

# Limitations and Improvement of Constructing Long Paired-end Libraries

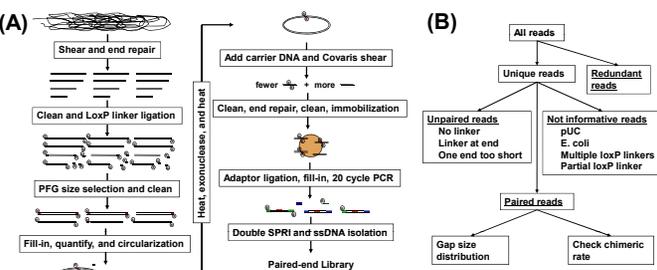
Jan-Fang Cheng, Jeff Froula, Aren Ewing, and Ze Peng



## Abstract

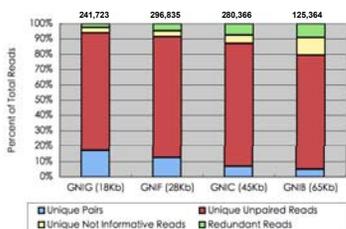
Long insert size paired-end (PE) sequences play a key role in *de novo* assemblies of complex genomes. In the attempt to understand the limitations of constructing PE libraries with greater than 30Kb gaps, we have purified 18, 28, 45, and 65Kb sheared DNA fragments from yeast and circularized the ends using the Cre-loxP approach described in the 454 Titanium Long Paired-end Library protocol. With the increasing fragment sizes, we found a general trend of decreasing library quality in several areas. First, the portions of 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. Our explanation for these quality changes is that as the fragment size increases, the efficiency of Cre recombinase joining two ends of the large fragments decreases. As a result of the PCR amplification, the redundancy of the library increases. On the other hand, the end joining of contaminated short fragments are much more efficient than the targeted long fragments. Therefore the short pairs becomes more prominent as the fragment size increases. Third, we have also observed that the chimeric rate increases with the increasing fragment sizes. It is conceivable that as the size of DNA fragments increases, the chance of joining ends of the same fragments decreases and the chance of joining ends between molecules increases, hence the increased chimeric rate. We have been testing several changes in the protocol. Here we present the progress of these changes.

## Paired-end Library Construction and Data Analysis Processes



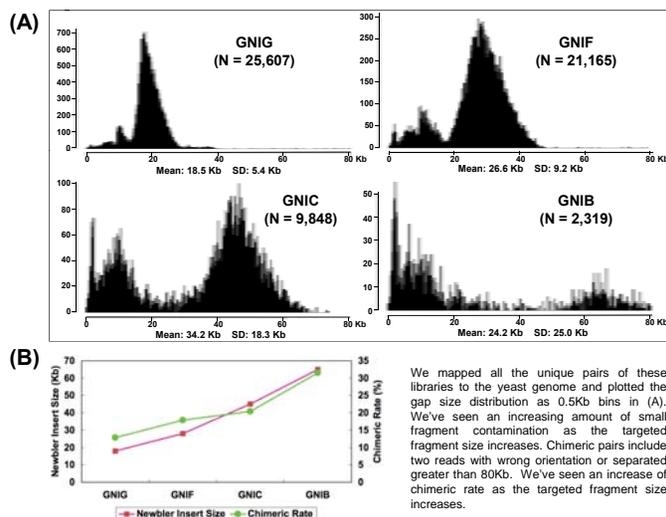
(A) This is a modified version of the 454 Recombi Paired-end Library Construction Protocol. We use pulsed-field gels to size select fragments greater than 20Kb in size. We increase the amount of DNA in the circularization step to 600ng. The pUC carrier DNA is treated with UV to reduce the chance of amplification. 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).

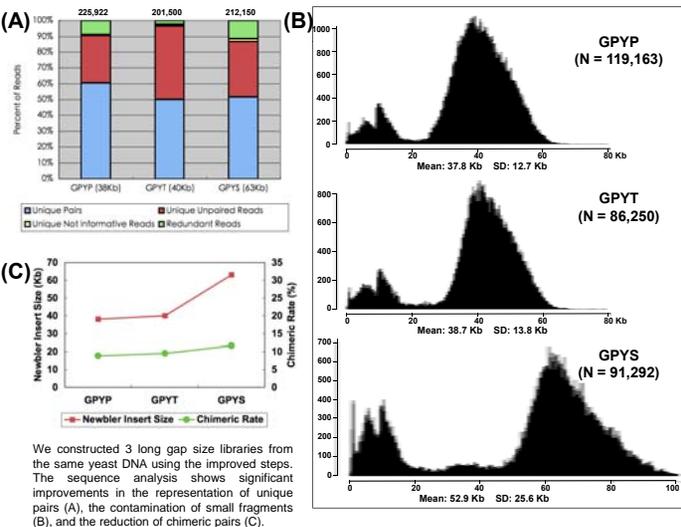
## Gap Size Distribution and Chimeric Rate



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.

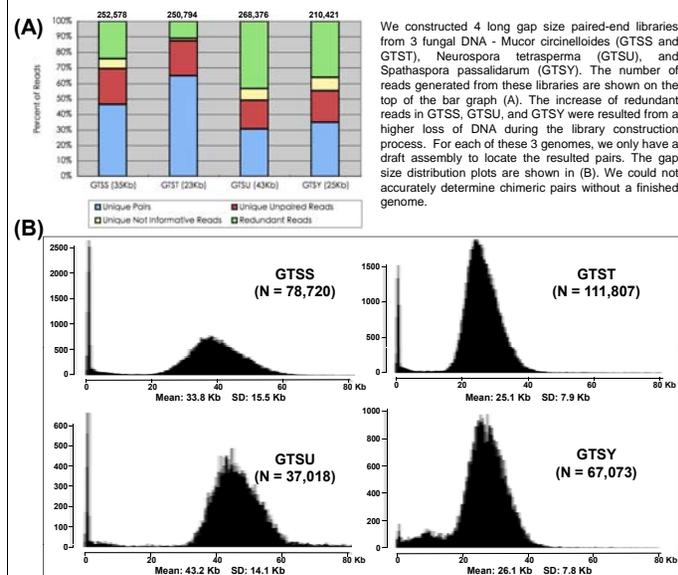
## Improvement of the Long Gap Size PE Library Construction

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
Double SPRI size selection	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 *Spathospora 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 a Fungal Genome with Long Gap Size Libraries

<i>Spathospora passalidarum</i> (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	22	22	22
N50 Contigs Length	153.94 Kb	211.37 Kb	205.22 Kb	196.77 Kb

We ran a set of test assemblies of the *Spathospora passalidarum* sequences using Newbler. They include the 454 shotgun reads (~33X depth) only (test assembly 1), the 454 shotgun and the new long 454 pairs (test assembly 2), the 454 shotgun and the fosmid pairs (test assembly 3), and the current assembly with the 454 shotgun, the fosmid pairs, and the old 454 pairs. The assemblies 2 and 3 resulted in comparable numbers of scaffolds, and large contigs, as well as the scaffold length and contig length. More assemblies with the *Mucor circinelloides* and *Neurospora tetrasperma* long 454 pairs are being investigated.

## Conclusions

1. We have observed some limitations of generating long gap size (≥ 30Kb) paired-end libraries using a modified the 454 Recombi Paired-end Library construction process
2. We further modified the process to improve the quality of the long gap size libraries
3. We have generated long gap size paired-end reads from 3 fungal genomes and the early evaluation of assemblies with these reads seems to be comparable with the fosmid end sequences

## 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, and Matt Hamilton and David Robinson of JGI for the sequencing support.