Multiple Long Mate Pair Approaches to Facilitate Short Read based *de novo* Genome Assembly

Julianna Chow*[1], **Jaya Rajamani***[1], James Han[2], Alicia Clum[1], Alex Copeland[1], Shweta Deshpande[1] and Chia-lin Wei[1]

1. DOE Joint Genome Institute / Lawrence Berkeley National Laboratory
2. DOE Joint Genome Institute / Lawrence Livermore National Laboratory

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

**To whom correspondence may be addressed. E-mail: JChow@lbl.gov or JRajamani@lbl.gov*

March 30, 2012

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

DISCLAIMER:

LLNL-ABS-546451

Multiple Long Mate Pair Approaches to Facilitate Short read based *de novo* Genome Assembly

Julianna Chow[1], Jaya Rajamani[1], James Han[2], Alicia Clum[1], Alex Copeland[1], Shweta Deshpande[1] and Chia-Lin Wei[1]

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

Abstract

Short reads based *de-novo* assembly is challenging for complex genomes with variable homologous regions and considerable repetitive elements. Long-Mate Paired (LMP) libraries of different jumping sizes offer potential solutions to bring short reads of different gap sizes into contigs and improve the intactness of the draft genome assembly. Furthermore, LMP sequences can facilitate the ordering of contigs into scaffolds and detect structural variations like indel and translocations. Here, we report the developments of two complementary LMP library construction methods: Cre-Lox Inverse PCR-Illumina Paired-End (CLIP-PE) and Ligation Free Paired-End (LFPE). In CLIP-PE, libraries are created by ligation of biotin-LoxP adaptors to the ends of fixed sizes gDNA and circularized by Cre recombinase mediated intra-molecule recombination. The circularized product is fragmented by a selection of 4-base pair cutting enzyme (*NlaIII, MseI or HpyCH4IV*). The ends of fragmented DNA are self-ligated, biotin-streptavidin selected and enriched by inverse PCR. Mate-Pairs generated by CLIP-PE libraries are identified by the 4-base pair enzyme cutting site. LFPE libraries are created by ligation of internal adaptors lacking 5' phosphate to the ends of gDNA fragments and circularization by hybridization. The circularized product is nick translated, digested by T7 Exonuclease/ S1 Nuclease into short paired tags of fix span sizes, biotin-streptavidin selected and finally, enriched by PCR. Mate-Pairs generated from LFPE libraries are identified by the internal linker junction site. Two organisms *Mycosphaerella fijiensis CIRAD86* and *Phycomyces blakesleeanus NRRL1555* were selected to test the effectiveness of these two LMP library creation methods. Sequencing data generated from the two methods were analyzed to evaluate the resulted assembled genomes for their qualities, coverage, redundancy, bias and accuracy. We conclude that both methods have their unique advantages and disadvantages for various genome assembly applications.

