Title

JGI Microbial Sequencing Process

Permalink

https://escholarship.org/uc/item/2hz6536f

Authors

Rio, Tijana Glavina del
Daum, Chris
Deshpande, Shweta
et al.

Publication Date

2009-05-27

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

JGI Microbial Sequencing Process

Tijana Glavina del Rio, Chris Daum, Shweta Deshpande, Alla Lapidus, Matt Nolan, Nicole Shapiro, Hope Tice, Lynne Goodwin, David Bruce, Susan Lucas.

Abstract:

JGI Microbial process has undergone through major changes in the last year. The implementation of the new sequencing technologies, Roche 454 and Illumina, have been the key factor in migrating the microbial draft sequencing from the older Sanger pipeline to the new 454 and Illumina pipelines. This poster will present the current sequencing process for the microbial genomes at the PGF, Walnut Creek facility. Scope of Work for microbial projects will be discussed as well as the production workflow process steps from the DNA receipt to finishing. A brief summary of each process step will be provided as well as important statistics for the area. Scheduling will be addressed to show how the two separate pipelines are managed for project scheduling and synched in order to produce a complete dataset at the end. The poster will also provide information on the current number of microbes in process and the PMO forecast for the year.

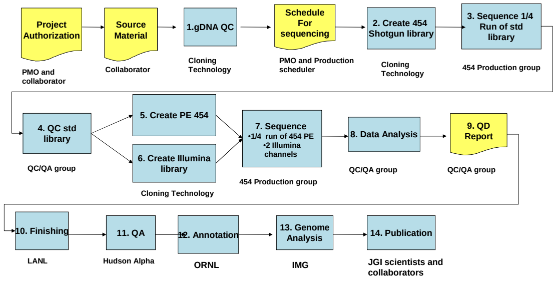
Introduction:

JGI engages users from the National Laboratory system, academic institutions, and private industry to perform sequencing projects that directly relate to the DOE Office of Biological and Environmental Research's mission in alternative energy, global carbon cycling, and bioremediation. JGI's user sequencing portfolio is drawn from its user programs, including the Community Sequencing Program, the new Bioenergy Research Centers, as well as more recently, a pilot project called GEBA, in order to meet programmatic objectives. Microbes sequenced by DOE JGI have far-reaching implications for addressing such DOE mission challenges as the remediation of radioactive and hazardous waste sites, sequestering heat-trapping carbon from the atmosphere, and developing renewable energy sources. Most microbes of interest to DOE are sequenced at JGI. **JGI is currently producing about 22% of the reported number of bacterial genome projects worldwide.**

Sequencing Strategy

Current JGI drafting strategy includes creation and sequencing of the 454 standard titanium, 454 paired end titanium, and Illumina data for gap closure and quality improvement (polishing). The 454 standard and PE libraries are sequenced to approximately 20x depth. An Illumina library is sequenced to 50x depth in order to reduce costly finishing reactions. For the 454 libraries, we run a quarter run for each library and for Illumina, we run 2 channels of the flow cell. In order to create a draft assembly that will ultimately be passed to finishers, the 454 reads are assembled using the Newbler assembler and Illumina using the Velvet assembler . For all projects, sequenced reads are deposited in the GenBank Trace Archive at NCBI. Los Alamos National Laboratory performs 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. For Prokaryotic Genome Portal link: http://genome.jgi-psf.org/mic_home.html

Figure 1. JGI Microbial Sequencing Process



Project Scheduling:

Scheduling of Microbial projects starts once DNA material has been received. All DNA samples go through the DNA QC process on a first come first serve basis. Generally, the QC is done once per week for all samples that arrive that week. Once DNA sample passes the QC, it is scheduled for library construction. The 454 standard library is constructed first to serve as a QC for the DNA sample material. Once it passes the sequencing QC, project is scheduled for a PE and Illumina library construction. The sequencing of the Illumina and 454 PE library is scheduled simultaneously as it is important that both components are completed at the approximately same time so there is no delay for the project.

DNA Library Preparation (454 shotgun library) (Fig 1, Step 2)

Current Throughput: 12-24 libraries per person per week.

Library Preparation:

Genomic DNA is sheared into randomly fragmented library of DNA fragments by ***nebulization***. Gel electrophoresis is run to select the DNA fragments of the appropriate size (500-800 bp). The dsDNA fragment frayed ends are ***polished and purified***. The fragments are then ***ligated to A & B adaptors*** (with PCR primer binding sites) and immobilized onto strepavidin-coated magnetic beads. The dsDNA that are bounded to the beads are denatured with NaOH and the complementary non-biotinylated strands containing a ligated A & B adaptor sequence are released and collected as single stranded DNA for the next step: Emulsion PCR.

DNA Library Preparation (454 PE library) (Fig 1, Step 5)

Current Throughput: 4 libraries per person per week.

