The Utility of 454 Long Paired-end Technology in De Novo Sequencing

Jan-Fang Cheng[1], Ze Peng[1], Matthew Hamilton[1], Ilya Malinov[1], Hank Tu[1], Edward Kirton[1], Eugene Goltsman[1], Alla Lapidus[1], Alex Copeland[1] and Susan Lucas[2]

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

US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA

We have constructed and sequenced several 454 paired end LT libraries with 3 Kb insert size to assess the quality and effectiveness of the technology in the whole genome assembly. A set of quality metrics, which include levels of redundant reads, linker positive, linker negative, half linker reads, and driver DNA contamination, and read length distribution, were used to measure the primary quality of these libraries. We have also assessed the quality of the resulted mate pairs including levels of chimera, distribution of insert sizes, and genome coverages after the assemblies are completed. In general, the technology is performing as expected with 3 Kb fragments. We have made several modifications in the procedure to improve the quality of the libraries, especially in the areas of reducing redundant reads and eliminating pUC driver DNA contamination. We have also compared two assembly tools for their performance in the assembly of mixed Sanger and 454 shotgun or paired-end data. We will report the quality of these libraries and the comparison of the assemblies with shotgun and paired-end reads. The current GS paired end LT library construction protocol would not work for fragment sizes larger than 3 Kb. The major obstacles are the low recovery of large DNA fragments after multiple purification steps and the lack of an efficient way to circularize large DNA fragments. We are in the process of modifying some steps to overcome these obstacles. We will present results of the improvements.

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.

UCRL-ABS-236476