaches to DOE-relevant challenges in energy and the environment.

Comments on poster