Title
Bonus Organisms in High-Throughput Eukaryotic Whole-Genome Shorgun Assembly

Permalink
https://escholarship.org/uc/item/5nv0q15b

Authors
Pangilinan, Jasmyn
Shapiro, Harris
Tu, Hank
et al.

Publication Date
2006-05-10
BONUS ORGANISMS IN HIGH-THROUGHPUT EUKARYOTIC WHOLE-GENOME SHOTGUN ASSEMBLY

Jasmyn Pangilinan¹, Harris Shapiro¹, Hank Tu¹, Darren Platt²

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

²Lawrence 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:

[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.
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 symbiotic, 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 present at different sequence depths, which is difficult to handle in most WGS assemblers. In addition, with multiple distinct genomes present, chimera 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, at 114 KB and 145 KB plasmids 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 insertions.

Library QC Process

Create QC configuration file

Check library plate count

Known bad plates

Manually remove bad data

Construction of trimmed, unique, insertful data sets

QC report HTML template

Library PARAMS file

As shown above, the GC content distribution for the initial 10 384-well plates of a Citrus sinensis 1.6 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.

Organellar 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 database. We are currently applying our screening techniques to an assembly of projects in the NCBI Trace Archive.