Title
Gap Closure with Roche/ 454 Reads

Authors
Cheng, Jan-Fang
Tice, Hope N.
Malfatti, Stephanie
et al.

Publication Date
2008-09-08

Gap Closure with Roche/454 Reads

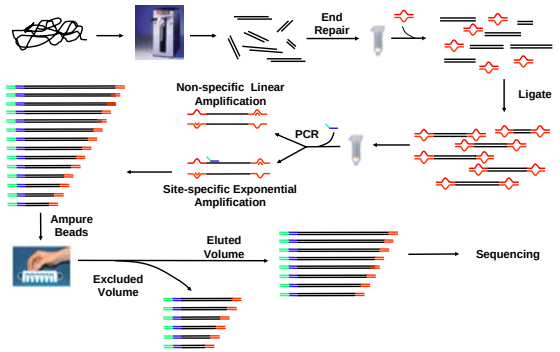
Jan-Fang Cheng, Hope N. Tice, Stephanie Malfatti, Wilson Phung, Matthew Hamilton, Alla Lapidus, and Susan Lucas

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Abstract

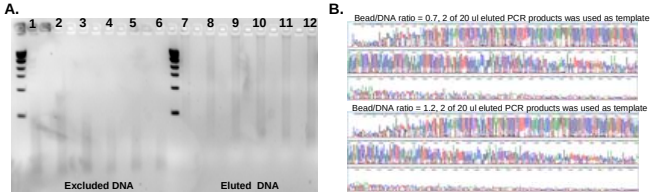
As the Sanger sequencing in the *de novo* assemblies is being replaced by the next generation sequencing reads, finishing of remaining gaps in the genomes will face two major challenges. First, the technology will need to be fast enough to handle many more drafted genomes. Second, it will have to be a clone-free approach. We have been testing a method that utilizes a universal "bubble-tag" to perform extended amplification from contig ends and gap closure in a clone-free condition. The "bubble-tag" method was first described by Doug Smith (PCR Methods Appl. 2, 1992) to sequence lambda DNA. The advantage of this approach is that gap closure can be performed from all contig ends without the prior knowledge of contigs' order and orientation. It has not been demonstrated, however, that this approach would work for the complex genomes. Here we describe the experimentation of this approach in closing gaps of the draft Ktedonobacter racemifer genome. Genomic DNA was sheared and ligated to the bubble adaptors. Primers derived from 96 contig ends were used with the universal bubble primer to amplify the gaps. We applied the AMPure beads to reduce the amount of small fragments. The remaining large amplified DNA appears to be suitable for sequencing. Different bead-to-DNA ratios were tested in order to generate long amplified templates. This approach enables the finishing of complex genomes in a clone-free process with the new sequencing platforms. More importantly, the uniformity of this approach is amenable for a massive, parallel finishing operation.

Bubble PCR Primer Walk Workflow



The DNA of interest is randomly sheared to 3 or 8kb fragments. After end repair the DNA is cleaned up by a column. Bubble adapters (red) are then ligated to the DNA ends. Another cleanup is done and PCR reactions are set up. The reactions contain the bubble primers and the site-specific primers (blue) with M13 tails (green). In this amplification, most of the DNA fragments would have a linear amplification whereas the region of interest would have an exponential amplification. The PCR reaction is then cleaned up using AMPure beads to remove small fragments. The amplification DNA that is eluted from the beads is subjected for sequencing.

The Bubble-PCR Approach Works in the Sanger Sequencing



The bead-to-DNA volume ratios of 0.7 to 2.0 were used to see which one provided the most effective cleaning of small fragments. (A) 20 µl PCR amplified products were cleaned by AMPure beads using the following beads-to-DNA ratios: 0.7 (lanes 1 and 7), 0.8 (lanes 2 and 8), 0.9 (lanes 3 and 9), 1.0 (lanes 4 and 10), 1.1 (lanes 5 and 11), and 1.2 (lanes 6 and 12). The chromatograms generated from sequencing templates of lanes 7 and 12 are shown in (B). The beads-to-DNA ratios in this range for cleaning the small fragments do not seem to significantly affect the quality of the sequencing reads. The Sanger approach has been successfully used in gap closure of several prokaryote genomes at JGI.

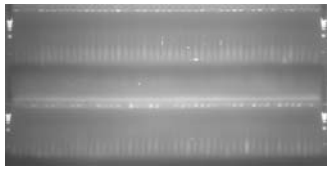
Does it Work with Roche/454 Reads? Test Case - Gap Closure of Ktedonobacter racemifer Genome

Draft assembly of K. racemifer	
Genome Size (est.)	13.7 Mb
GC content	53.80%
8Kb Sanger depth	9.3X
40Kb Sanger depth	0.4X
454 std FLX depth	24.6X
Total phrap contigs	446
Major phrap contigs	393

96 primers designed	
Aligned to genome	350 sites
Multiple hits	46 primers
Unique hits	50 primers
From contig ends	73 primers

A.

B.

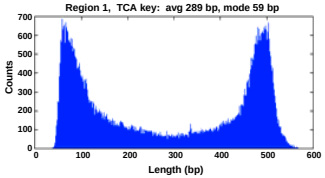


Ninety six primers were selected from the draft genome for the test of primer walk using the bubble-PCR approach. Many of them have multiple matches to the genome, and not all primers were derived form the ends of contigs. Two ug of gDNA was sheared to 8Kb, end-polished, and ligated to a bubble adaptor. About 0.8 ng of the ligated DNA was used in the each of the 96 PCR reactions. After AMPure bead cleaning, 2 ul of the PCR products were loaded on the gel (A), and the excluded DNA were also loaded in gel (B) for comparison. It appears that DNA fragments less than 500 bp were effectively removed by this process. We pooled 24 ul of the eluted product from each amplification and constructed a 454 Titanium library.

