A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost

Darren Heavens, Gonzalo Garcia Accinelli, Bernardo Clavijo, and Matthew Derek Clark
The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK

BioTechniques 59:42-45 (July 2015) doi 10.2144/000114310
Keywords: long mate pair (LMP) construction; insert sizes; genome assembly; wheat

Supplementary material for this article is available at www.BioTechniques.com/article/114310.

Long mate pair (LMP) or "jump" libraries are invaluable for producing contiguous genome assemblies and assessing structural variation. However the consistent production of high quality (low duplication rate, accurately sized) LMP libraries has proven problematic in many genome projects. Input DNA length and quantity are key issues that can affect success. Here we demonstrate how 12 libraries covering a wide range of jump sizes can be constructed from <10 µg of DNA, thus ensuring production of the best LMP libraries from a given DNA sample. Finally, we demonstrate the accuracy of the insert sizes by mapping reads from each library back to an existing assembly.

Standard paired-end next-generation sequencing projects can produce long continuous sections of sequence (contigs), but these alone lack the long-range information required to produce single contig assemblies of even bacterial chromosomes (1). Assemblies based on paired-end data alone are unable to resolve repeated sequences that are bigger than the insert size of the library (typically ~500 bp). The genomes of some higher eukaryotes can consist of >80% repeated sequences (2), and this can result in highly fragmented genome assemblies containing many thousands or even millions of small contigs.

In order to increase assembly contiguity, many projects use long mate pair (LMP) libraries to jump over repeated sequences to connect contigs, a process known as scaffolding (3). Depending on the quantity and quality of the available input DNA it is possible to generate LMP libraries with insert sizes ranging from 1.5 kb to 40 kb. High quality assemblies typically use multiple LMP libraries of different insert sizes, which is costly in terms of input DNA quantity, time, and money. LMP libraries are also notoriously difficult to make, especially for the larger insert sizes.

Using the Illumina Nextera Mate Pair Sample Preparation Kit (Illumina, San Diego, CA), libraries can be constructed from as little as 1 µg of genomic DNA (gDNA) using the Nextera transposase to fragment DNA and tag the molecules with known sequences (a process known as tagmentation). However, these libraries tend to have a broad insert size which can range from 1 kb to 12 kb (Supplementary Figure S2). As a result, many labs employ gel-based size selection to generate specific insert sizes that can be supplied to the scaffolding algorithm, thereby simplifying the scaffolding step. Semi-automated gel approaches such as BluePippin (Sage Science, Beverly, MA) improve this process but limit throughput to four libraries at a time and use more input DNA. Constructing 4 LMP libraries, could require >18 µg of DNA, and if insert sizes >10 kb are targeted, each size selection run would last longer than 6 h, meaning that library construction could take up to 3 days to complete (Figure 1). Furthermore, in our experience it is hard to predict how a specific DNA sample will perform in a tagmentation reaction, so more than one reaction is often needed to obtain a specific size. Finally, there can be 10%–20% variance between the targeted and recovered DNA size on a BluePippin.

We optimized the Nextera based LMP Library Construction kit to maximize fragmentation across the largest possible size range using the minimum amount of input material. Using gDNA isolated from the bread wheat (*Triticum aestivum*) variety Chinese Spring 42, we performed just 2 Gel Plus tagmentation reactions and subsequent strand displacements to construct 12 LMP libraries. This allows us to construct 60 LMP libraries from 5 samples using a 10-reaction kit. As fragment size in a Nextera reaction is controlled by the ratio of DNA and Nextera enzyme, one reaction was performed with 3 µg of input DNA, and another with 6 µg. The two Nextera reactions were then pooled post strand displacement, and the range of fragment sizes confirmed by analyzing the profiles on

METHOD SUMMARY

We present a method to simultaneously size select and construct up to 12 long mate pair (LMP) libraries at a time and then map the generated reads back to the available assembled sequences to accurately calculate insert sizes. These calculations can then be used to determine which libraries to sequence to greater depth and to use the accurate insert size information in de novo genome assemblies to improve outputs.

