Title
Gap Resolution: A Software Package for Improving Newbler Genome Assemblies

Authors
Trong, Stephan
LaButti, Kurt
Foster, Brian
et al.

Publication Date
2009-05-27

Gap Resolution: A Software Package for Improving Newbler Genome Assemblies

Stephan Trong[2], Kurt LaButti[1], Brian Foster[1], Cliff Han[3], Tom Brettin[3], and Alla Lapidus[1]

[1]Lawrence Berkeley National Laboratory, [2] Lawrence Livermore National Laboratory, [3] Los Alamos National Laboratory
DOE Joint Genome Institute, Walnut Creek, CA 94598

INTRODUCTION

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.

METHODOLOGY

To expedite gap closure and assembly improvement on large numbers of these assemblies, we developed the following protocol and have written software to automate these steps.

1. Identify and distribute the reads for each captured gap into sub-projects based on read pairing information.

2. Assemble the reads associated with each sub-project using a secondary assembler, such as Newbler or PGA. Validate assembly using anchor sequences.

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.

SOFTWARE DETAILS

• This software package, written in Perl, was designed specifically to help automate the process of gap closure and assembly improvement in next generation assemblies.

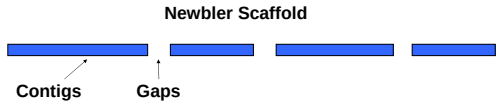
• Designed to be modular, the software currently contains modules to reassemble the gaps using either Newbler or PGA. Primer3 is used design primers for finishing reactions in gaps that could not be closed automatically.

• The software was written to run on Linux. Other operating systems have not been tested yet.

• On a typical microbial genome of size 3.5Mb, gap resolution takes approximate 0.5 hrs to complete.

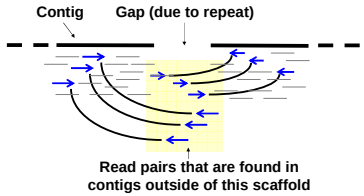
Automating assembly improvement using "Gap Resolution" software:

– Microbial genomes with 454 only sequences are assembled using the Newbler assembler.
– Newbler does not resolve repetitive regions. These regions are collapsed into individual contigs, thus creating gaps in the assembly.
– How can we help automate the process of closing these gaps?

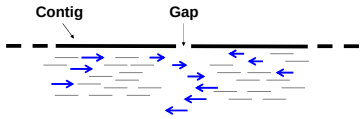


GAP RESOLUTION STRATEGY:

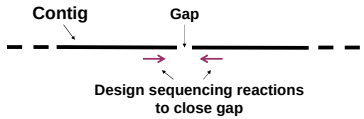
Step 1: For each gap, identify read pairs from contigs found on different scaffolds

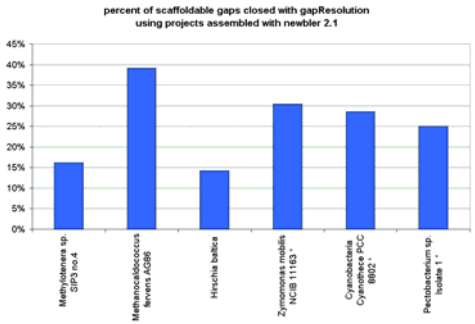


Step 2: Assemble reads in contigs adjacent to the gap and reads obtained from contigs outside the scaffold; validate for correctness using anchor sequences adjacent to the gap



Step 3: If gap is not closed, design sequencing reactions





Step 4: Iterate as necessary

CONCLUSION

• Use of this software on microbial genome assemblies has demonstrated effectiveness in reducing manual finishing on each project. We are currently testing the software for use on fungal projects.

• Preliminary results from an analysis of 6 microbial genomes showed a reduction in the number of gaps ranging from 14% to 39% (31% avg) when using the Newbler assembler for reassembling the gaps. Results were comparable using an alternative assembler such as PGA on microbial genomes, but improved noticeably when applied to fungal genomes, possibly due to differences in repeat structure.

• The software has shown an accuracy of approximately 99% for the gaps that were automatically closed.

LLNL-POST-413451

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.