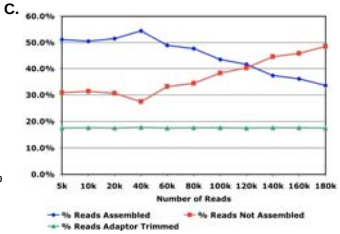
Generation, Assembly, and Gap Filling of 454 reads

A. Load 1/2 of a bulk run → Run Newbler assemblies with randomly picked 5k, 10k, 20k, and up to 180k reads → Align gap filling contigs with the draft assembled contigs by blast and MUMmer

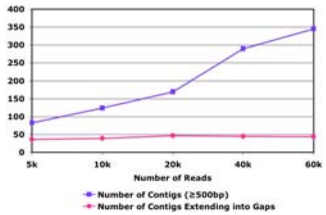
B.



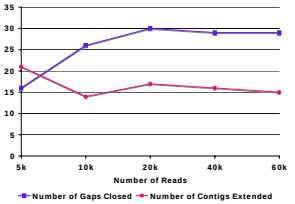
C.



D.



E.



A. A brief overview of the experimental processes.

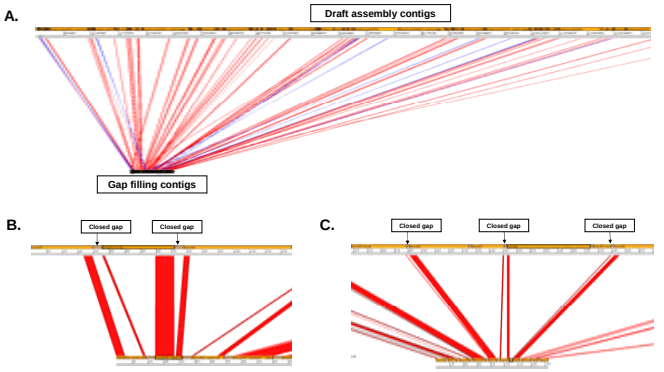
B. About 50% of the reads are shorter than 300bp. Based on the distribution of the sequence depth within the resulted contigs, we believe that the extremely short read length is caused by the nature of this library.

C. About 18% of the reads were trimmed by the bubble-tag sequence. In the cases of low sequence depth (5k to 40k reads), about 30% of the reads were not assembled into contigs. This percentage increases with an increasing amount of sequence coverage. The reads that were not assembled include singletons, repeats, too short, etc.

D. We are particularly interested in the gap closing or extension results generated by different amount of sequence coverage. Although the number of major contigs (≥500bp) increases with a higher sequence depth, the number of contigs that extend into the gaps reaches the highest number at 20k reads.

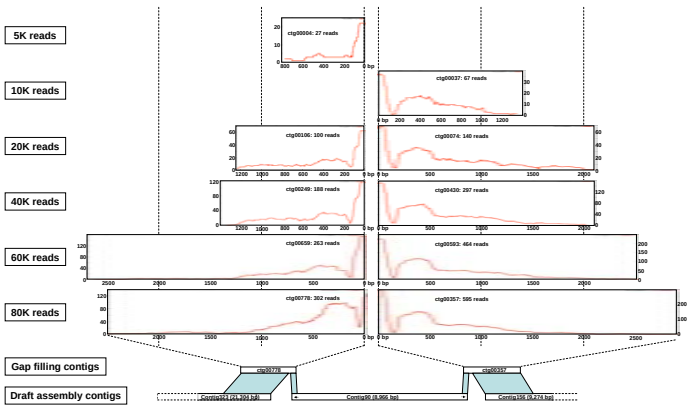
E. At 20k reads, 30 gaps were closed and 17 gaps gain new sequences.

Alignment of the Gap Filling Contigs and the Draft Contigs



Shown here are a distant view (A) and two close-up views (B and C) of the ACT (Artemis Comparison Tool) display of the gap filling contigs and the draft assembly of the *K. racemifer* genome based on the MUMmer alignment. The contigs that extend or bridge the gaps in the draft genome is located at bottom of each graph.

Sequence Depth Distribution in the Gap Filling Contig



This diagram shows that two gaps were closed by walking out of a draft assembled contig using the bubble-PCR approach. Two primers derived from the contig ends (bottom of the diagram) were used to generated templates, which were then used in the 454 sequencing to generate contigs. The contig length increases as the sequence depth increases. The distribution of the 454 reads within the assembled contigs has a similar pattern. The high sequence coverage at the beginning of the contigs accounts for the high percent of the short read length in this library.

Conclusions

1. We have demonstrated that the bubble-tag PCR can be used to generate targeted templates for the finishing process. The sequencing and assembly can be performed by using the 454 platform.
2. This clone-free finishing process can be achieved by pooling a large number of DNA fragments generated from contig ends and poor quality regions. However careful quantification and equal mixing of DNA samples is important to minimize the representation bias.

Acknowledgements

We would like to thank Linda Meincke and Cliff Han of LANL for providing the draft assembly and primer sequences, Eugene Goltsman and Alex Copeland for suggestions of running the assemblies, and Dorothy Lang for running the MUMmer analysis.

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.