Genome Improvement with Bubble-PCR and Roche/454 Reads

[1]Hope N. Tice, [2]Janey Lee, [2]Matt Hamilton, [2]Brian Foster, [1]Stephanie Malfatti, [1]Patrick Chain, [2]Alla Lapidus, [1]Susan Lucas, and [2]Jan-Fang Cheng

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

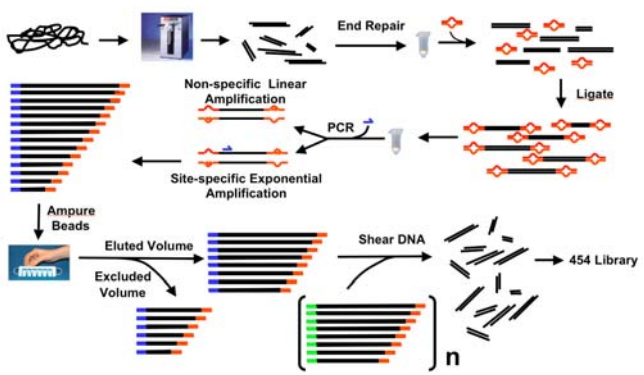
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Abstract

As Sanger sequencing is being replaced by higher throughput and lower cost of the second-generation sequencing, finishing of microbial genomes and eukaryotic genome improvement 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 incorporate a clone-free approach to fill gaps. The method that utilizes a universal "bubble-tag" to perform primer walking and gap closure in a clone-free condition has been implemented at the Joint Genome Institute (JGI) and JGI-LANL using the Sanger based sequencing. The "bubble-tag" method was first described by Doug Smith (PCR Methods Appl. 2: 21-27, 1992) to amplify and sequence lambda DNA. We have started experimenting the bubble-PCR approach using sequences from the Roche/454 platform.

Here we describe the experimentation of this approach in primer walking of the finished microbes, Micrococcus luteus NCTC 2665, and Methanococcus voltae A3. This test is to see if there is a sequencing bias toward high GC or low GC microbes. We also describe this experimentation method on a fungal project Aspergillus carbonarius. Genomic DNA was sheared, blunt-end repaired, or digested with frequent cutters, and ligated to bubble adaptors. Site specific primers were used together with the universal bubble primer to amplify and sequence the regions of interest. Roche/454 libraries were generated from the pooled amplification products. This approach enables primer walking and gap filling in a clone-free draft sequencing process. As described in the past the uniformity of this approach is amenable for an automated finishing/genome improvement process.

Bubble PCR Primer Walk Workflow



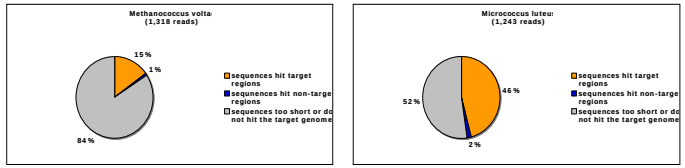
The gDNA of interest is randomly sheared to 3 or 8kb fragments. After end repair, bubble adapters (red) are ligated to the DNA ends. After cleanup, bubble PCR (bPCR) reactions are set up with site-specific primers (blue) and the universal bubble primer (red). Most DNA fragments would have a linear amplification whereas the regions of interest would have an exponential amplification. Small fragments are removed by using AMPure beads and the remaining DNA is used for sequencing. Thousands of bPCR amplified targets can be pooled together for the construction of a single 454 library.

Experimental Data

	Micrococcus luteus NCTC2665	Methanococcus voltae A3	Aspergillus carbonarius (cleaned)	Aspergillus carbonarius (not cleaned)
Genome Size (Mb)	2.6	1.9	34.9	34.9
G+C (%)	72.9	28.6	51	51
Number of Primers Used	20	20	546/530 Unique	576/544 Unique
total reads after trimming	1243	1318	297234	181569
numAlignedReads	531	188	220586	146607
numberSingleton	36	166	12516	9712
numberRepeat	19	22	32970	17581
numberTooShort	676	964	64132	25250
numberOfLargeContigs	13	2	1332	1131
numberOfBases	9420	1135	1308242	1105890
numberOfContigs	19	14	2958	2450
numberOfBases	12147	4934	1877674	1587383

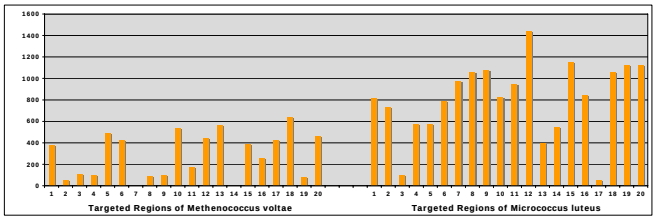
Four barcoded Titanium libraries were constructed, each containing a set of pooled bPCR products generated from the above organisms. Two bacterial genomes were included to test the bPCR in the context of high and low GC contents, and two Aspergillus libraries were constructed to test the impact of removing small fragments after the bPCR. A total of 481,364 reads were survived after the Newbler quality trimming. Some parameters of the Newbler assembly are shown in the Table.

