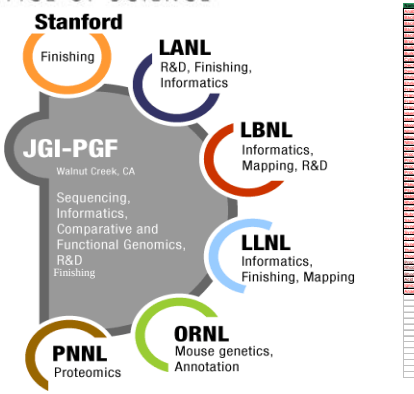
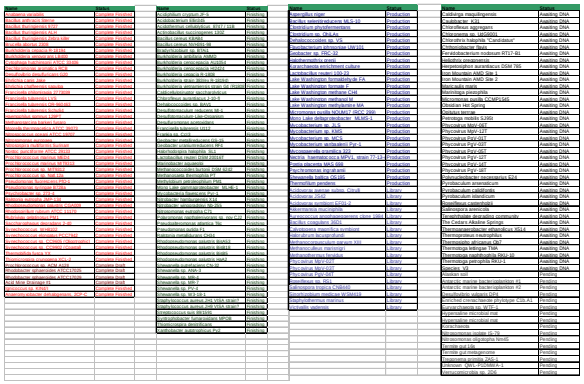
Microbial Finishing at JGI

Alla Lapidus, Patrick Chain, Cliff Han, Eugene Goltsman, Michele Martinez, Stephanie Malfatty, Olga Chertkov, Stephan Trong, Tom Brettin, Roxanne Tapia, Alex Copeland, Paul Richardson.

*Correspondence alapidus@lbl.gov

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE





Finishing Statistics

Status	Total	Description
Complete Finished	44	Project completed to finish standards
Complete Draft	4	Project completed to draft standards
Finishing	55	Draft phase complete. Finish phase underway
Production	27	Libraries completed and draft sequencing underway
Library	19	Libraries under construction for drafting
Awaiting DNA	40	Awaiting DNA for initiation of library construction
Pending	16	Statement of Work incomplete for project
Grand Total	205	

Taxa	Total
archaea	16
bacterium	150
environmental s	19
eukaryon	7
symbiont	2
virus	11
Grand Total	205

Microbial Finishing steps

- QD -> draft assembly
- Repeat resolution (automated and manual)
- Primer walk
- More repeat resolution
- Assembled genome -> closed
- Polishing (quality improvement) -> polished
- Assembly QA (Stanford) -> FINISHED

Finishing Check List/Standards

- All low quality areas (<Q30) reviewed and re sequenced.
- Final error rate < 0.2 per 10 Kb.
- No single clone coverage (minimum of 2X depth everywhere)
- Manually inspected and quantified single stranded regions.
- Checked all high quality discrepancies.
- Final sequence has no strings of xxxx anywhere
- All repeats verified (paired ends and PCR if necessary).
- Check ends of final contigs (chromosomes, plasmids)
- Final Assembly QC

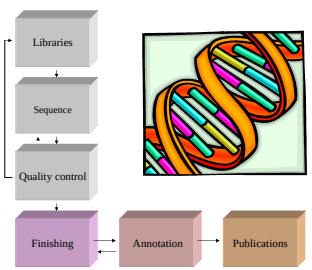
The U.S. Department of Energy (DOE) established the Microbial Genome and GTL programs to determine the complete genome sequence of a number of microbes that may be useful to DOE in carrying out its missions (which include research of new energy sources, sequester excess atmospheric carbon affecting global climate, and to clean up contaminated environments). Another program was established to study hard-to-culture individual microbes and microbial communities. They are very difficult to study but play critical roles in the Earth's ecology.

JGI is a leader in performing sequences to support these programs. Another important specialization of the JGI is the recently started Community Sequencing Program (http://www.jgi.doe.gov/CSP/index.html).