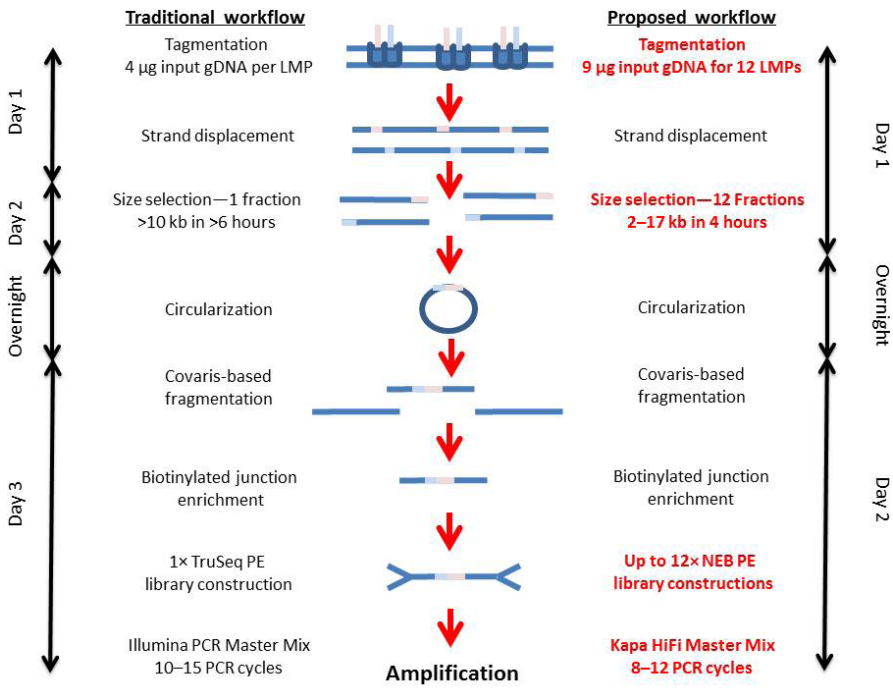

Figure 1. Nextera-based long mate pair (LMP) workflow. The traditional LMP workflow compared with our proposed workflow with differences between the two workflows highlighted in red.

an Agilent BioAnalyser 12000 chip (Agilent, Stockport, UK) (Supplementary Figure S1). By using 2 independent tagmentation reactions, we ensured the material entering size selection ranged from 1.5 kb to >17 kb with a good distribution, allowing us to construct LMP libraries from a wide range of insert sizes.

Size selection was performed on a Sage Science Electrophoretic Lateral Fractionator (SageELF), which is unique in its ability to simultaneously isolate 12 different discrete size fractions from a single sample loading. The pooled, strand-displaced reactions were loaded onto a 0.75% cassette, which was configured to separate the sample for 3 h 30 min and then elute 12 fractions over 35 min. Post size selection, the size of each of the 12 isolated fractions was measured on an Agilent BioAnalyser Chip 12000 (Figure 2A and Table 1), and the yield was determined using a High Sensitivity Qubit Assay (Thermo Fisher, Cambridge, UK) (Supplementary Table S1).

We loaded 5 µg of DNA onto the SageELF and recovered >2 µg across the 12 fractions, which represents >40% of the starting material. Fraction 5 encompassed an important LMP target insert size of 8 kb (a very common transposon in wheat is ~7 kb). For this size, we managed to recover >180 ng of material (Supplementary Table S1). To compare this against our standard approach, we tagmented and strand displaced 4 µg of the same wheat gDNA, and confirmed the fragmentation profile on a 12000 BioAnalyser Chip (Supplementary Figure S1). After targeting an 8 kb (7.4–8.6 kb) size selection on a BluePippin, with the improved recovery protocol we recovered only 56 ng of material. When we ran this out on a 12000 BioAnalyser Chip, it estimated the fragments to be centered on 9.5 kb and spanning 8.0–10.5 kb (data not shown), which illustrates the problem with targeting specific insert sizes. For the comparable SageELF fraction (Fraction 4) we recovered 261 ng of material centered on 9.5 kb and spanning 8.5- 10.6 kb (data not shown) highlighting that size selection is not only tighter, but we also observed significantly higher recoveries when using the SageELF.

