Title
Large Gap Size Paired-end Library Construction for Second Generation Sequencing

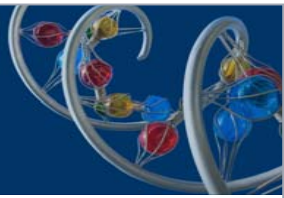
Author
Peng, Ze

Publication Date
2010-06-02

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Large Gap Size Paired-end Library Construction for Second Generation Sequencing

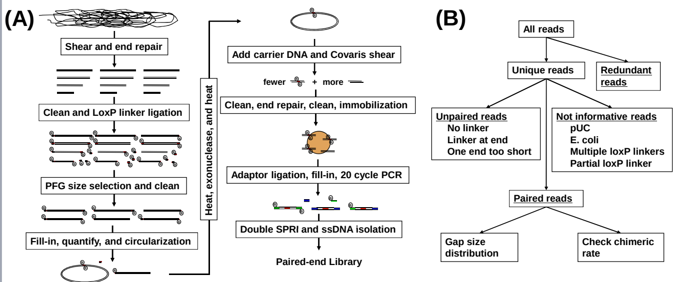
Ze Peng[1], Matthew Hamilton[1], Jeff Froula[1], Aren Ewing[1], Brian Foster[1], and Jan-Fang Cheng[1]
1Lawrence Berkeley National Laboratory



Abstract

Fosmid or BAC end sequencing plays an important role in de novo assembly of large genomes like fungi and plants. However construction and Sanger sequencing of fosmid or BAC libraries are laborious and costly. The current 454 Paired-End (PE) Library and Illumina Jumping Library construction protocols are limited with the gap sizes of approximately 20 kb and 8 kb, respectively. In the attempt to understand the limitations of constructing PE libraries with greater than 30Kb gaps, we have purified 18, 28, 45, and 65Kb sheared DNA fragments from yeast and circularized the ends using the Cre-loxP approach described in the 454 PE Library protocol. With the increasing fragment sizes, we found a general trend of decreasing library quality in several areas. First, redundant reads and reads containing multiple loxP linkers increase when the average fragment size increases. Second, the contamination of short distance pairs (<10Kb) increases as the fragment size increases. Third, chimeric rate increases with the increasing fragment sizes. We have modified several steps to improve the quality of the long span PE libraries. The modification includes (1) the use of special PFGE program to reduce small fragment contamination; (2) the increase of DNA samples in the circularization step and prior to the PCR to reduce redundant reads; and (3) the decrease of fragment size in the double SPRI size selection to get a higher frequency of LoxP linker containing reads. With these modifications we have generated large gap size PE libraries with a much better quality.

Paired-end Library Construction and Data Analysis Processes



(A)

Shear and end repair

Clean and LoxP linker ligation

PFG size selection and clean

Fill-in, quantify, and circularization

Heat, exonuclease, and heat

Add carrier DNA and Covaris shear

fewer + more

Clean, end repair, clean, immobilization

Adaptor ligation, fill-in, 20 cycle PCR

Double SPRI and ssDNA isolation

Paired-end Library

(B)

All reads

Unique reads

Redundant reads

Unpaired reads
- No linker
- Linker at end
- One end too short

Not informative reads
- pUC
- E. coli
- Multiple loxP linkers
- Partial loxP linker

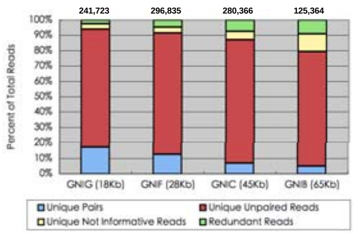
Paired reads

Gap size distribution

Check chimeric rate

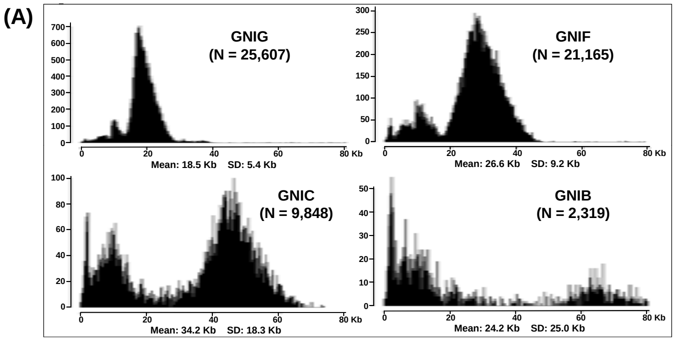
(A) This is a modified version of the 454 Recombi Paired-end Library Construction Protocol. We use pulsed-field gels to size select fragments greater than 20Kb in size. We increase the amount of DNA in the circularization step to 600ng. The pUC carrier DNA is treated with UV to reduce the chance of amplification. We also use Covaris sonicator to shear circularized DNA. The flow of sequence data analysis is shown in (B). All reads are cross-matched against each other to identify redundant reads (greater than 95% nucleotide matches). The remaining reads are grouped into 3 major categories including “not informative”, “unpaired”, and “paired” reads. The paired reads must have more than 15 bases of sequences on both sides of the loxP linker. Only unique paired reads are used to check for chimera and gap size distribution.

