BONUS ORGANISMS IN HIGH-THROUGHPUT EUKARYOTIC WHOLE-GENOME SHOTGUN ASSEMBLY

Jasmyn Pangilinan[1], Harris Shapiro[1], Hank Tu[1], Darren Platt[2]

[1]Lawrence Berkeley National Laboratory / DOE Joint Genome Institute
2800 Mitchell Drive, Walnut Creek, CA 94598

[2]Lawrence Livermore National Laboratory / DOE Joint Genome Institute
2800 Mitchell Drive, Walnut Creek, CA 94598

February 6, 2006

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:

Bonus Organisms in High-Throughput Eukaryotic Whole-Genome Shotgun Assembly

Jasmyn Pangilinan, Harris Shapiro, Hank Tu, Darren Platt

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

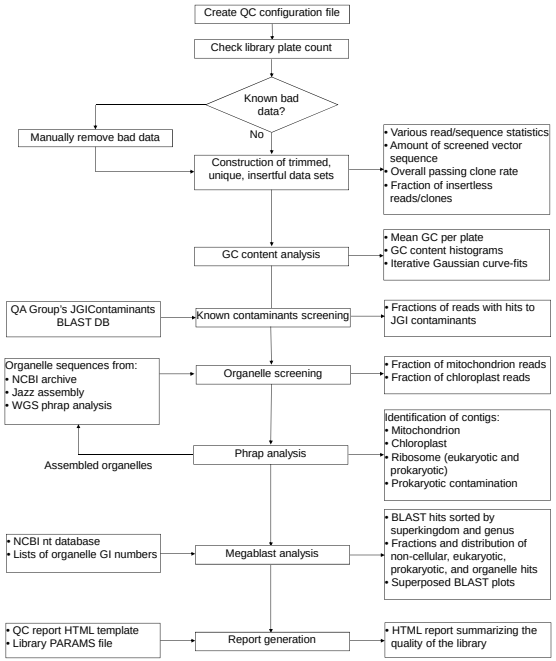
Abstract

The DOE Joint Genome Institute has sequenced over 50 eukaryotic genomes, ranging in size from 15 MB to 1.6 GB, over a wide range of organism types. In the course of doing so, it has become clear that a substantial fraction of these data sets contains "bonus" organisms, usually prokaryotes, in addition to the desired genome. While some of these additional organisms are extraneous contamination, they are sometimes symbionts, and so can be of biological interest.

Therefore, it is desirable to assemble the "bonus" organisms along with the main genome. This transforms the problem into one of metagenomic assembly, which is considerably more challenging than traditional whole-genome shotgun (WGS) assembly. The different organisms will usually be present at different sequence depths, which is difficult to handle in most WGS assemblers. In addition, with multiple distinct genomes present, chimerism can produce cross-organism combinations. Finally, there is no guarantee that only a single "bonus" organism will be present. For example, one JGI project contained at least two different prokaryotic contaminants, plus a 145 KB plasmid of unknown origin.

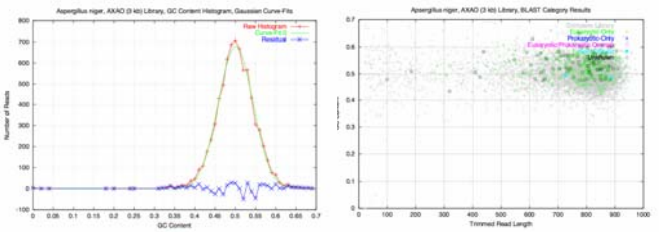
We have developed techniques to routinely identify and handle such "bonus" organisms in a high-throughput sequencing environment. Approaches include screening and partitioning the unassembled data, and iterative subassemblies. These methods are applicable not only to "bonus" organisms, but also to desired components such as organelles. These procedures have the additional benefit of identifying, and allowing for the removal of, cloning artifacts such as *E.coli* and spurious vector inclusions.