Circularization of the SageELF fractions was then performed overnight at 30°C, followed by exonuclease (Illumina) treatment at 37°C for 30 m, incubation at 70°C for 30 m to denature the enzyme, and then addition of Stop Ligation buffer (Illumina). Circularized fragments were then sheared on a Covaris S2 (Covaris, Woburn, MA), targeting a 450 bp shear, and then library molecules containing the biotinylated junction adapter were bound to M280 streptavidin-coated beads (Thermo Fisher). Fragmented molecules from each of the 12 size-selected fractions were end repaired and A-tailed using the relevant NEB modules (NEB, Hitchin, UK) and then Illumina TruSeq adapters (Illumina) were ligated (each size fraction received a different index) with NEB Blunt T/A ligase (NEB).

We used Kapa HiFi polymerase (Kapa Biosystems, London, UK) for its improved performance, especially in GC rich regions, instead of the Illumina PCR master mix (4). Post size selection, we calculated the copy number of each fraction based on the predicted size from the SageELF and the yield to measure the library complexity. For samples with a copy number >3.75 × 10^{10} we performed 8 PCR cycles, for samples with a copy number between 2 × 10^{10} and 3.75 × 10^{10}, 10 cycles were performed, and for samples with a copy number <2 × 10^{10}, 12 cycles were performed. The library molecules were amplified directly from the

Table 1. Sizes of long mate pair (LMP) inserts for each fraction as determined by the SageELF, BioAnalyser, and mapping reads back to the wheat chromosome 3B assembly.

Fraction	ELF library size (kb)	BA 12000 library size (kb)	Mapped insert size (kb)
1	16.18	Not determined	Insufficient data
2	13.31	Not determined	14.8
3	11.74	12.52	11.3
4	9.81	9.24	9.0
5	8.00	8.03	7.3
6	6.46	6.68	5.9
7	5.16	5.37	4.8
8	4.28	4.31	3.8
9	3.70	3.46	3.2
10	2.93	2.66	2.4
11	2.22	2.16	1.9
12	1.71	1.67	1.4