Library preparation:

Genomic DNA is sheared to desired size and ligated to loxP adaptors and circularized via recombination by a Cre excision reaction. The circular DNA templates are then randomly fragmented by nebulization, followed by streptavidin-mediated capture of the desired biotinylated fragment and ligation of adaptors for amplification. The paired end library is then amplified via PCR using a primer set, one of which is biotinylated. Following AMPure (SPRI) bead size exclusion and streptavidin magnetic bead immobilization of dsDNA amplicon, ssDNA templates are isolated by alkaline treatment. The resulting ssDNA Paired End library is suitable for input to emPCR and subsequent sequencing, on the Titanium platform.

454 Sequencing Process (Fig 2)

Current Throughput: 10 runs per week. Each titanium run gets an average of 340MB and an average read length of 349.

Sequencing Process:

Emulsion PCR step: the goal of this step is to achieve a capture bead with clonally amplified fragments attached that originated from one fragment of the DNA library.

Breaking: after the amplification step, the emulsion is broken chemically and the beads carrying the amplified DNA library are recovered and washed to remove the emulsion oil and any unincorporated nts and polymerase molecules.

Enrichment: the goal of this step is to enrich the total bead population for amplified DNA-carrying beads, which is done by removing beads that do not have DNA fragments bound to them or did not amplify well.

Primer Annealing: In this step we also anneal sequencing primers to the bead immobilized sstDNA fragments.

Sequencing run: procedure is comprised of three main steps: Pre-wash run, PTP Preparation and the Sequencing Run. Input in this process is a library that is clonally amplified, bead immobilized sstDNA fragments with annealed sequencing primers. Output is a set of digital images ready for processing and elucidation of sample sequence.

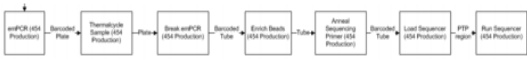

emPCR (454 Production)
Barcoded Plate
Thermalcycle Sample (454 Production)
Plate
Break emPCR (454 Production)
Barcoded Tube
Enrich Beads (454 Production)
Tube
Anneal Sequencing Primer (454 Production)
Barcoded Tube
Load Sequencer (454 Production)
PTP region
Run Sequencer (454 Production)

Figure 2. JGI 454 Sequencing Process Flowchart

Illumina Sequencing Process (Fig 1, Step 7)

Current Throughput: Production has 5 Illumina machines in operation. Illumina GA*II* instrument can generate over 1 billion bases of DNA sequence per 36-cycle standard run and over 4 billion bases per paired-end run. For our microbial projects, the amount of sequencing we generate based on 2 channels from the flow cell and a 36 cycle single read run is on average 689 MB. Our weekly maximum run capacity for 36 cycle single read runs is 8 per week which would require an input of 28 libraries. Currently we run on average 2-3 runs per week (7-11 projects per week).

Sequencing Process:

Cluster Generation: consists of the following steps; hybridization, amplification, linearization, blocking and primer hybridization. During cluster generation, a single DNA fragment is attached to the surface of an oligonucleotide coated flow cell and amplified to form a surface-bound colony (the cluster). The result is a heterogeneous population of clusters, with each cluster consisting of many identical copies of the original template molecule. All of these reactions occur on a cluster station.

Sequencing-by-synthesis (SBS): analyzer run is a combination of SBS and reverse-terminator-based sequencing. Only one base per cycle is incorporated, therefore the read length is limited to the number of cycles performed.

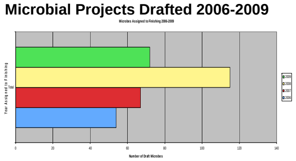
Paired-End Sequencing: When performing a paired-end run, after the initial cycles (Read1), an additional cluster generation is performed on the analyzer, and the template is sequenced in the opposite direction. After Read1, the DNA present is double stranded, consisting of the template strand and the sequence strand. The sequence strand is stripped off, and the 3' ends of the template and extra primers are unblocked. Bridge amplification takes place on the analyzer for Read2. The bridges are linearized, and the original forward template is cleaved off and washed away. The template DNA is then sequenced in the opposite direction.

Data analysis: There are three tools that can be utilized to asses the performance of a run: Run Browser, IPAR and Genome Analysis Pipeline. JGI uses specific aspects of each tool and the data metrics they provide to asses the quality of a sequencing run.

Microbial Forecast for FY2009-2011

Count of Genus	Program Name and Year														
Estimated Delivery	BESC 2008	BESC 2009	CSP 2007	CSP 2008	CSP 2009	DD 2007	DOEM 2007	GEBA 2007	GLBRC 2008	GLBRC 2009	GTL 2007	JBEI 2008	LSP 2007	GEBA - PCC 2009	Grand Total
FY09 Q3								30	5					15	50
FY09 Q4								15	6	8				15	44
FY10 Q1	4	2		1	8	1	1	15		15				5	52
FY10 Q2			5					15		16		7			43
FY10 Q3				3				15		15	2	7	2		44
FY10 Q4								15		15					30
FY11 Q1								16		25					41
FY11 Q2										20					20
FY11 Q3										23					23
FY11 Q4										24					24
Grand Total	4	2	5	4	8	1	1	121	11	161	2	14	2	35	371