Library QC Process

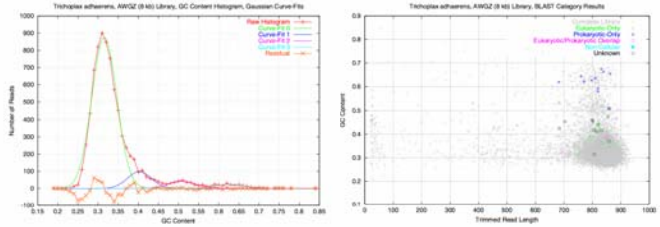


GC Content + Kitchen Sink BLAST Analysis

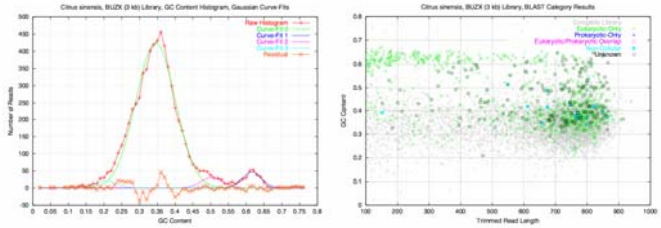
The GC content distribution of a given data set can be fit to a set of Gaussian curves to detect outliers such as prokaryotic or organellar contamination. The fraction of the area under each curve-fit Gaussian can then be used to estimate the amounts of the various types of contamination.



Above is a GC content distribution for 10 384-well plates of an *Aspergillus niger* 3 kb library. Based on this initial QC, the overall data set looks very clean with almost all the reads belonging to the mean GC peak at 0.5. This is confirmed by the kitchen-sink BLAST results, which contain few anomalous hits.



Multiple, or unusually wide, peaks in the GC content distribution can often indicate the presence of a contaminant. As shown above, the GC content distribution for the initial 10 384-well plates of a *Trichoplax adhaerens* 8 kb library contains a suspicious high-GC tail. The kitchen-sink BLAST results suggest the presence at least one high-GC prokaryote.



Sometimes there are false alarms. As shown above, the GC content distribution for the initial 10 384-well plates of a *Citrus sinensis* 3 kb library contains a well-separated high-GC peak. However, superposition of the kitchen-sink BLAST results reveals that both peaks belong to the desired organism.

Organelle sequences can sometimes masquerade as prokaryotic contamination. With experience, the types of hits that lend themselves to misidentification can be quickly excluded. In addition, as organelles are often much smaller than prokaryotes, subassembling the suspect sequences can usually distinguish between the two cases.

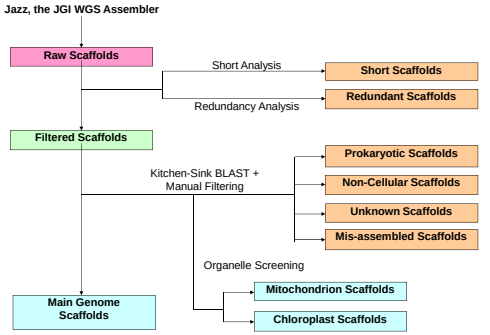
Sometimes contamination creeps in during the library creation process, e.g. from contaminated sequencing reagents. Such cases can often be distinguished by examining the fraction of apparent contamination by library and plate, with true contamination present in all of a project's libraries, and across most (or all) of its plates.

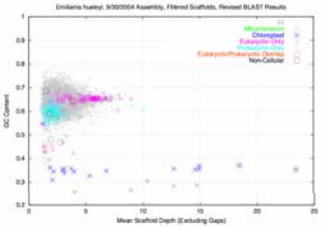
More pathological cases arise due to mislabeled sequence in the NCBI databases themselves. During the course of analyzing several dozen sequencing projects, a number of anomalous sequences have been identified in the NCBI nt database. Labeled as an assortment of eukaryotes, these sequences apparently contain substantial segments of *E.coli*. Cataloguing these anomalies avoids the spurious inclusion of contaminant sequence. Examples include:

•gi|2315165|emb|Y14328.1|EHY14328: Submitted as "Entamoeba histolytica mRNA for 3E1 protein", bases 263-839 align with 98.1% identity to *E.coli*.
•gi|54653446|gb|BT018665.1|: Submitted as "Zea mays clone EL01N0510B04.d mRNA sequence", bases 4-365 align with 98.6% identity to *E.coli*.
•gi|54652349|gb|BT017568.1|: Submitted as "Zea mays clone EL01N0426D07.c mRNA sequence", bases 1-224 align with 100% identity to *E.coli*.

