

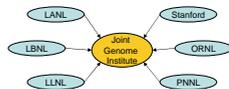
Abstract

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. This poster will present current JGI sequencing projects and describe process steps necessary for executing a project from initiation to completion. We will present quality control measures and metrics that have been implemented at different steps in the process to evaluate projects prior to large scale sequencing. This will include project initiation and specification, project management, and sequencing data analysis. The ultimate goal is to ensure quality data, efficiency and the timely completion of projects.

Introduction

The DOE Joint Genome Institute (JGI) is a "virtual institute" that integrates the genomic capabilities of six partner institutions: Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), Oak Ridge National Laboratory (ORNL), Stanford University, and Pacific Northwest National Laboratory (PNNL) (Fig. 1). 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.

Figure 1.



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 a 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 alga (*Chlamydomonas reinhardtii*), two diatoms, a white rot fungus (*Phanerochaete chrysosporium*), filamentous fungus (*Trichoderma reesei*), poplar tree (*Populus trichocarpa*) as well as a number of plant pathogens. The current capacity of the PGF is approximately three billion bases per month and this year will generate up to 52 million lanes.

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 (OBER) Microbial Sequencing Program and the Laboratory Science Program (LSP).

Discussion

Overview of the OBER, CSP, and LSP Sequencing Programs

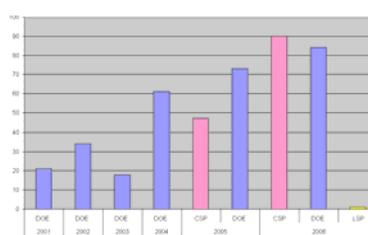
The OBER Microbial Sequencing Program began in 2001. Its focus is to provide DNA sequence infrastructure to address issues relevant to DOE missions of environmental remediation, carbon sequestration, and alternative energy production. Program candidates are microbes, microbial consortia, and organisms that are 250 Mb or smaller in size that are sequenced to 6 to 8x coverage. A subset of those organisms are 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 (Fig. 2). We are currently working on 100 additional microbial projects. A list of all of the Microbial sequencing projects and their status can be found at <http://microbial.jgi.doe.gov/organisms.shtml>.

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 read length less than 1 Gb. A large-genome program for shotgun sequencing of genomes larger than 250 Mb (Fig. 2). A list of all of the current sequencing projects and their status can be found at <http://www.jgi.doe.gov/sequencing/cspseplans.html>.

The Laboratory Science Program (LSP) is a new program that was created in FY2005. It provides the DOE national laboratories with broad access to high-throughput DNA sequencing to support their biology programs. The LSP has been allocated approximately six billion bases of raw sequence for this fiscal year. The program is 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 visit <http://www.jgi.doe.gov/programs/LSP/index.html>.

Figure 2: Total Program Projects by Year

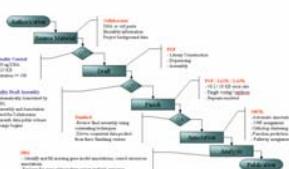


Project Workflow

Once a project is approved through one of the three programs, it is ready to enter the process. A statement of work is prepared and approved by the project committee and the collaborator. This document lists the project budget, targeted sequencing and finishing dates and the overall goal of the project. In addition, a project specification document is generated that details all relevant information that is required to ensure the success of the project. Such information includes: organism background information, collaborator information, the completed statement of work, and information specific to library creation, sequencing, assembly, finishing and annotation. Once the documentation has been completed, the JGI is ready to receive the gDNA from the collaborator.

The 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 the utmost quality of the DNA sample before large scale sequencing begins. Figure 3 shows different process steps and their subsequent QC steps for a microbial project. The ultimate goal is to have all 3 libraries run concurrently through the process so that all of the data is ready for final assembly and analysis.

Figure 3.



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. Projects are prioritized based on project readiness, DOE mission relevance, and rank order determined by peer review.

As projects move forward through the process, they are documented via LIMS at each process step. The current status of each project is tracked by each group using an excel system. 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 and picked, QC pass/fail date, total number of plates needed, sequencing date and most current project status. This system is soon being replaced by a Global Project Tracking System (GPTS), a database application, that is being designed in house by the informatics team.