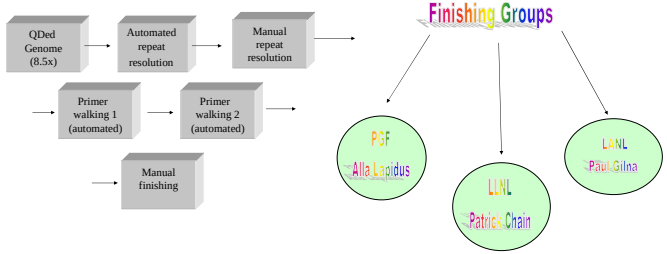
All of the above projects include sequence and detailed analysis of the genomes of the different representatives of the microbial world. A completely sequenced, high quality genome is a perfect starting point for the genome annotation (http://img.jgi.doe.gov/v1.1/main.cgi), microarrayes, knockout experiments etc. Despite the fact that drafted genome contains a sufficiently large amount of information to be usable, a completed and polished one is overall a better product especially if it will be used to analyze previously unknown and difficult-to-cultivate microbes; for the comparative analysis of the clinical isolates or for the creation of microbial strains overproducing different proteins and amino acids. Knowledge of the completely finished genome will allow scientists to modify specific regions of the genome and therefore to affect the expression of the gene being studied. The study of the potential usage of microorganisms as new energy sources requires a complete knowledge of the genomes sequence as well.

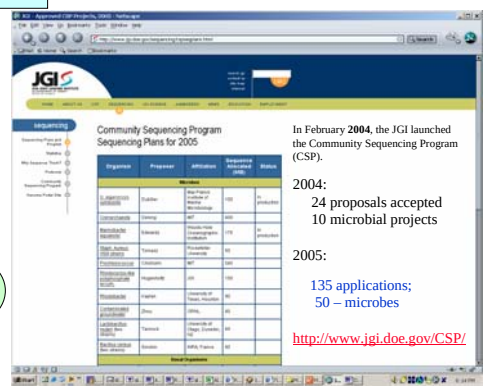
Thus in order to be able to realize these and many other studies, it is necessary to close most (if not all) of the genomes being sequenced at JGI. To this date JGI has accumulated a significant amount of experience closing the microbial genomes. During the last two years the combined efforts of three groups (LANL, LLNL and PGF) allowed to finish 45 microbial projects. Our goal is to fulfill the needs of all of the projects undertaken by JGI and complete no less than 50 genomes per year.

Genomics Project



Microbial Finishing





In February **2004**, the JGI launched the Community Sequencing Program (CSP).

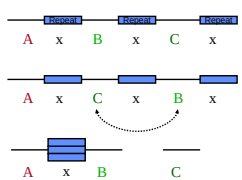
2004:
24 proposals accepted
10 microbial projects

2005:
135 applications;
50 – microbes

http://www.jgi.doe.gov/CSP/

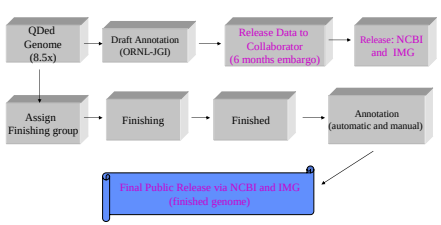
The Problem With Repeats

- Repeats can make assembly ambiguous, with several possible layouts being equally likely
- Repeats can be up to 5 kb and more in length - plasmid clone inserts are too short to span them (~3 kb). Libraries with larger inserts may provide an answer



An average sequencing read is ~650-700 bp. When in a repeat it is impossible to tell which copy it belongs to

Data flow - Data Release



(Poster: "Integrated Microbial Genomes (IMG): New Data and Functionality in Version 1.2", Victor M. Markowitz, Frank Korzeniewski et.al)

Finished Genomes: 44 Draft Genomes: 128

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. W-7405-ENG-36. LBNL-58674 Poster

US DOE Joint Genome Institute