Preliminary High-Throughput Metagenome Assembly

Hank Tu [1], Craig Furman [2], Jasmyn Pangilinan [1], Harris Shapiro [1], Surge Dusheyko [1]

[1] Lawrence Berkeley National Laboratory/JGI, [2] Lawrence Livermore National Laboratory/JGI

March 28, 2007

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

DISCLAIMER:

[**LBNL**] This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

[**LLNL**] This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

PRELIMINARY HIGH-THROUGHPUT METAGENOME ASSEMBLY

Serge Dusheyko, Craig Furman, Jasmyn Pangilinan, Harris Shapiro, Hank Tu

Abstract

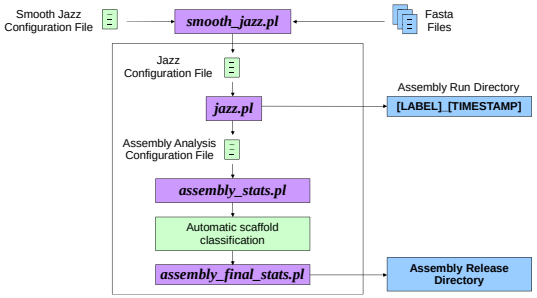
Metagenome data sets present a qualitatively different assembly problem than traditional single-organism whole-genome shotgun (WGS) assembly.

The unique aspects of such projects include the presence of a potentially large number of distinct organisms and their representation in the data set at widely different fractions. In addition, multiple closely related strains could be present, which would be difficult to assemble separately. Failure to take these issues into account can result in poor assemblies that either jumble together different strains or which fail to yield useful results.

The DOE Joint Genome Institute has sequenced a number of metagenomic projects and plans to considerably increase this number in the coming year. As a result, the JGI has a need for high-throughput tools and techniques for handling metagenome projects.

We present the techniques developed to handle metagenome assemblies in a high-throughput environment. This includes a streamlined assembly wrapper, based on the JGI's in-house WGS assembler, Jazz. It also includes the selection of sensible defaults targeted for metagenome data sets, as well as quality control automation for cleaning up the raw results. While analysis is ongoing, we will discuss preliminary assessments of the quality of the assembly results (http://fames.jgi-psf.org).

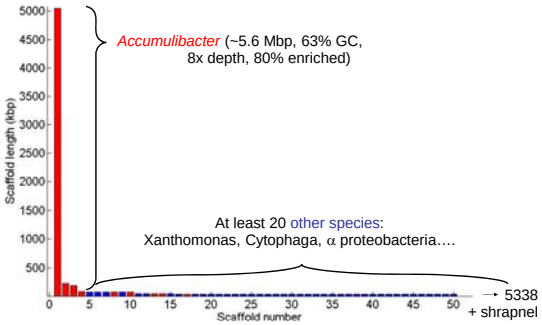
Metagenome Assembly Process



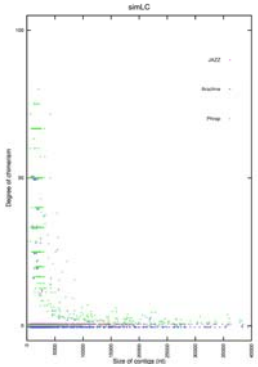
Smooth Jazz Configurable Parameters

Keyword	Required?	Description
plabel	Yes	A string for naming/identifying assembly files.
library	Yes	Entries describing the type & insert size of each library.
project_ID	No	The JGI project number for the data set.
mitochondrion	No	A reference mitochondrion sequence to use for screening.
chloroplast	No	A reference chloroplast sequence to use for screening.
cDNA_file	No	A set of cDNA sequences to use for completeness estimates.
genome_size	No	The estimated genome size for the project. If not specified, set to 5 MB if the metagenomic flag isn't toggled, and 25 MB if it is.
depth	No	The estimated sequence depth for the project. If not specified, set to 8.0 if the metagenomic flag isn't toggled, and 0.0 if it is.
metagenomic	No	Whether or not the data set is a metagenomic project. Toggled off by default.

Metagenome Assembly of US Sludge



Assembly Method Comparison[1]



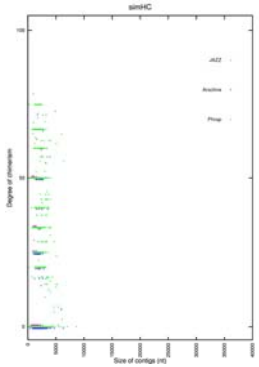
The following graphs show the degree of chimerism for each of the contigs in a simulated data set as assembled by three commonly used programs at JGI.

The *simLC* dataset was designed to simulate low complexity communities. It is dominated by a single population along with several low abundance ones.

Phrap incorporated more reads and produced more homogeneous contigs than its competitors. However, it is extremely aggressive and prone to create chimeric joins as well. JAZZ represents a nice compromise, producing more contigs than Arachne while avoiding the chimeric mistakes of Phrap.

The *simHC* dataset was designed to simulate high complexity communities. It contains over 100 different populations, none of which is dominant.

Due to low depth of coverage and the overall complexity of the problem, all three assemblers produced minimal assembly.



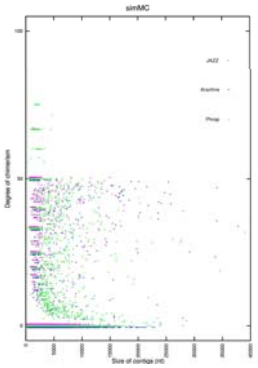
	simLC			simMC			simHC		
	Phrap	Arachne	JAZZ	Phrap	Arachne	JAZZ	Phrap	Arachne	JAZZ
Reads	100628	100628	100546	125652	125652	125513	118064	118064	117890
Contigs(>1 reads)	10320	1400	1953	12644	4892	5840	19236	558	1066
Homogeneous contigs	9055	1333	1941	11026	4329	5282	16451	477	1012
Reads in contigs	56843	33065	29390	82693	55022	39808	47338	2744	2832
Degree of assembly in contigs (%)	56.49%	32.86%	29.23%	65.81%	43.79%	31.71%	40.10%	2.32%	2.40%
Major contigs (≥10 reads)	482	367	503	1980	1372	876	86	20	11
Homogeneous contigs	287	330	502	1380	1133	605	5	18	9
Reads in major contigs	30852	28060	24405	50267	39190	19719	1521	290	192
Degree of assembly in major contigs (%)	30.66%	27.88%	24.28%	40.00%	31.19%	15.71%	1.29%	0.25%	0.16%

Summary of metagenome assembly. The total number of reads given to each assembler is reported on the first line. When this comparison was made, JAZZ was unable to incorporate finished reads that may help prevent chimeric joins. That shortcoming has since been rectified, and JAZZ is now expected to be more competitive with Arachne.

The *simMC* dataset was designed to simulate medium complexity communities. It contains five dominant populations along with several low abundance ones.

Again, Phrap incorporated more reads and produced more homogeneous contigs, but at a heavy chimeric cost. Most *simMC* chimeric contigs were due to the co-assembly of different strains belonging to the same species, and none of the assemblers could effectively discriminate closely-related sequences.

Similar to the *simLC* dataset, JAZZ produced more contigs than Arachne while committing fewer chimeric mistakes than Phrap.



Future Work

Multiple sets of default parameters, each optimized for a different expected level of metagenome complexity.

Improved compatibility and better integration with finishing tools.

Rewrite the JAZZ assembler to improve scalability, so that it can handle the data load from next-gen sequencing machines.

[1]Mavromatis, Konstantinos et al. *On the fidelity of processing metagenomic sequences using simulated datasets.*

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California