A method to simultaneously construct up to 12 different sized Illumina Nextera long mate pair libraries with reduced DNA, time and costs.

Darren Heavens [1], Gonzalo Garcia Accinelli [1], Bernardo Clavijo [1], Will Deacon [2], Chris Boles [3] and Matthew Clark [1]

(1)The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, UK, NR4 7UH
(2) Labtech, 2 Birch House, Brambleside, Bellbrook Ind. Est, Uckfield, East Sussex, TN22 1QQ
(3) Sage Science, Suite 2400, 500 Cummings Center, Beverly, MA 01915, USA

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. For some higher eukaryotes, which can contain over >80% repeated sequence[2], this can result in highly fragmented genome assemblies consisting of 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 DNA it is possible to generate LMP libraries with insert sizes from 1.5kbp to 40kbp. High quality assemblies typically use multiple LMP libraries of different insert sizes, which is costly both in DNA, time and money. They are also notoriously difficult to make, especially for the larger insert sizes.

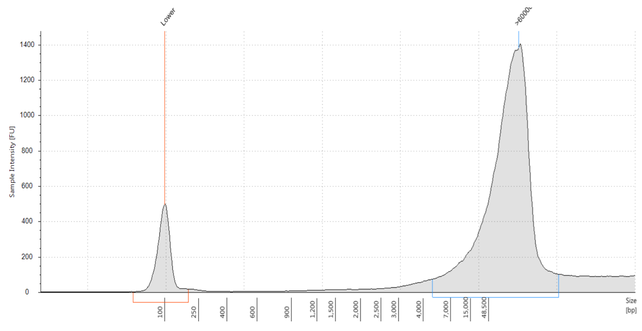

Figure 1. Size assessment of DNA prior to LMP construction. Agilent Genomic Tape electropherogram for Bread Wheat.

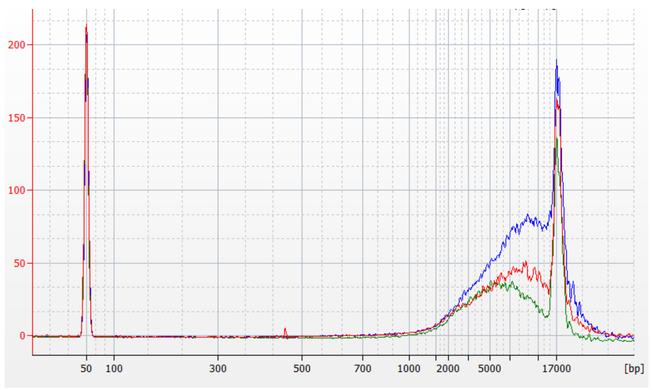

Figure 3. BioAnalyser Images of tagmented and strand displaced DNA pre size selection. Electropherograms of tagmented DNA with 3µg (green) and 6µg (blue) input DNA and the pooled strand displaced DNA (red).

Fraction	ELF library Size (kbp)	BA 12000 Library Size (kbp)	Mapped Insert Size (kbp)
1	16.1	Not Determined	Insufficient data
2	13.3	Not Determined	14.8
3	11.7	12.5	11.3
4	9.8	9.2	9.0
5	8.0	8.0	7.3
6	6.4	6.6	5.9
7	5.1	5.3	4.8
8	4.2	4.3	3.8
9	3.7	3.4	3.2
10	2.9	2.6	2.4
11	2.2	2.1	1.9
12	1.7	1.6	1.4

Table 1. Sizes of LMP inserts for each fraction as determined by the ELF, BioAnalyser and when reads were mapped back to the Bread Wheat Chromosome 3B reference.

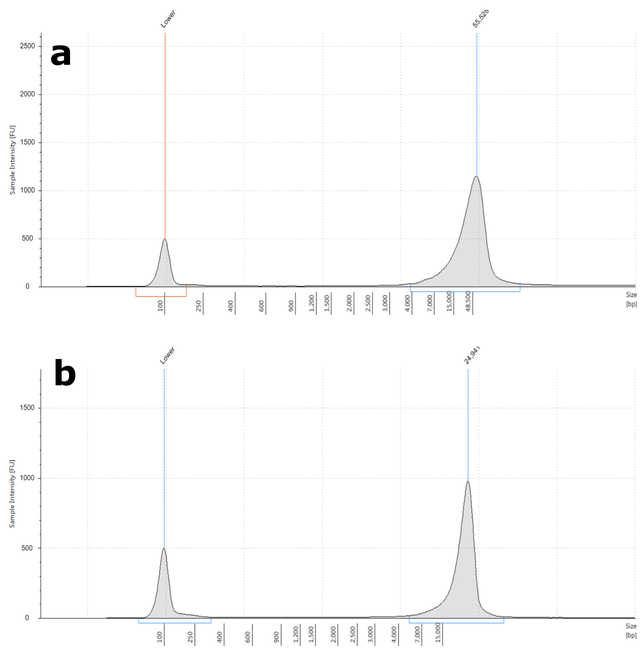

Figure 6. Size assessment of DNA prior to LMP construction. Agilent Genomic Tape electropherograms for a) Durum Wheat and b) European Ash