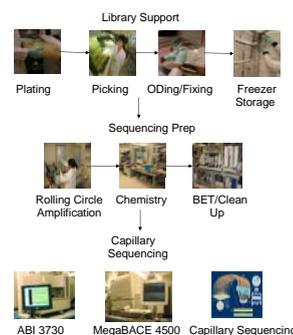
Production Process

Our current sequencing strategy is to shotgun sequence 3 and 8kb libraries to 8x draft coverage and to sequence 40kb fosmid libraries to 0.5x sequence coverage. For each project we construct 3 libraries: 3 and 8kb are cloned into pCC18 and pMCL200 vector respectively and 40kb are cloned into pCC18s vector. All three libraries are created in the Cloning Technology group and passed on to Library Support group as transformation stocks. In Library Support, the transformation stocks are plated on large agar bioassays and the individual colonies are picked using automated Genetix Qx colony picking machines. The colonies are inoculated into a 384 well destination plate containing LB glycerol and an antibiotic. The fully grown plates are held and QC'd to ensure all wells contain an adequate amount of cells and there is no contamination. (Refer to Fig. 4)

The plates are then passed to Sequencing Prep group where the cells are lysed and the plasmid DNA is amplified using Rolling Circle Amplification (RCA). The amplified DNA is aliquoted into 2 new plates to which the forward or reverse direction sequencing chemistry is added. The plates are placed on thermocyclers to allow the Sanger reaction to proceed. The samples are then cleaned using the modified magnetic bead protocol in order to remove any leftover unincorporated dyes, salts, and nucleotides along with cell debris and other waste by-products (Refer to Fig. 4).

The clean purified DNA is transferred to a new plate and ready to be loaded onto one of two Capillary Sequencing platforms (Refer to Fig. 4). We sequence plates on either the ABI 3730 or MegaBACE 4500 instruments. The data generated is assessed and analyzed by the Quality Assurance (QA) team.

Figure 4. Overview of the Production Process



Quality Control Steps

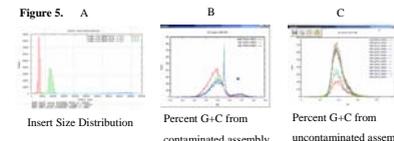
All projects undergo the following Quality Control Steps:

Initial DNA QC

Under DNA preparation guidelines, 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.

Library QC

Before large scale sequencing of a project begins, 10 plates are initially sent for sequencing. The objectives of library QC are to determine the insertless clone rate, estimate the insert size and distribution, access contamination levels and attempt to confirm that the correct organism has been cloned. 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. Overall sequence quality of the library is also examined based on the average Q20 readlength and pass rate. Once 10 plate QC is passed the large scale sequencing of a project can begin. Figure 5A, B and C show insert size distribution, percent GC from a contaminated assembly and percent GC from an uncontaminated assembly respectively.



Assembly QC (QD)

Consensus sequences are assembled by Phrap (projects < 30 MB) and by Jazzer assembler (projects > 30MB) in order to examine the quality of the assembly. The core of the assembly QC has been fully automated and is carried out by a single script which runs BLAST jobs, parses and filters the output, converts filtered hits into lists of reads to exclude from the final assembly, produces a clean fasta file and a summary report. Project or source DNA contamination is identified again by assessing GC plots of all reads. Low quality reads are excluded from the final assembly (less than 100 total phred Q20 bases). Once the draft assembly has passed quality assessment, the sequence is submitted to NCBI's GenBank for public use.

For larger projects (>30MB), there is one more QC step at 2x sequencing depth. At this step, sequencing results together with depth coverage and the average sequencing readlength and pass rate for the project, 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 genomic annotations.

For all microbial and metagenomic projects, 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 in December 2005. <http://img.jgi.doe.gov/cgi-bin/pub/main.cgi>

Genome Analysis and System Modeling Group of the Life Sciences Division of Oak Ridge National Laboratory conducts genetics research and system development in genomic sequencing, computational genome analysis, and computational protein structure analysis. They provide bioinformatics and analytic services and resources to collaborators, predict prospective gene and protein models for analysis, provide user services for the general community, including computer-annotated genomes in Genome Channel. <http://genome.ornl.gov/microbial/tdm/>

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 to 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 genome-scale/systems-based scientific approaches to DOE-relevant challenges in energy and the environment.