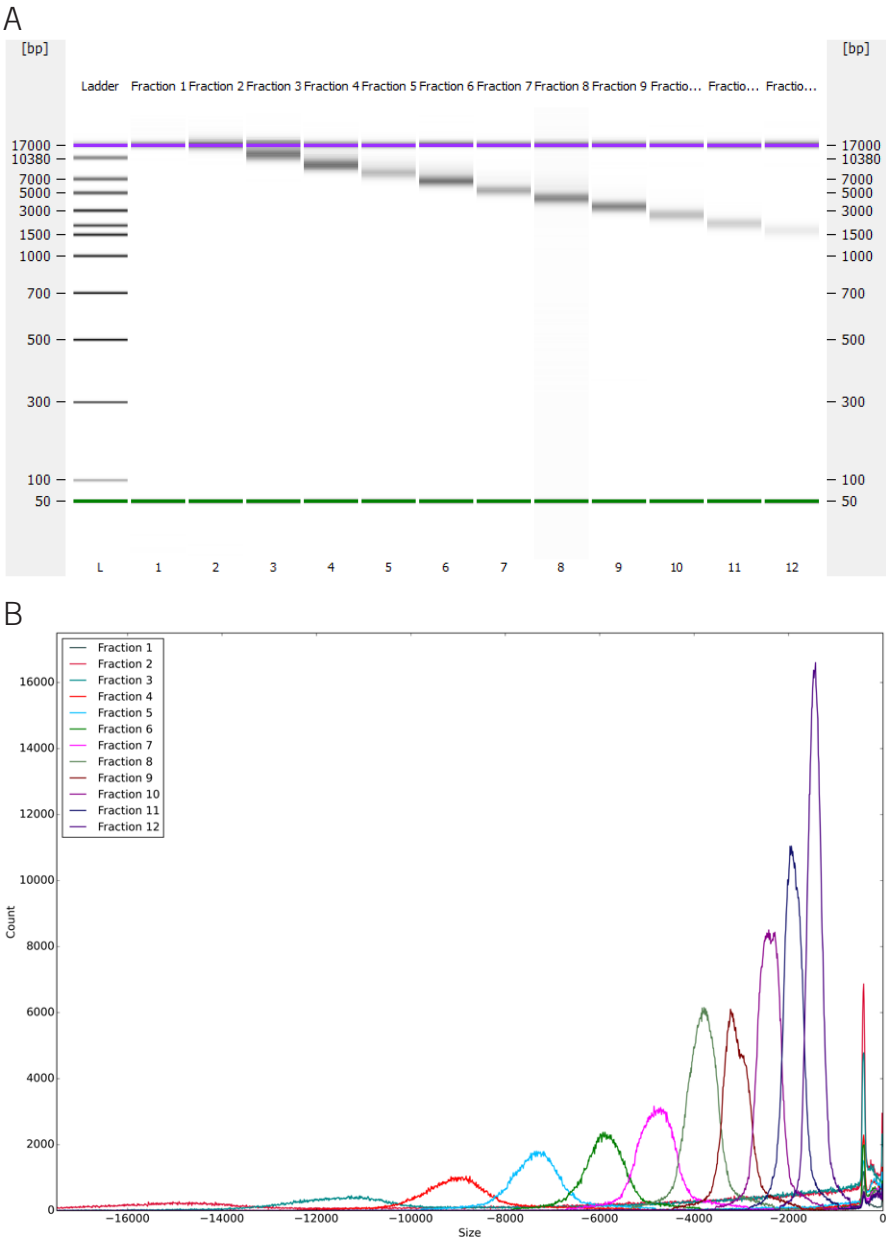

Figure 2. BioAnalyser images of DNA post size selection and size distribution of BWA mapped reads. (A) SageELF size-selected fractions were analyzed to estimate fragment length prior to circularization. (B) NextClip filtered reads from each size-selected fraction library were aligned against the wheat chromosome 3B assembly and the number of reads vs. insert size plotted.

streptavidin beads using Kapa HiFi and the Illumina primer cocktail (Illumina). We aimed to maintain library complexity and reduce PCR duplication rates while generating sufficient material for multiple HiSeq runs (Supplementary Table S1).

Post amplification, a CleanPCR (GC Biotech, Alphen aan den Rijn, The Netherlands) bead clean-up was carried out, and the final library was eluted in 20 µl resuspension buffer (Illumina). Library quality controls were performed by running an Agilent BioAnalyser High Sensitivity chip, and the DNA concentrations were measured using the High Sensitivity Qubit assay (Supplementary Table S1). Equimolar amounts of LMP libraries from fractions 2–12 were then pooled, with the library from fraction 1 spiked in at one-tenth the concentration of the others due to it being relatively weaker (Supplementary Table S1). The 12 pooled libraries were size selected on a BluePippin to ensure that all library fragments would have insert sizes between 370 and 470 bp (maximizing usable mate pairs) and then quantified using the Kapa qPCR Illumina Quantification kit.

