Low input mate-paired sequencing and its application on FFPE tumor samples

Zhoutao Chen, Tanya Biorac, Melvin Wei, George Marnellos, Charles Scafe, Bin Li, Tony Xu, Christopher Adams, Warren Tom, Jeffrey Ichikawa, Rob Bennett
Life Technologies, 5791 Van Allen Way, Carlsbad, CA 92008

ABSTRACT

Mate paired sequencing is critical for shotgun based whole genome sequencing and structural variation study using next generation massively parallel sequencing technologies. To construct long mate-paired (LMP) library with high complexity it usually requires tens of micrograms of input DNA, which is rarely the case for primary tumor samples. The least efficient step in the LMP protocol is a circularization step, which joins two ends of a molecule together to an internal adaptor to form a mate pair. We have developed a new intramolecular circularization (NIC) method, which improves efficiency of this circularization step dramatically. We have also found a way to stabilize these circularized DNA molecules. We used the nick translation method in our original LMP protocol to generate even-sized mate-paired tags with defined average length, and further optimized the protocol to generate longer mate tags and tighter tag size distribution for 2x60 mate-paired sequencing. Using the NIC method and improved procedure to construct LMP libraries with HuRef genomic DNA, we demonstrate several folds improvement on yield of LMP library over our original LMP method, and a reduced number of false positive mate reads. The reduced false positive mate rate potentially reduces the sequencing depth requirement for mapping accuracy. Furthermore, we have constructed LMP libraries from a couple of micrograms of genomic DNA isolated from formalin fixed paraffin embedded tumor samples. The quality and complexity of these libraries will be presented with SOLiD™ sequencing data. This low input mate-paired library construction method will further broaden the application of next generation sequencing technology for cancer genome research.

INTRODUCTION

SOLiD™ mate paired sequencing can generate paired end sequencing reads with a distance ranging from a few hundreds to tens of hundreds of base pairs apart depending on the researcher's preference. It provides critical information for shotgun based whole genome sequencing and structural variation study in a massively parallel fashion. SOLiD™ long mate paired (LMP) library construction method uses a patent pending nick translation approach to generate even sized mate paired tags. This approach not only produces most balanced tag sizes on each end, but is also very scalable on tag length to accommodate consistent increase on sequencing read length. Current SOLiD™ 4 LMP protocol requires 10-20µg genomic DNA to generate highly complex library for large genomes, such as human. This amount of DNA is very scarce for some cases, e.g. primary tumors.

Here we present a new and improved LMP protocol, which uses a new intramolecular circularization (NIC) method and optimized workflow which removes most column based purification steps. This new LMP protocol improves the library yield 3 to 10 folds depending on the insert size and input amount. It has eliminated adaptor dimer carry-over in the final library and further decreased false mate pairing rate. The sequencing quality of LMP libraries constructed from 1 and 5µg of HuRef genomic DNA are presented. Very highly complex LMP library can be successfully constructed from 5µg HuRef gDNA. The complexity of LMP libraries constructed with 1µg HuRef DNA is also reasonable for applications requiring relatively low depth of sequence coverage. Furthermore, we evaluate the NIC LMP method on FFPE samples from a breast cancer patient. The preliminary results from these libraries are very encouraging. We are looking for collaborations to further develop this approach for cancer genome research.

MATERIALS AND METHODS

For LMP library preparation, 5500 SOLiD™ Mate-Paired Library Construction Kits (Beta) were used. Emulsion beads were prepared using SOLiD™ EZ bead™ system or manual ePCR procedure. SOLiD™ sequencing data were generated from SOLiD™ 4 System. Figure 1 shows the major changes in the 5500 SOLiD™ mate-paired library construction workflow compared with SOLiD™ 4 mate-paired library construction workflow.