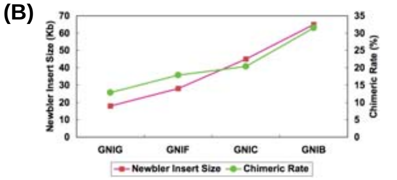
Quality of Long Gap Size PE Libraries



We constructed 4 libraries with DNA isolated from Saccharomyces cerevisiae S288C. We used this completed genome of 12,156,676 bases to evaluate the limitations of the current approach for constructing long gap size paired-end libraries. The fragments isolated from the PFG range from 18 to 65Kb. The names and fragment sizes of these libraries are shown in the left bar graph. The number of reads generated from these libraries are shown on the top of the graph. The large amount of unpaired reads seen in all 4 libraries were caused by the short sequence read length (avg. 300bp) and the long library inserts (avg. 600bp).

Gap Size Distribution and Chimeric Rate

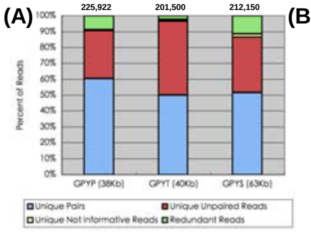


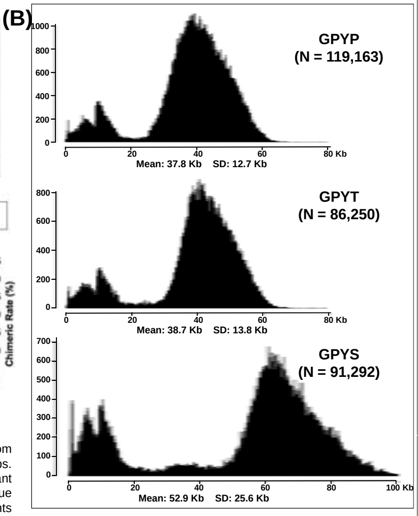


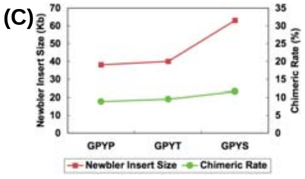
We mapped all the unique pairs of these libraries to the yeast genome and plotted the gap size distribution as 0.5Kb bins in (A). We’ve seen an increasing amount of small fragment contamination as the targeted fragment size increases. Chimeric pairs include two reads with wrong orientation or separated greater than 80Kb. We’ve seen an increase of chimeric rate as the targeted fragment size increases.

Improvement of the Long Gap Size PE Library Construction

Steps to improve	Old process	New Process	Effect
Pulsed-field gel size selection	Once	Twice or two discontinuous pulse cycles	Reduce small fragments
DNA concentration in circularization	6 ng/ul	3 ng/ul	Reduce chimeric rate
Sonication shearing	500-700 bp	200-400 bp	Increase reads with loxP linkers

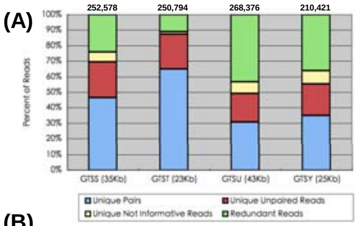




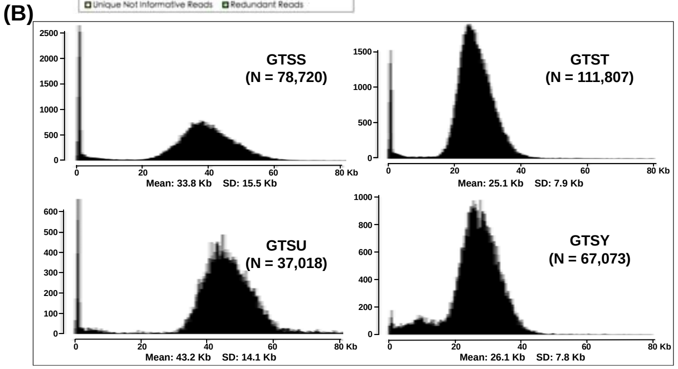


We constructed 3 long gap size libraries from the same yeast DNA using the improved steps. The sequence analysis shows significant improvements in the representation of unique pairs (A), the contamination of small fragments (B), and the reduction of chimeric pairs (C).