CLIP-PE and LFPE Library Construction Approaches

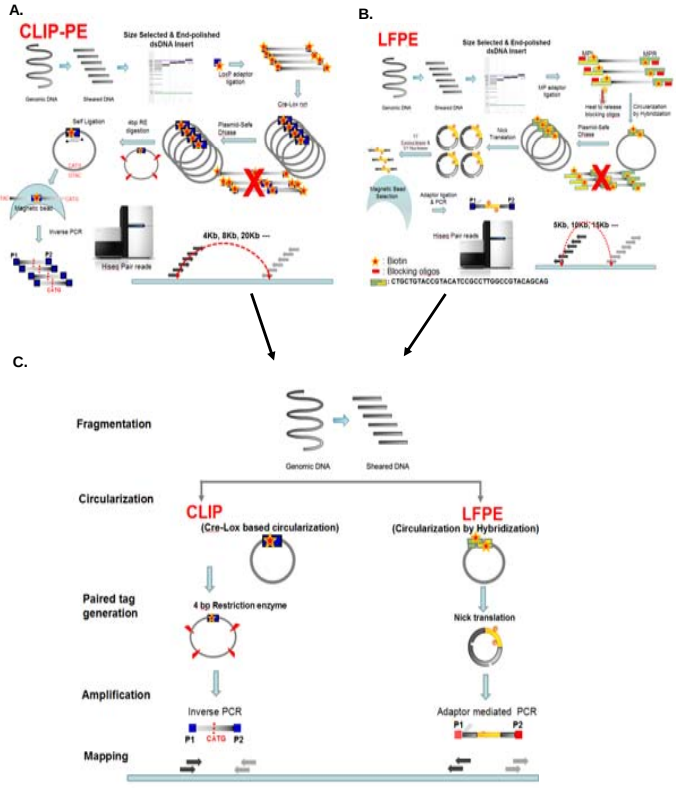

Figure 1. A schematic representation of CLIP-PE and LFPE library construction process (A and B). CLIP-PE is a Cre-LoxP recombination system based circularization, using a 4-bp restriction enzyme followed by inverse PCR to generate long paired-end tags. Circularization of LFPE is based on a ligation free hybridization of internal adpators. Long pair-end tags are created by Nick Translation, T7-Exonuclease/S1 Nuclease following circularization (C). The organisms Phycomyces blakesleeanus NRRL1555 and Mycosphaerella fijiensis CIRAD86 was carefully selected for this experiment to compare the CLIP-PE and LFPE protocols.

Insert Size Distribution

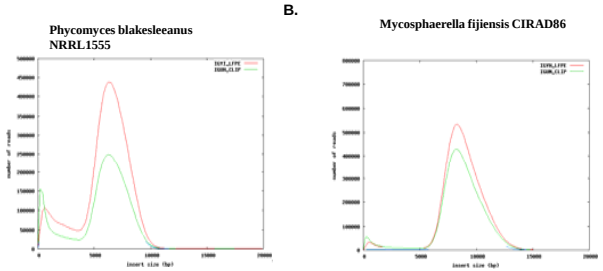

Figure 2. Histogram of insert sizes from Phycomyces blakesleeanus NRRL1555 and Mycosphaerella fijiensis CIRAD86 6-8 kb libraries. The distribution of insert lengths were determined by aligning the reads to the reference genome using the BWA aligner. The genomic DNA were sheared by hydroshear and ran on agarose gel. The samples were size selected between 6-8kb and divided evenly for CLIP-PE and LFPE library creation. Therefore, the size distribution for the two process is expected to be identical.

Coverage Distribution

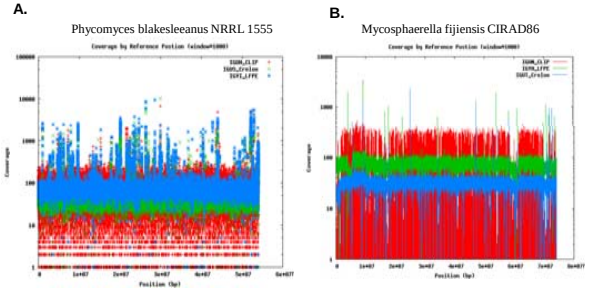

Figure 3. The above chart describes the total coverage when each library were mapped back to the reference. According to the above chart, sequencing data generated from LFPE library creation process yield a tighter coverage, were as CLIP-PE has a broad range. The third library, Cre-lox based, is not a part of the experiment and is not covered in the poster.

Redundancy

Organism	Library	Type	Size (kb)	Reads (M)	Net Reads (M)	Redundant (%)
Phycomyces blakesleeanus NRRL1555	IGUN	CLIP	8	58.5	33.6	62.9
Phycomyces blakesleeanus NRRL1555	IGYI	LFPE	8	78.9	59.7	30.3
Mycosphaerella fijensis CIRAD86	IGUN	CLIP	8	66.4	40.6	55.7
Mycosphaerella fijensis CIRAD86	IGYH	LFPE	8	74.7	63.2	15.7

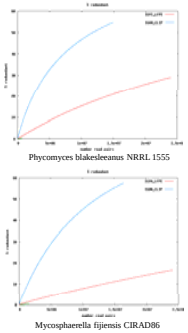

Figure 4. The libraries created by CLIP-PE and LFPE process for Phycomyces blakesleeanus NRRL1555 and Mycosphaerella fijiensis CIRAD86 were pooled based on organism and sequenced 2 x150 run on Illumina HiSeq platform, sharing one lane for each organism. The idea is to generate equal amount of reads for a fair comparison of redundancy between the two process. Non-redundant pairs are those that have unambiguous mapping coordinates. According to the data- chart and graphs, LFPE library creation is significantly less redundant for both organisms.

