JGI Sequencing Projects: Statistics and Timelines

Authors: ***Tijana Glavina del Rio***, Kerrie Barry, Lynne Goodwin, Miranda Harmon-Smith, Harris Shapiro, Susan Lucas and David Bruce.

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, generating a total of 55 million lanes this year. JGI sequencing projects are initiated through one of three peer reviewed programs: Community Sequencing Program (CSP), DOE Microbial Program and the Laboratory Science Program (LSP). This poster will present an overview of project statistics for 2006 and current projects for 2007. In 2006, the JGI processed a collection of DOE mission relevant sequencing projects ranging from prokaryotes to eukaryotes as well as several microbial communities. The poster will also describe how projects are scheduled for production sequencing and display tools used for tracking projects to their completion. Project timeline from initiation to completion will also be presented.

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. 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), two diatoms (Phaeodactylum tricornutum and Thalassiosira pesudonana), a white rot fungus (Phanerochaete chrysosporium), many filomentous fungi (Trichoderma reesei, Aspergillus niger, Laccaria bicolor, Mychosphaerella graminicola, Nectria haematococca, Phycomyces blakesleeanus), poplar tree (Populus trichocarpa) as well a number of fungal- like plant pathogens (Phytophthora ramorum and Phytophthora sojae). There are three major DOE-directed sequencing programs that utilize the high-throughput sequencing of the JGI: **Community Sequencing Program (CSP), DOE Office of Biological and Environmental Research (BER) Microbial Sequencing Program Laboratory Science Program (LSP)**.
You can find more information about these programs on our website: http://www.jgi.doe.gov/programs/index.html

Sequencing Strategy

PGF sequencing strategy employs the whole-genome shotgun sequencing method to produce high-quality draft sequence. For each project a 3-Kb, 8-Kb, and 40-Kb DNA library is created and sequenced from both sides of the library insert, producing paired ends, resulting in approximately 8-9X depth. In support of the eukaryotic projects, cDNA libraries are generated in order to understand gene structure of the genome and help with annotation efforts. For each large genome 15 fosmid clones are selected and subcloned for better size estimation. Recently, we have changed the sequencing strategy for the prokaryotic projects by replacing a 3kb library with a 454 library for additional coverage and increased coverage for the 8kb library from 4x to 6x. For all projects, sequenced reads are deposited in the GenBank Trace Archive at NCBI. Reads are aligned by using various genome assemblers to produce the primary draft assembly, which consists of contigs linked into larger scaffolds by paired-end information. The PGF and several partner institutions (Stanford University, Los Alamos National Laboratory, Lawrence Livermore National Laboratory and Lawrence Berkeley National Laboratory) perform finishing work (gap closing, quality improvement, and assembly verification) for organisms that require this level of refinement. All genomes have at least a minimal automated annotation, and most may be searched on genome browsers via the Genome Portal.
Eukaryotic Genome Portal link: http://genome.jgi-psf.org/.
For Prokaryotic Genome Portal link: http://genome.jgi-psf.org/mic_home.html

Project Flow

Once a project is approved through one of the three programs, it is ready to enter the sequencing process. The starting material can be converted to four different library types, depending on the project scope: WGS library, 16/18S library, EST library and 454 library. Samples will go through the following 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 and constructed library before large scale sequencing begins. Figure 1 shows the JGI sequencing project pipeline on a high level. If a project requires more then one library type to be constructed, the ultimate goal is to have all libraries for each project run concurrently through the process so that all of the different data is ready for final assembly and analysis at the same time.

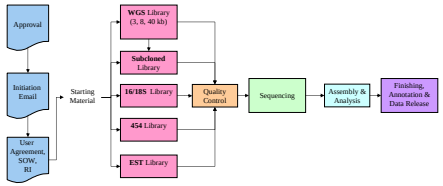


Fig 1. JGI Sequencing Project Pipelines

Project Management

As high throughput sequencing centers move from managing a small number of large projects to managing many simultaneous small projects, the ability to govern schedule, cost, quality, and project specification becomes more difficult. To better handle the tracking, organization and flow of sequencing projects through the JGI sequencing pipeline, the JGI Project Management Office (PMO) was formed in January, 2006. The PMO consists of a team of five project managers. Each project manager is dedicated to a group of projects based on specific microbial communities. Throughout the life cycle of a sequencing project, the project managers oversee its stepwise progression throughout the JGI sequencing pipeline with particular emphasis on active, ongoing communications with the PI and associated collaborators. The JGI Global Project Tracking System (GPTS) database is the project manager's primary resource for tracking on a project level and communicating project progress through the sequencing pipeline process to both the internal and external collaborators. For tracking projects on a library level, a simple excel spreadsheet is used for this purpose (Fig 2 and 3). This is necessary in order to know the status of a project and all of its many components.