References

1. Tanja Magoc, Stephan Pabinger, Stefan Canzar, Xinyue Liu, Qi Su, Daniela Puiu, Luke J. Tallon, Steven L. Salzberg. 2013. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics. July 15; 29(14): 1718–1725
2. Todd J. Treangen, Steven L. Salzberg. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. November 29; 13(1): 36–46
3. Niranjan Nagarajan, Mihai Pop. 2013. Sequence assembly demystified. Nature Reviews Genetics 14, 157-167
4. Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, Swerdlow HP, Oyola SO. 2011. Optimal enzymes for amplifying sequencing libraries. Nat Methods. Dec 28; 9(1):10-1
5. Richard M. Leggett, Ricardo H. Ramirez-Gonzalez, Bernardo J. Clavijo, Darren Waite, Robert P. Davey. 2013. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Front Genet. 4: 288.
6. Richard M. Leggett, Bernardo J. Clavijo, Leah Clissold, Matthew D. Clark, Mario Caccamo. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics. February 15; 30(4): 566–56
7.Li H. and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60
8. Frédéric Choulet, Adriana Albert, Sébastien Theil, Natasha Glover, Valérie Barbe, Josquin Daron, Lise Pingault, Pierre Sourdille *et al*. 2014. Structural and functional partitioning of bread wheat chromosome 3B. Science. Jul 18;345(6194):124972

Methodology

Starting with High Molecular Weight (HMW) Bread Wheat DNA, >60kbp (Figure 1), we optimised the Nextera based LMP Library Construction kit to fragment across the largest possible size range whilst using the minimum amount of input material (Figures 2 and 3). The tagmented samples were pooled post strand displacement and size selection performed on a Sage Science Electrophoretic Lateral Fractionator (ELF) resulting in the isolation of 12 discrete size fractions (Figure 4 and Table 1).

Following circularisation, fragmentation and enrichment of the fragments containing the biotinylated Nextera junction adapter, molecules from each of the 12 size selected fractions were independently end repaired, A-tailed and had unique Illumina compatible adapters ligated. Kapa HiFi polymerase [4] was used to amplify viable library molecules for each individual fraction and then all 12 LMP libraries pooled. The pool was then size selected to ensure that all library fragments would have insert sizes between 370 and 470bp and sequenced as a 2x300bp MiSeq run.

Data Analysis

Post sequencing the duplication rate and the presence of over represented sequences were determined[5] and the data processed through NextClip[6]. True mate pairs were then mapped using BWA-mem[7] to the Bread Wheat (*Triticum aestivum*) variety Chinese Spring 42, chromosome 3B reference[8] and the insert size for each library determined and plotted (Table 1 and Figure 5).

Although the BioAnalyser and ELF both estimate the size of fraction 5 to be 8kbp when the sequence data is mapped backed to the Bread Wheat chromosome 3B assembly it suggests that it is in fact 7.2kbp (Table 1). This demonstrates the benefit of the ELF based approach both in terms of accuracy in determining insert size but also being able to sequence a slightly larger or slightly smaller insert libraries without having to repeat the whole process if one library isn't deemed suitable. It also give the flexibility of running any combination of the 12 fractions if desired.

For genome projects requiring multiple insert size LMP libraries the ability to construct up to 12 discretely sized, individual libraries for a combined reagent cost of £800 compared to the reagent cost of £450 for a single insert size LMP library highlights the potential cost savings.

Advantages Over The Traditional Approach

With single size LMP libraries we targeted libraries with inserts up to one third of the size of the starting material (Figures 1 and 6). Analysing the results we achieved when using the ELF for Bread Wheat, Durum Wheat and European Ash LMP libraries (Table 2) shows the benefit of this global approach. We aim for LMP libraries with a high proportion of true mate pairs (>65%) and a low duplication rate (<2%). For the European Ash (25kbp starting material) we have been able to construct a LMP library with a much larger insert size (15kbp) than we would have traditionally targeted (8kbp).