As is clear from the table to the right, a substantial fraction of JGI sequencing projects contains bonus organisms. To investigate whether this is an unusually high rate, we are currently applying our screening techniques to an assortment of projects in the NCBI Trace Archive.

High-Level Assembly QC Process





To the left, the results of the assembly QC process for an early *Emiliania huxleyi* assembly are shown.

Organellar scaffolds are present in distinct GC content bands. This makes it straightforward to extract them from the assembly and, if necessary, use them as the starting points for subassemblies.

Multiple prokaryotic contaminants are also present, at different sequence depths and embedded in the main GC content peak. This makes it difficult to extract all of the contaminant scaffolds. Iterative subassemblies could improve the contiguity of the prokaryotic sequences and allow for better screening, but it is likely that additional sequencing would be required to fully extract them.

Bonus Organisms in JGI WGS Sequencing Projects

Sequencing Project	Bonus Organisms?
Arabidopsis lyrata	None.
Aspergillus niger	High-GC prokaryote (40 kb only).
Aureococcus anophagefferens	Probable high-GC prokaryote. The prokaryotic scaffolds are embedded in the main organism's GC content peak, so confirmation is awaiting the next round of assembly.
Batrachochytrium dendrobatidis	Probable high-GC prokaryote. The prokaryotic scaffolds are embedded in the main organism's GC content peak, so confirmation awaiting the next round of assembly.
Branchiostoma floridae	None.
Capitella sp. I	High-GC prokaryote (40 kb only).
Chlamydomonas reinhardtii	Possible high-GC prokaryote.
Chlorella sp. NC64A	High-GC prokaryote.
Citrus sinensis	Probably none.
Daphnia pulex	High-GC prokaryote.
Emiliania huxleyi	At least 2 high-GC prokaryotes, plus a plasmid of unknown origin.
Glomus intraradices	One or more high-GC prokaryotes, mostly due to contamination of some of the source DNA.
Glycine max Williams 82	Possible high-GC prokaryote. Confirmation awaiting the initial assembly.
Helobdella robusta	Possible high-GC prokaryote. Confirmation is awaiting the final assembly.
Laccaria bicolor	High-GC prokaryote.
Lottia gigantea	None.
Melampsora larici-populina	Probable high-GC prokaryote. Confirmation is awaiting the initial assembly.
Micromonas pusilla CCMP1545	Potential high-GC prokaryote.
Micromonas pusilla NOUM17	High-GC prokaryote.
Mimulus guttatus	None.
Monosiga brevicollis	Low-GC prokaryote.
Mycosphaerella graminicola	None.
Mycosphaerella fijiensis	None.
Naegleria gruberi	None.
Nectria haematococca mpVI	High-GC prokaryote.
Nematostella vectensis	High-GC prokaryote.
Ostreococcus sp. RCC809	Possible low-GC prokaryote. Confirmation is awaiting the initial assembly.
Phaeodactylum tricornutum	None.
Phycomyces blakesleeanus	None.
Physcomitrella patens	1-2 high-GC prokaryotes.
Pichia stipitis	None.
Postia placenta	None.
Reniera sp.	High-GC prokaryote.
Selaginella moellendorfii	None.
Sorghum bicolor	1-2 high-GC prokaryotes.
Sporobolomyces roseus	None.
Trichoderma atroviride	None.
Trichoderma virens	None.
Trichoplax adhaerens	1-2 high-GC prokaryotes.
Volvox carteri	High-GC prokaryote.
Xenopus tropicalis	1-2 high-GC prokaryotes, plus a possible low-GC prokaryote.

Projects with confirmed bonus organisms, where those organisms appear to be intrinsic to the source DNA, are highlighted in green. Projects with non-intrinsic bonus organisms, or where the bonus organism has not yet been confirmed, are highlighted in blue. 17 additional JGI WGS sequencing projects are not listed, either because the necessary analyses have not yet been performed, or because insufficient sequence has been generated.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231