Lawrence Berkeley National Laboratory

Recent Work

Title

DOE Joint Genome Institute Production Genomics Facility (PGF) Finishing Pipeline

Permalink

https://escholarship.org/uc/item/8vp5k4dv

Authors

Sun, H.
Clum, A.
Goltsman, E.
et al.

Publication Date

2007-05-04

eScholarship.org Powered by the California Digital Library
University of California

DOE JOINT GENOME INSTITUTE PRODUCTION GENOMICS FACILITY (PGF) FINISHING PIPELINE

H. SUN, A. CLUM, E. GOLTSMAN, S. LOWRY, B. FOSTER, S. TRONG, P. KALE, P. RICHARDSON, A.L. LAPIDUS

DOE JOINT GENOME INSTITUTE, WALNUT CREEK, CA

INTRODUCTION

The DOE Joint Genome Institute is an integrated high-throughput sequencing facility which provides high-quality genomic sequences to the scientific community. Currently there are approximately 300 microbial genomes in the JGI production genomics facility (PGF) pipeline in Walnut Creek and to date, 160 have been completely finished. The PGF microbial finishing group is involved in improving the microbial genomic sequences that come directly out of production. The finishing team is responsible for solving misassemblies, closing gaps, improving low-quality areas in the DNA assemblies and submitting high-quality data to the annotation group which in turn submit the sequence to the public domain. The finishing pipeline at PGF starts with resolving DNA repeats using an automated pipeline developed at JGI/PGF. Low quality and gap regions in the assemblies are resolved by directed reactions mainly using primers designed by applying Autofinishing functionality of Consed and other software developed in-house. Finally, manual checks are performed to ensure a minimum consensus phred quality of 30 or above, error rate is less than 1 base per 50,000bp and 2 times coverage throughout the DNA sequence assemblies. The strategies for solving repeats, misassemblies and difficult to finish regions will be further discussed in detail in this poster.

Finishing Pipeline

Draft QC

Assemblers — Assemble draft sequences with 3 assemblers: PGA, Phrap and Arachne.

Automated Repeat resolution

Autofinish — 2 rounds of primer walk

Finishing — Gap closure, quality improvement

Annotation

Quality

All low-quality areas (<Q30) are reviewed and re-sequenced.
The final error rate must be less than 1 per 50 Kb.
Minimum of 2x coverage everywhere.
All high quality discrepancies are checked.
All repeats are verified by either forward/reverse pairs or PCR.
The ends of final contigs (chromosomes, plasmids) are checked
Less than 5% of the whole genome is covered by 454 only.
The final assembly passes a QC check.

Lab Techniques

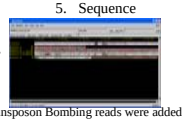
Main Problematic regions:	Approach:
GC rich, AT rich	RCA and sequencing chemistry with 5% DMSO
Hairpins	Sequence Finishing Kit (SFK) Amersham for
Large duplications (>10kb)	high GC and hard stops
Collapsed tandem repeats	PCR using iProof and Failsafe
Hard stops	Qiagen plasmid miniprep sequenced with dGTP
Homopolymer runs	Shatter Library and subcloning
	Transposon Bombing of plasmid templates

Shatter Library Creation

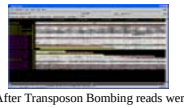
1. Hydrashear
2. Blune End Repair
3. Agarose Gel Separation/Purification
4. Vector Ligation
5. Transformation
6. Plating, Selection, Sequencing

EZ-Tn5 <KAN-2> Transposon Insertion

1. Incubate 2 hrs. at 37 degrees
2. Transformation
3. Select Kan^R clones
4. Prepare plasmid DNA by RCA
5. Sequence



Before Transposon Bombing reads were added



After Transposon Bombing reads were added

Microbial Finishing Software Development

Repeat Finding

Vmatch - a software tool for solving large scale sequence matching tasks.
getBait - Identify repeats by blasting contig sequences against each other and determining the best unique copy.
getAceRepeat – Software program to identify repeats using the 'matchElsewhereHighQual' tags in the ace file generated by PHRAP assembler.

Repeat Resolution

goFishing – Software program to identify and creates fake traces to resolve misassembled regions and optionally reassemble fakes with draft reads to produce an assembly.
goneFishing – Software program to reassemble fake traces created by goFishing or getphrapreads.
getphrapreads – Software program to grab reads from the contigs of an assembly for creating fake traces.
getacereads – Software program to grab reads from the contigs of an assembly for creating fake traces.
misassemblyReporter – Software program to report misassembled regions based on mate-pair information.
tagRepeatsInAceFile - Software program to tag repeat regions in the ace file given a repeat fasta file.

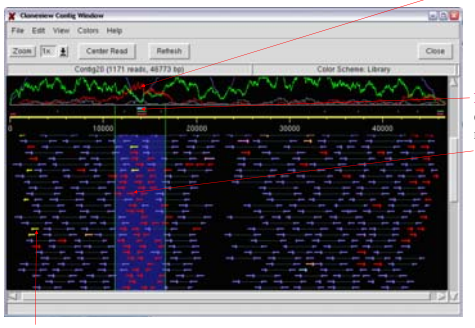
Polishing

tagAceForPolishing - Script to identify target regions for polishing and add those regions as consensus tags to an ace file.
runAutofinishForPolishing - Script to run Consed/Autofinish on polishing regions identified by tagAceForPolishing.
addPrimerTagsToCloneviewFile - script to add primers designed by Autofinish to a cloneview file for displaying in Cloneview.