To validate the pool and accurately determine the insert size of each LMP, the pools were run on a MiSeq (Illumina) with 2 × 300 bp reads. Sequence data were screened via a primary analysis pipeline to demultiplex reads based on library indexes and to determine basic run metrics, including duplication rate, GC content, and the presence of over-represented sequences (5). The data were then processed through NextClip (6) to classify LMP reads. Those deemed as true mate pairs, based on the presence of the Nextera junction sequence within the reads with sufficient sequence either side, were then mapped using BWA-mem (7) to the bread wheat (*Triticum aestivum*) variety Chinese Spring 42 chromosome 3B reference sequence (8) using default parameters, and the insert size for each library determined and plotted (Figure 2B and Table 1).

Using the SageELF streamlines the library construction process, allowing LMP libraries >10 kb to be constructed in under 2 days with <10 µg input material. For many genome projects, multiple insert size LMP libraries are required, and the ability to construct up to 12 discretely sized libraries for a combined reagent cost of $1270 compared with the reagent cost of $715 for a single insert size LMP library highlights the potential cost savings. We also observe significant improvements with increased yield and tighter size selection than when using the BluePippin, especially when looking to construct LMP libraries with insert sizes >10 kb.

Accurately determining the size and span of the inserts for mate pair libraries simplifies the scaffolding problem, enabling the assembly of longer, more precise sequences with fewer non-determined bases (runs of N bases), empowering all subsequent downstream analysis. Although the BioAnalyser and SageELF both estimate the size of fraction 5 to be 8 kb, mapping the sequence data back to the wheat chromosome 3B assembly suggested that the size is in fact 7.2 kb (Table 1). This demonstrates the benefit of this approach both in terms of accuracy in determining insert size and also the ability to sequence slightly larger or slightly smaller insert libraries without having to repeat the whole process if one library isn't deemed suitable. It also gives the flexibility of running all 12 libraries if desired.

Author contributions

D.H. wrote the manuscript and carried out the experiments. G.G.A. and B.C. analyzed the sequence data. D.H., B.C., and M.D.C. had the original idea and designed the study. D.H., G.G.A., B.C., and M.D.C. edited the manuscript. B.C. and M.D.C. supervised the study.

Acknowledgments

Wheat gDNA was provided by Neil Mckenzie and Mike Bevan, John Innes Centre. Library quantification, the MiSeq Sequencing, and Primary Analysis Pipeline were run by the Platforms and Pipeline Team at TGAC**.** This work was supported by a BBSRC Triticeae Genomics for Sustainable Agriculture Grant, BB/J003743/1, and a BBSRC National Capability Grant, BB/J010375/1.

Competing interests

The authors declare no competing interests.

References

1. **Magoc, T., S. Pabinger, S. Canzar, X. Liu, Q. Su, D. Puiu, L.J. Tallon, S.L. Salzberg**. 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics. *29*:1718-1725.
2. **Treangen, T.J. and S.L. Salzberg.** 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. *13*:36-46.
3. **Nagarajan, N. and M. Pop.** 2013. Sequence assembly demystified. Nat. Rev. Genet. *14*:157-167.
4. **Quail M.A., T.D. Otto, Y. Gu, S.R. Harris, T.F. Skelly, J.A. McQuillan, H.P. Swerdlow, S.O. Oyola** 2011. Optimal enzymes for amplifying sequencing libraries. Nat Methods. *9*:10-11.
5. **Leggett, R.M., R.H. Ramirez-Gonzalez, B.J. Clavijo, D. Waite, and R.P. Davey.** 2013. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front Genet. *4*:288.
6. **Leggett, R.M., B.J. Clavijo, L. Clissold, M.D. Clark, M. Caccamo.** 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics. *30*:566-569.
7. **Li, H. and R. Durbin.** 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics *25*:1754-1760.
8. **Choulet, F., A. Albert, S. Theil, N. Glover, V. Barbe, J. Daron, L. Pingault, P. Sourdille, et al.** 2014. Structural and functional partitioning of bread wheat chromosome 3B. Science. *345*:1249721.

Received 20 February 2015; accepted 13 April 2015.

Address correspondence to Darren Heavens, The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK, NR4 7UH. E-mail: darren.heavens@tgac.ac.uk