Microbial Projects Drafted 2006-2009



Current Project Status

Production Status	# of Microbial Projects
Awaiting Lib construction	46
In Lib construction	5
In PE lib queue (not started)	35
In 454 sequencing	13
Ready for 454 sequencing	29
Failed	0
Ready for QD	20
Problem Projects-need analysis	13
Total	161

Microbial Quality Analysis (Fig 1, Step 4)

The QA group works closely with finishers to insure that projects have met internal quality specifications before finishing work begins. They also work closely with production staff and collaborators to identity mix ups and contamination and to identify bad libraries before investing significant resources into sequencing them. The goal of the JGI sequencing effort is to provide high quality sequence data to the public in a timely and cost effective manner and ultimately, annotated assemblies of microbial projects containing the absolute minim of errors. The QA group has a key role in defining and maintaining project quality standards.

Initial QC of microbial projects:

The first step of a microbial QC process is to QC the 454 standard library. The process starts by checking for gDNA contamination. GC content analysis is used to identify possible contamination and mixed DNA sources in a project. Multi-modal GC content and contig depth graphs are typically indicative of multiple genomes. The genome of interest is confirmed by presence of megablasts hits to a 16S and NT databases as are potential contaminants. Previously when gDNA contamination is suspected, a request for a new gDNA sample was made. Now if the genome of interest is sufficiently dominant and distinct enough from the contaminant, sequencing proceeds, as only on occasion are subsequent samples devoid of contamination. When we suspect JGI process contamination we perform a megablast against the other projects the genomes of interests libraries were created with. Our Illumina data rarely contains JGI process contamination so having this dataset also helps discern between gDNA and process contamination in the 454 dataset. Read and contig depth distribution is also reviewed to ensure sufficient coverage over the whole genome, not just in aggregate. The 454 shotgun assembly is also checked for excessive number of sequencing gaps, in which case more may be ordered. The sequencing quantity and type of the 454 PE library ordered depends on repeat content, contig depth distribution, and GC content. We must quantify the redundant PEs to assure we have acquired a sufficient quantity of PEs to properly scaffold and accurately solve the repetitive regions of the genome. In the absence of sufficient PEs, the same library may be run again if the redundancy rate of PE reads is low, typically though a new PE library must be constructed. At the end of the analysis, a report is submitted which includes all of the project's performance metrics. This report also contains the amount of data collected for each library as well as a request for more specific sequencing in most cases.

QD: Final QC of microbial projects (Fig 1, Step 8)

This stage starts upon completion of data acquisition. That includes 454 standard and PE library and Illumina library. Before a microbial project can be sent to the finishers, contamination must be removed. It is important to recognize and remove contamination from a project during this phase. One way to do this is to compare two datasets for the same project. If the contamination is present to just one dataset, then typically we can be confident it should be removed. We also remove any redundant reads, primarily identifiable in the PE libraries. Illumina contig data output from the Velvet assembler is integrated into the 454 newbler assembly by using shredded contigs. Including them helps close gaps in the high GC regions of the genome.

Microbial Finishing: (Fig 1, Step 10)
JGI Finishing process:
Once a draft assembly has been generated and QDed, it is ready for finishing to begin. An in-house developed software tool creates subprojects for each gap. In-silica attempts are made to close gaps using existing unassembled pyrosequence and Illumina data. Any remaining gaps are tackled by PCR based methods. These include standard PCR, bubble PCR, multiplex PCR, combinatorial PCR, and long range PCR. Once products are generated they can be sequenced, cloned or shattered as needed. Currently gap closing data is still generated using Sanger. Once a genome is closed, Illumina data is used to polish the genome. Any areas that are still substandard are subjects for resequencing.

LANL Finishing Process:
During finishing, the initial draft assembly contains many contigs and scaffolds. During the first round of finishing, the initial 454 assembly is converted into a phrap assembly by making fake reads from the consensus, collecting the read pairs in the 454 paired end library. The Phred/Phrap/Consed software package is used for sequence assembly and quality assessment in the following finishing process. After the shotgun stage, reads are assembled with parallel phrap. Possible mis-assemblies are corrected with gap resolution (Cliff Han, unpublished), Dupfinisher (Han, 2006), or sequencing cloned bridging PCR fragments with subcloning or transposon bombing (Epicentre Biotechnologies, Madison, WI). Gaps between contigs are closed by editing in Consed, by PCR and by Bubble PCR primer walks. Additional reactions are necessary to close gaps and to raise the quality of the finished sequence. The final genome sequences is completed with an error rate less than 1 in 100,000 bp.

IMG: (Fig 1, Step 13)

IMG is a powerful data management platform that supports timely analysis of genomes from a comparative, functional, and evolutionary perspective. The key mission of the IMG system is to provide a data management platform that supports timely and comprehensive analysis of JGI-sequenced genomes in a comparative genomics context. IMG aims to provide a high level of system and data comprehensibility, in terms of documentation and clarity regarding data and structural and operational semantics, and eventually to contribute to the common goal of improving the overall quality of microbial genome data for the scientific community.

LLNL-POST-413434

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 Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.