Visualization Tools

cloneview – Assembly viewer for displaying reads within an assembly. Reads are displayed by contig. Inconsistent read pairs are optionally highlighted in red. Use in conjunction with ace2cloneview.

Cloneview display of misassembled region in Contig20:



Depth of coverage: A spike in inconsistent read coverage (red) followed by a dip in consistent read coverage (green) often indicates a region that is incorrectly assembled. Consistent clone coverage is shown in purple and unpaired read coverage is shown in gray.

Repeat region: Repeat regions are displayed as colored bars on top of the yellow line representing the subject contig.

Read layout: Inconsistent read pairs are highlighted in red color. The current color display is set to coloring reads by library. Alternative color schemes include coloring by paired reads, unpaired reads and library insert sizes. Within the highlighted area (blue background), reads could be marked in order to save the read names to a file for further processing (not shown).

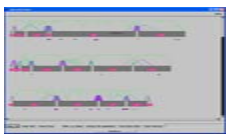
Linking Reads: Reads are highlighted yellow if its pair reside in another contig and the sum of the distance between the ends of the two read pairs to the end of the contig is less than the maximum library insert size. This indicates that the read pair form a link to join the ends of two contigs.

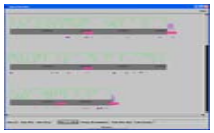
Cloneview Features

Viewing multiple contigs in separate windows.
Displaying the read layout in either an overlap or tiling formation.
Coloring reads by different color schemes (paired, unpaired, library, library insert size).
Highlighting inconsistent read pairs with a red color.
Coloring read pairs that link contig ends yellow.
Marking reads within a user defined area, or singly to save the read names to a separate file.
Displaying graphs of the depth of coverage for consistent reads, inconsistent reads, unpaired reads, consistent clones.
Displaying consensus tags as bars on top of the contig line (e.g., a tag could represent a repeat area within a contig).

Use 454 technology for finishing

The JGI now uses 454 data in every genome it sequences. 454 technology eliminates cloning bias besides its well known advantage in throughput and cost effectiveness. However, confirming concensus sequences with 454 reads only has not been completely successful due to the overall inconsistent and unpredictable error rates. Therefore, currently, only 5% or less of the entire genome is allowed to have 454 data only. How to balance out 454 data and Sanger reads as well as how to use 454 data for polishing microbial genomes are going to be ongoing studies at the JGI.

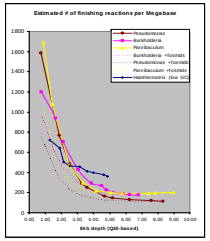


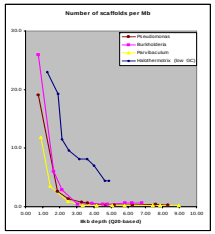


Assembly Views of Thermoanaerobactor ethanolicus X514

Repeat resolution + 2 round of prefinishing **Repeat resolution + 2 round of prefinishing + 454 data**

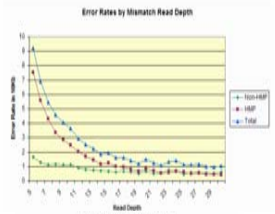
Balancing 454 and Sanger reads



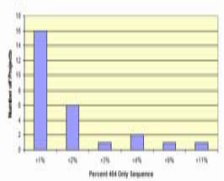


For each criteria, the genomes under study are combined. Datasets with 11x of 454 depth are shown. With both criteria, there is a strong correlation between the high-GC genomes, and the highest saturation point is at 5x. The low-GC Halothermotrix orenii has a significantly larger number of both uncaptured gaps and projected reactions, and the saturation point could be beyond its maximum depth. Fosmid data was excluded in the case of H.orenii.

454 error rate analysis



This shows the error rate in 454 Newbler assembly by 454 mismatch read depth. The error rate is computed by dividing the number of 454 errors at a given mismatch read depth by the number of bases observed at that read depth. The data is obtained from 29 microbial projects.



Our analysis of 29 microbial projects sequenced with 8x-3kb, 8x-8kb, 1x-fosmid and 1-2 runs 454 shows that the majority of projects contained less than 1% of regions covered only by 454 sequences. The graph below shows a distribution of percent 454 only areas that were observed in these projects.

Future Development

454 technology has proven to be a great tool for contigs scaffolding especially for genomes with low GC content. However, the error rates in 454 only regions have not been predictable or consistent. Also, since 454 data does not cover repeat regions, it is essentially important to include fosmid and 8Kb libraries in the datasets. Future development should focus on balancing out 454 and sanger data, using fosmid di-tag libraries to help orient contigs, developing tools to use 454 data for polishing so that the length of polishing process can be further reduced. Another direction is to utilize and implement Solexa technology in genome assembly.

LBNL-62690 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. LBNL-62531 Poster