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\(^2\). 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\(^3\), 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.

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 tagged 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 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 duplicate 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 assembly 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 library without having to disrupt 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 Wheat Libraries (Table 2) shows the benefit of this global approach. We aim for LMP libraries of 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).

References


Acknowledgements

Wheat gDNA was provided by Neil McKenzie and Mike Bevan, JIC. Library quantification of Sequencing and Primary Analysis Pipeline run by the Platforms and Pipeline Team at TGAC. This work was supported by a BBSC Triticaceae Genomics for Sustainable Agriculture Grant, BB/J003734/1, a BBSC National Capability Grant, BB/J003735/1 and an EU Framework 7 'TransPLANT' (award 283496).

**Contact Details**

Email: darren.heavens@tgac.uk

Tel: +44 (0) 1603450838