Title
POLISHER: a Tool for Using Ultra Short Read in Micorbial Genome Finishing

Authors
LaButti, Kurt
Foster, Brian
Lowry, Steve
et al.

Publication Date
2008-05-28

POLISHER: a tool for using ultra short reads in microbial genome finishing

Kurt LaButti[1], Brian Foster[1], Steve Lowry[1], Stephan Trong[2], Eugene Goltsman[1], and Alla Lapidus[1]

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

DOE Joint Genome Institute, Walnut Creek, CA 94598

Polishing is one of the major steps of genome finishing at the JGI (Joint Genome Institute). Along with repeat resolution and gap closure, it is required to produce a fully sequenced high quality genome. Polishing consists of consensus error correction and quality improvement such that the resulting consensus meets a pre-defined standard. This has traditionally been done through targeted Sanger clone based sequencing, making it the most time consuming and resource intensive stage in finishing. Our pilot experiments, conducted using Illumina data produced by the JGI for several microbes, demonstrated that aligning ultra short reads against unpolished contigs help correct a significant amount of consensus errors and greatly reduce the amount of quality improvement necessary to produce a finished genome. A prototype tool named the "Polisher" was developed in order to automate this process. It facilitates polishing and error correction of a subject assembly (acefile), typically a draft or closed assembly, using Illumina read data.

The Polisher corrects consensus errors and supports correctly called bases that would normally be targeted for polishing due to their below standard quality by analyzing Illumina alignment data and automatically modifying the consensus. The Illumina read data in Fasta format is aligned to the subject sequence and simultaneously parsed for the best hit based on percent identity. The best hit alignment results are then used to determine coverage and discrepancy information per base in the subject. A list of errors is generated where there is overwhelming evidence that the consensus base is wrong and needs to be changed. Such areas represent mismatches, deletions, and insertion. Mismatches and deletions are automatically corrected in the acefile while insertions, for now, remain as tags for manual inspection. In addition to the previously described error correction, other areas of the genome that would normally be targeted for polishing are inspected. These areas currently represent low quality, single subclone, and 454 only regions and exist as tags in the acefile. If there is overwhelming evidence that a particular base targeted for traditional polishing is correct, then that base is termed Supported and retagged. If there is not enough evidence for support, then the original polishing tag remains for traditional manual polishing.

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.

LLNL-ABS-402508