454 Bubble PCR Under High GC vs. Low GC Content



These two genomes were selected in this study due to their extreme GC contents. The GC contents in the Methanococcus voltae and the Micrococcus luteus genomes are 28.6% and 72.9%, respectively. We randomly selected 20 primers that were previously used in the gap closure task to evaluate the performance of the primer walking in the bubble-PCR and 454 sequencing approach. A similar number of trimmed reads were generated from the two bacterial genomes. We have found that a significantly greater portion of reads in the Micrococcus luteus genome are mapped to the targeted regions. There is no evidence to support that the greater number of mapped reads generated in the Micrococcus genome was caused by the difference in GC content.

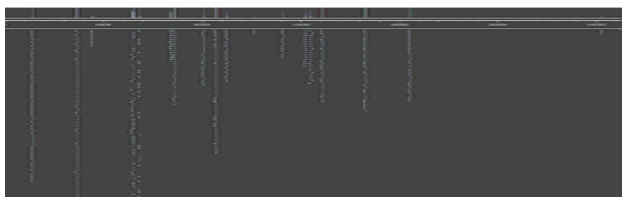
Distribution of the Reads Mapped to the Targeted Regions



We have also tried to determine if there is any bias in the distribution of the mapped reads. The number of bases extended from each of the targeted regions was determined by aligning the reads to the primers. Of the 20 targeted regions in the Methanococcus voltae genome, we saw 18 regions containing sequences extended from the original primers. The average length of the extension is 316 bp. We have detected that 20 out of 20 regions in the Micrococcus luteus genome contain sequences extended from the original primers. The average length of the extension is 809 bp. This difference in length in these two cases may be resulted from the collection of a greater number of reads in the targeted regions of the Micrococcus luteus genome.

Methanococcus voltae

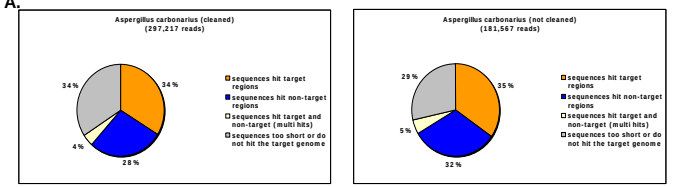
Micrococcus luteus



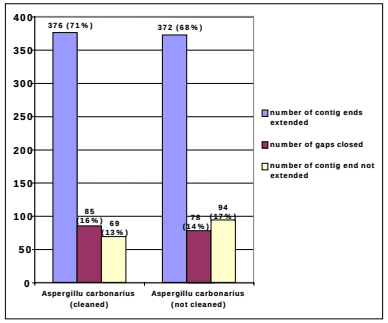
The distribution of the 454 reads that were mapped to the two bacterial genome are shown in the above CloneView panels. Incidentally the 20 primers selected from the Methanococcus voltae genome in this study are all derived from regions of a close proximity. These panels illustrate the specificity of the sequences generated by the bubble-PCR products. We have not seen any significant bias of the mapped reads in these two bacterial genomes.

Aspergillus carbonarius Bubble PCR 454 Sequencing Gap Closure

A.



B.

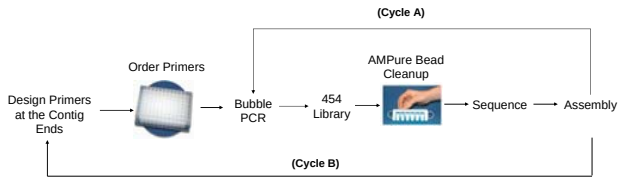


The feasibility of using the bPCR approach to perform primer walking in a fungal genome was tested here. In this study, we designed 1,122 primers from contig ends of the draft Aspergillus carbonarius genome. Of them, 1,074 are unique and 48 are derived from repeats.

We split the bPCR products into two groups to test the effects of a AMPure bead cleaning step before the 454 library construction. In both cases, the percents of the trimmed reads that were mapped to the targeted regions are very similar (panel A) suggesting that the benefit of the AMPure cleaning before library construction is very low. This observation is also supported by the percents of gaps that were either closed or extended in both cases (panel B).

In summary, we have closed a total of 155 gaps and extended 748 contig ends in one run of bPCR and 454 sequencing.

Proposed Gap Closure & Primer Walking Process Using Bubble PCR and 454 Sequencing



The gap closure and primer walking can become a scalable process if integrated with some automated processes. For examples, primers can be ordered at a known concentration and PCR performed by adding the universal bubble primer and adaptor-ligated DNA template. Bead cleanup can also be done using robots. After assembly the missing targets can be identified and another library can be created using the missing bubble PCR procdcts (Cycle A), or new primers can be designed for additional runs of primer walks.

Conclusions

1. We have demonstrated the feasibility of using the bubble-PCR approach and the 454 sequencing to improve the draft assemblies of bacterial and fungal genomes
2. There is room for improvement in this approach. We would like to see a higher percent of reads generated from the targeted regions. Another area of improvement would be the length of sequences extended into the gaps
3. The high and low GC contents do not seem to affect the distribution of the 454 finishing reads
4. This approach enables a clone-free finishing process that is amenable for automation and the second generation sequencing platforms

Acknowledgements

We would like to thank Alicia Clum for providing the primers and primer information, David Robinson for helping with the analysis of spectrophotometer data for pooling, and Diane Bauer for providing training on the spectrophotometer.

LLNL-POST-413440

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