Fig 2.

Fig 3.

Prioritization and Scheduling of Projects

The following considerations and guiding principles are used to determine the order in which projects enter the pipeline for sequencing.

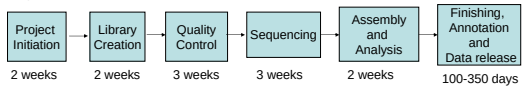
1. Project Initiation completed-User Agreement signed, Statement of Work (SOW) finalized, Required Information (RI) sheet received from collaborator.
2. DNA sample quality- DNA is received and passed DNA QC.
   If the first two requirements are satisfied, the project is scheduled for sequencing on a first come first serve basis.

Scheduling of Projects:
The production line runs at the monthly average rate of 4.6 million lanes and 2.9 Billion bases. Medium (10-100MB) and Large genome projects (>100MB in size) are scheduled at a 60% production capacity and everything else is scheduled as a filler. This way we can ensure we are always running a few of the medium and large genome projects with lots of smaller projects in the background. Small projects are scheduled on a weekly basis based on availability (passing libraries from the Cloning group). This scheduling method helps keep large and medium projects on target and is easier to manage due to the limited amount of total projects in the queue for tracking.

Project Timeline:

The average time it takes for a sequencing project to be completed varies significantly. This is generally due to the project scope of work, sample type (EST, 16S, WGS, clone or 454), DNA sample quality, genome size, requested coverage, special sequencing and analysis requests. For the purpose of generating timelines and forecasting project completion dates, we use a project such as a microbial project for our estimates. Currently, it takes an average of 12 weeks (3 months) to get this type of project completed. This average does not include projects that fail library QC or for any other reason and are sent back to the beginning of the process. Fig 4. shows an approximate project timeline for a microbial project.

Fig 4.



Status of 2006/2007 Large Eukaryotic Projects

Program/Year	Genus Species	Status
CSP 2006	Arabidopsis lyrata	Production Final 8x
CSP 2006	Capsella rubella	Production Lib QC
CSP 2006	Mimulus guttatusIM62	Production Final 8x
DOE 2006	Pseudo-nitzschia multiseries CLN-47	Awaiting Sample
CSP 2006	Sorghum bicolor	Final 8x Assembly
DOE 2006	Dunaliella salina UTEX-1644	Awaiting Sample
CSP 2007	Aquilegia formosa	Awaiting Sample
CSP 2007	Brachypodium distachyon	Production Lib QC
CSP 2007	Gossypium raimondii	Production Lib QC
CSP 2007	Manihot esculenta	Production Lib QC
LSP 2007	Glycine max Williams 82	In 4x Assembly
CSP 2007	Thellungiella (salt crest)	Awaiting Sample

Fig 5. JGI finishing effort includes three teams: PGF, LLNL and LANL. The JGI attempts to finish the genome sequence for all prokaryote projects. In 2006, JGI has completed 123 genomes to its finished quality.
Fig 6. shows the status of 2006 projects. We have 28 projects in production, 9 in final assembly and 25 pending.
Fig 7. shows the status of 2007 projects. We have 20 projects in production, 91 pending and 2 completed.

Fig 5. Finished Projects in 2006

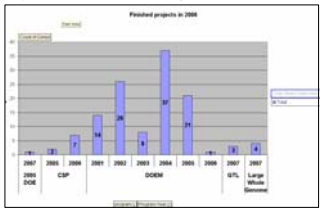


Fig 6. JGI Project Status 2006

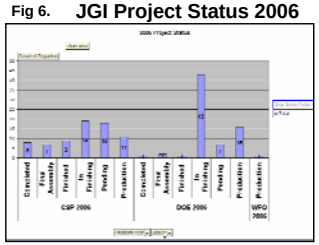
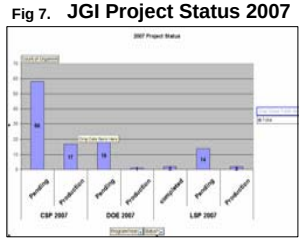


Fig 7. JGI Project Status 2007



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. As we grow, 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.

UCRL-POST-229322
LBNL-62524 Poster

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.