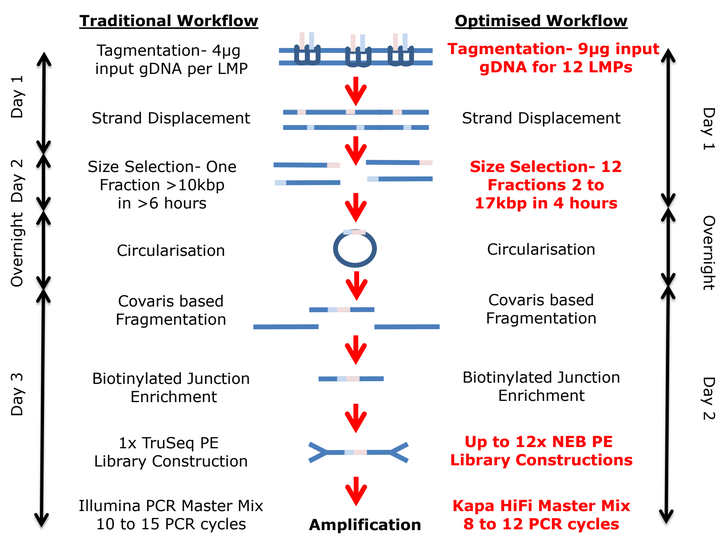

Figure 2. Nextera- based LMP workflow. The traditional LMP workflow compared to our optimised workflow with differences between the two highlighted in red.

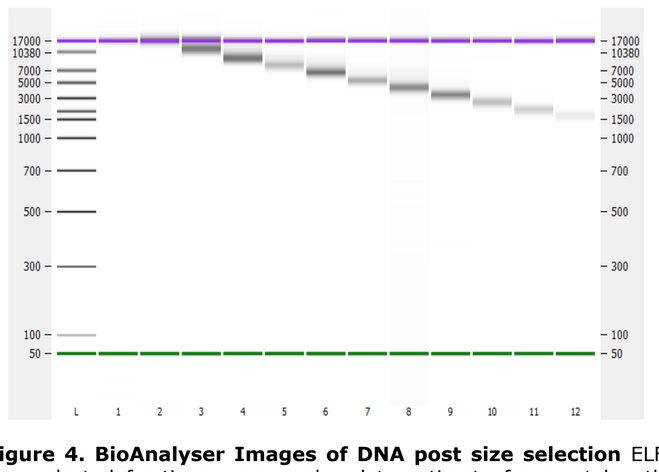

Figure 4. BioAnalyser Images of DNA post size selection ELF size selected fractions were analysed to estimate fragment length prior to circularisation.

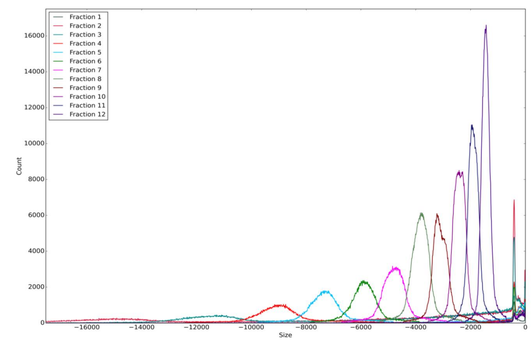

Figure 5. Size distribution of BWA mapped reads. NextClip processed reads from each size selected fraction library were aligned against the Bread Wheat Chromosome 3B assembly and the number of reads v insert size plotted.

Sample	ELF predicted Size (Kbp)	Duplication Rate (%)	True Mate Pairs (%)	Mapped Insert Size (Kbp)
Bread Wheat	11.7	1.7	68.9	11.4
Durum Wheat	9.8	1.1	80.7	9.9
European Ash	13.3	0.3	73.7	15.1

Table 2. Elf predicted insert size, Duplication Rates and True Mate Pairs as determined by NextClip and actual insert size of libraries when mapped back to the available assembly for the optimal ELF size selected fraction LMP libraries for Bread Wheat, Durum Wheat and European Ash.

Acknowledgements

Wheat gDNA was provided by Neil Mckenzie and Mike Bevan, JIC. Library quantification, the Sequencing and Primary Analysis Pipeline 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, a BBSRC National Capability Grant, BB/J010375/1 and an EU Framework 7 Programme (TransPLANT, award 283496).

Contact Details

Email:darren.heavens@tgac.ac.uk

Tel: +44 (0) 1603450838

TGAC The Genome Analysis Centre

BBSRC Greater Norwich Development Partnership