Gap Resolution: A Software Package for Improving Newbler Genome Assemblies

Stephan Trong, Kurt LaButti, Brian Foster, Cliff Han, Tom Brettin, and Alla Lapidus

DOE Joint Genome Institute, Walnut Creek, CA 94598

With the continued improvements of next generation sequencing technologies and their advantages over traditional Sanger sequencing, the Joint Genome Institute (JGI) has modified its sequencing pipeline to take advantage of the benefits of such technologies. Currently, standard 454 Titanium, paired end 454 Titanium, and Illumina GAII data are generated for all microbial projects and then assembled using the Newbler genome assembler. This allows us to efficiently produce high quality draft assemblies at a much greater throughput than before. However, it also presents us with new challenges. In addition to the increased throughput, we also have to deal with a larger number of gaps in the Newbler assemblies. Gaps in these assemblies are usually caused by repeats (Newbler collapses repeat copies into individual contigs, thus creating gaps), strong secondary structures, and artifacts of the PCR process (specific to 454 paired end libraries). Some gaps in draft assemblies can be resolved merely by adding back the collapsed data from repeats. To expedite gap closure and assembly improvement on large numbers of these assemblies, we developed a software package, called Gap Resolution, to perform the following process automation.

1. Identify and distribute the data for each gap into sub-projects.
2. Assemble the data associated with each sub-project using a secondary assembler, such as PGA.
3. Determine if any gaps are closed after reassembly, and either design fakes (consensus of closed gap) for those that closed or lab experiments for those that require additional data.

This software package was designed specifically to help automate the process of gap closure and assembly improvement in next generation assemblies. Use of this software on microbial genome assemblies has significantly alleviated manual finishing on each project. We are currently testing the software for use on fungal projects.

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.