RESULTS

Figure 1. Major improvements in the 5500 SOLiD™ mate paired library protocol

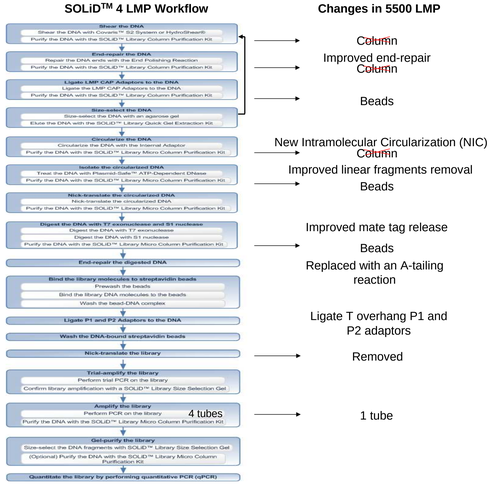
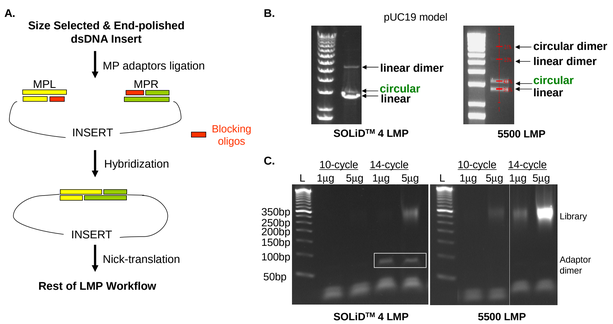


Figure 2. New intramolecular circularization improves circularization efficiency significantly



A. Schematic of NIC method in the 5500 LMP. ***B***. 5500 LMP shows significant improvement on circularization efficiency over SOLiD 4 LMP. ***C***. LMP trial PCR products. 3kb insert HuRef LMP libraries using 5500 LMP protocol show much higher yield than those using SOLiD 4 protocol and no detectable adaptor dimers.

Table 1. 5500 LMP libraries use less starting material but show better sequencing quality

A.

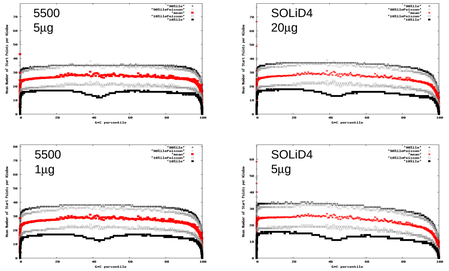
	Single Tag Mapping Statistics (2x50)	5500 LMP (Quad)		SOLiD 4 LMP (Quad)	
		5ug	1ug	20ug	5ug
F3	Total mapped reads (millions)	65.3	63.9	61.4	65.6
	Megabases of coverage	3017.7	2897.3	2816.7	3029.0
	Average number of uniquely mapped reads per start point	1.06	1.20	1.10	1.14
R3	Total mapped reads (millions)	75.8	77.5	76.1	77.6
	Megabases of coverage	3,483.2	3,532.4	3,496.3	3,573.4
	Average number of uniquely mapped reads per start point	1.06	1.24	1.10	1.15

B.

Pairing Statistics	5500 LMP (Quad)		SOLiD 4 LMP (Quad)	
	5ug	1ug	20ug	5ug
Mapped single reads (millions)	143.4	144.5	139.8	145.9
Properly paired	73.5%	70.7%	67.2%	72.7%
Singletons	12.7%	17.0%	15.8%	12.8%
Mate mapped to a different chromosome	3.9%	2.2%	6.5%	4.6%

1kb insert LMP libraries were constructed using HuRef genomic DNA and 2x50 sequenced on SOLiD 4 system. Human Genome Assembly hg19 was used for mapping with Bioscope v1.3.

Figure 3. Improved G/C coverage on 5500 LMP libraries



The same libraries, sequencing and mapping conditions as those in the Table 1.

To assess the library performance at higher sequencing coverage and library quality for SNP calling for a human sample, we used 5µg HuRef genomic DNA to prepare a 1kb insert mate paired library using the 5500 LMP protocol. 2x60 mate paired sequencing was performed on a full slide with Exact Call Chemistry (ECC) on a SOLiD™ 4 system. More than 55Gb mapped sequences were generated (Table 2). Excellent genotype concordance was observed compared with HuRef genotypes (Fig 4).