Long Gap Size PE Library Construction of Fungal Genomes

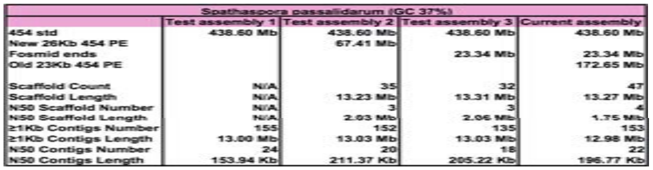


We constructed 4 long gap size paired-end libraries from 3 fungal DNA - Mucor circinelloides (GTSS and GTST), Neurospora tetrasperma (GTSU), and Spathaspora passalidarum (GTSY). The number of reads generated from these libraries are shown on the top of the bar graph (A). The increase of redundant reads in GTSS, GTSU, and GTSY were resulted from a higher loss of DNA during the library construction process. For each of these 3 genomes, we only have a draft assembly to locate the resulted pairs. The gap size distribution plots are shown in (B). We could not accurately determine chimeric pairs without a finished genome.



Test Assembly of Two Fungal Genomes with Long Gap Size Libraries

A



Spathaspora passalidarum (GC 37%)				
	Test assembly 1	Test assembly 2	Test assembly 3	Current assembly
454 std	438.60 Mb	438.60 Mb	438.60 Mb	438.60 Mb
New 26Kb 454 PE		67.41 Mb		
Fosmid ends			23.34 Mb	23.34 Mb
Old 23Kb 454 PE				172.65 Mb
Scaffold Count	N/A	35	32	47
Scaffold Length	N/A	13.23 Mb	13.31 Mb	13.27 Mb
N50 Scaffold Number	N/A	3	3	4
N50 Scaffold Length	N/A	2.03 Mb	2.06 Mb	1.75 Mb
≥1Kb Contigs Number	155	152	135	153
≥1Kb Contigs Length	13.00 Mb	13.03 Mb	13.03 Mb	12.98 Mb
N50 Contigs Number	24	20	18	22
N50 Contigs Length	153.94 Kb	211.37 Kb	205.22 Kb	196.77 Kb

B

Mucor circinelloides CBS277.49							
	Test assembly 1	Test assembly 2	Test assembly 3	Test assembly 4	Test assembly 5	Test assembly 6	Test assembly 7
454 std	2,255 Mb	2,255 Mb	2,255 Mb	2,255 Mb			2,255 Mb
New 23 Kb 454 PE		81 Mb				81 Mb	81 Mb
New 35 Kb 454 PE			80 Mb				
Fosmid ends				39 Mb	39 Mb		
pMCL ends					229 Mb	229 Mb	
pUC ends					156 Mb	156 Mb	
Old 5kb 454 PE							677 Mb
Old 8kb 454 PE							868 Mb
Scaffold Count	N/A	766	1075	850	501	602	565
Scaffold Length	N/A	38.01 Mb	36.06 Mb	42.34 Mb	38.99 Mb	38.16 Mb	37.62 Mb
N50 Scaffold Number	N/A	6	111	8	5	5	5
N50 Scaffold Length	N/A	2.50 Mb	0.97 Mb	1.41 Mb	2.78 Mb	3.15 Mb	3.35 Mb
≥1kb Contigs Number	2,507	2,451	2,506	2,447	2	2,007	2,296
≥1kb Contigs Length	34.92 Mb	34.88 Mb	34.87 Mb	34.98 Mb	35.75 Mb	35.63 Mb	35.02 Mb
N50 Contigs Number	464	440	457	455	329	340	415
N50 Contigs Length	27 Kb	28 Kb	27 Kb	27 Kb	37 Kb	34 Kb	30 Kb

We ran two sets of test assemblies of the Spathaspora passalidarum (Table A) and Mucor circinelloides (Table B) sequences using Newbler. The top halves of the tables show the type of libraries and amount of sequences used in the assemblies. The bottom halves of the tables show the assembly stats. The results show that 454 large insert paired-ends and the Sanger fosmid ends generate similar number of scaffolds and scaffold sizes in the whole genome assemblies (test assemblies 2 and 3 of Table A, and test assemblies 2 and 4, and test assemblies 5 and 6 of Table B).

Acknowledgements

We would like to thank Roche/454 Life Science for providing early access to the Titanium Recombi Paired-end Library construction reagents and protocol.

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.