Assembly Metrics

	Library Type	Contigs	Scaffolds	CL50	BL50	
P. blakesleeanus NRRL1555	--	371	81	368	1500	Nuclear Genome Assembly v2.0
	CLIP-PE	2920	690	42	339	
	LFPE	2804	479	50	526	
M. Fijiensis CIRAD86	--	778	56	211	5900	Nuclear Genome Assembly v2.0
	CLIP-PE	4944	3817	27	64	
	LFPE	4390	538	50	443	

Figure 5. Nuclear Genome Assembly v2 is an improved assembly produced by the JGI Finishing Pipeline. Assembly v2 reports P. blakesleeanus NRRL1555 has 81 scaffolds and M. fijiensis CIRAD86 has 56 scaffolds.The N50 contig and scaffold score using only CLIP-PE library for assembly is much lower than LFPE library. Assembly using LFPE reduce scaffold size and slightly improve the number of contigs.

Quality of Reads

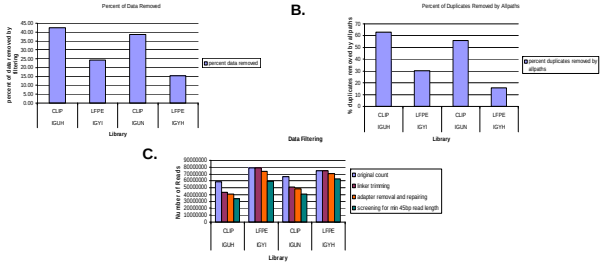

Figure 6. The above graphs detail of how much data is left after each filtering step: total percent data removed from filtering (A and B) and percent duplicates removed by AllPaths (C). The minimum read length distributions is determined after all filtering steps, which includes middle linker trimming, filtering out artifacts and removing singlet reads, and screening out pairs with one read is less than 45 bp (C).

Annotation Results

A.

Phycomyces blakesleeanus NRRL1555

IGUH - CLIP-PE
run.std_50.jmp_50
number of genes - 16537
number of genes with deletion defects - 3722
number of completely missed genes - 2363
number of genes with deletions less than 10bp - 54
number of genes with missing bases in introns, but not coding - 221
number of affected genes with 0gap - 0
number of genes with insertions - 70
Total Fraction of Affected Genes - 22.93 %

IGYI - LFPE
run.std_50.jmp_50
number of genes - 16537
number of genes with deletion defects - 3469
number of completely missed genes - 2254
number of genes with deletions less than 10bp - 54
number of genes with missing bases in introns, but not coding - 201
number of affected genes with 0gap - 0
number of genes with insertions - 93
Total Fraction of Affected Genes - 21.54 %

B.

Mycosphaerella fijiensis CIRAD86

IGUN CLIP-PE
run.std_50.jmp_50
number of genes - 13107
number of genes with deletion defects - 585
number of completely missed genes - 187
number of genes with deletions less than 10bp - 117
number of genes with missing bases in introns, but not coding - 162
number of affected genes with 0gap - 0
number of genes with insertions - 97
Total Fraction of Affected Genes - 5.20 %

IGYH LFPE
run.std_50.jmp_50
number of genes - 13107
number of genes with deletion defects - 328
number of completely missed genes - 95
number of genes with deletions less than 10bp - 79
number of genes with missing bases in introns, but not coding - 161
number of affected genes with 0gap - 0
number of genes with insertions - 88
Total Fraction of Affected Genes - 3.17 %

Figure 7. The annotation results provide a detailed comparison of the quality of each library.

Conclusions

CLIP-PE and LFPE 8kb libraries were created in parallel; however, data show LFPE libraries have better: coverage/complexity distribution, uniqueness, and assembly compared with CLIP-PE libraries. However, both library creation processes have limitations. CLIP-PE is limited as it relies on 4-bp restriction site which can introduce coverage bias in genomes with extremely high or low GC. Furthermore, there maybe concerns of potential gaps in genome coverage if the restriction enzyme site is unevenly distributed throughout the genome. LFPE is hybridization mediated circularization followed by nick translation which makes the process sensitive to time and temperature. Circularization can be less efficient as the target insert size increase to above 10kb due to the nature of self-hybridization of internal adaptors. Additionally, DNA quality of starting material is far most important. Samples with poor quality DNA may not produce a high quality library as a result of nicks within the genome inhibiting nick translation reaction. CLIP-PE on the other hand can be modified to replace enzyme restriction with random shearing (process is under testing) to eliminate coverage biases. LFPE process produces better quality sequencing data, however process is not scalable and not robust as compared to CLIP-PE libraries.

Acknowledgements

We would like to thank Alicia Clum and Alex Copeland for the data analysis – charts and graphs, Jaya Rajamani for creating the CLIP-PE libraries.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231