Table 2 . Mapping statistics of a full slide 2x60 sequencing run with ECC

	Single Tag Mapping Statistics (2x60, one slide)	5ug HuRef LMP, 1kb insert
F3	Total mapped reads (millions)	508
	Megabases of coverage	27,514
	Average number of uniquely mapped reads per start point	1.40
R3	Total mapped reads (millions)	507
	Megabases of coverage	27,773
	Average number of uniquely mapped reads per start point	1.36

Figure 4. High Concordance with HuRef genotypes

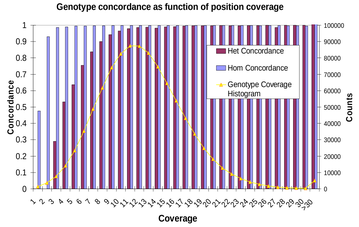
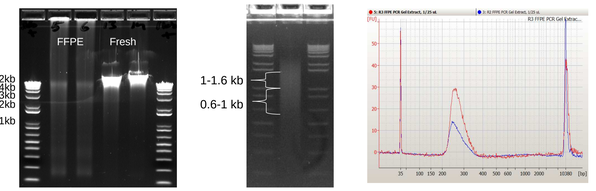


Figure 5. Preparation of LMP libraries from 2µg of genomic DNA extracted from FFPE samples

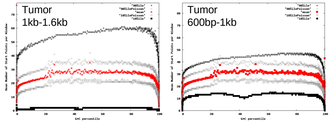


To evaluate 5500 LMP protocol on FFPE tumor samples, we used one year old FFPE tumor and control samples from a breast cancer case. **A.** Illustration of the genomic DNA quality on an E-gel. PureLink Genomic DNA Mini kits were used to extract gDNA from both FFPE and fresh frozen tissue samples. **B.** Illustration of size selection method. Each sample was sheared to ~1kb using Covaris S2. The DNA fragment ranging from 0.6-1kb and 1-1.6kb were used to construct two LMP libraries per sample. **C.** Example Bioanalyzer electropherogram of the final LMP libraries using high sensitive DNA chip. The final libraries were amplified either 12 cycles or 14 cycles.

Table 3. 2x50 Sequencing mapping statistics of R3 tag of FFPE LMP libraries

Sample	Size Fraction	PCR cycle #	Total reads (millions)	% Uniquely Mapped	Unique Reads Per Start
Tumor	1kb-1.6kb	14	97	63.1%	7.62
	600bp-1kb	12	104	65.5%	2.45
Control	1kb-1.6kb	14	95	66.5%	5.46
	600bp-1kb	12	98	62.3%	2.78

Figure 6. G/C coverage of FFPE LMP Libraries



We observed many more large indels in the tumor samples than the normal control. Further analyses on these FFPE LMP sequencing data are underway to address the biological implications discovered by SOLiD LMP sequencing.

CONCLUSIONS

We demonstrated that 5500 SOLiD™ LMP library protocol had superior performance over the SOLiD™ 4 LMP protocol. In the 5500 LMP protocol we developed a more efficient circularization method, further optimized key enzymatic reactions and removed or replaced six column purification steps. With these improvements, we observed 3 to 10 folds library yield increase depending on the insert size and the amount of starting material, and further reduced false positive mate rate using the 5500 LMP kit. Using Exact Call Chemistry we demonstrated excellent HuRef genotype concordance.

Using the 5500 LMP protocol, we also successfully constructed LMP libraries from 2µg DNA extracted from FFPE tumor samples for the first time. Although the complexity of these libraries dropped significantly compared to libraries from non-FFPE samples, we were still able to get over 15× clone coverage from two quads of sequencing data. With further improvements, we believe SOLiD™ mate paired sequencing will become a powerful new tool for investigators to mine the valuable FFPE collection and unlock the mysteries of many diseases.

ACKNOWLEDGEMENTS

Many thanks to the unconditional support from SOLiD molecular biology team and 5500 development team which made the work presented here possible.

TRADEMARKS/LICENSING