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

Permalink
https://escholarship.org/uc/item/34f6p03s

Author
Chow, Julianna

Publication Date
2014-04-21
Multiple Long Mate Pair Approaches to Facilitate Short Read based de novo Genome Assembly

Julianna Chow*1, Jaya Rajamani*1, James Han2, Alicia Clum1, Alex Copeland1, Shweta Deshpande1 and Chia-Lin Wei1

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:

[LBNL] This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of 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.

[LLNL] This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of its employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

LLNL-ABS-546451
Multiple Long Mate Pair Approaches to Facilitate Short read based de novo Genome Assembly

Julianna Chow¹, Jaya Rajamani¹, James Han², Alicia Clum¹, Alex Copeland¹, Shweta Deshpande¹ and Cha-Lin Wei¹
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 integrity 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 (NtBIl, Mool or HpyC41F48). The ends of fragmented DNA are self-ligated, biotin-strepavidin 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 ±500 sizes, biotin-strepavidin selected and finally, enriched by PCR. Mate-Pairs generated from LFPE libraries are identified by the internal linker junction sites. Two organisms Phycomyces blakesleeanus NRRL1555 and Mycosphaerella fijiensis CIRAD86 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.

Assembly Metrics

The above chart compares the total coverage when each library were mapped back 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 of the two process is expected to be identical.

Conclusions

CLIP-PE and LFPE 86b libraries were created in parallel; however, data show LFPE libraries have better: coverage/completeness 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-base restriction sites which can introduce coverage bias in genomes with extremely high or low GC. Furthermore, there may be concerns of potential gaps in genome coverage if the restriction enzyme site is unwiely distributed across the genome. LFPE, on the other hand, introduces circularization followed by nick translation which makes the process sensitive to time and temperature. Circulation can be less efficient as the target insert size increases to above 10kb. Additionally, CLIP-PE is more sensitive to time and temperature. Therefore, DNA quality of starting material is far more important. Samples with poor quality DNA may not produce a high quality library as a result of nick 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 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.