Random Shear BAC Library Enables Gap Closing in the Arabidopsis Genome

Lucigen Corporation, Middleton, WI 53562, USA, *Monsanto Company, St. Louis, MO 63167, USA

Abstract

The Arabidopsis thaliana genome was first sequenced in 2000. The reference was built using multiple sub cloning methods paired with Sanger sequencing and has been considered a gold standard plant genome for many years. However, there are numerous gaps in the Arabidopsis reference genome due to inherent biases in the techniques and the repetitive nature of these areas. New technologies such as third generation single molecule sequencing merged with techniques that do not have the same biases allow us to better sequence and assemble these previously difficult or impossible genomic locations. Lucigen Corporation (WI) created a randomly sheared Arabidopsis BAC library and screened it for the flanking sequences near known gaps in the reference. These BACs were then sequenced using Illumina short reads and Pacific Biosciences long reads. This approach has enabled us to close or extend into many of the known gaps and provide resolution and annotation in these highly repetitive regions.

Lucigen’s Random Shear BAC Library

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size (Mb)</th>
<th>BAC libraries</th>
<th>No. of chromosomes</th>
<th>No. of contigs</th>
<th>Genomic gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis</td>
<td>125</td>
<td>Two -17X</td>
<td>5</td>
<td>27</td>
<td>&lt;5%</td>
</tr>
<tr>
<td>Rice</td>
<td>430</td>
<td>Two -26X</td>
<td>12</td>
<td>284</td>
<td>&lt;10%</td>
</tr>
<tr>
<td>Soybean</td>
<td>1,155</td>
<td>Three -10X</td>
<td>20</td>
<td>2,905</td>
<td>&lt;10%</td>
</tr>
<tr>
<td>Maize</td>
<td>2,500</td>
<td>Three -15X</td>
<td>10</td>
<td>3,488</td>
<td>unknown</td>
</tr>
<tr>
<td>Drosophila</td>
<td>97</td>
<td>One -14X</td>
<td>4</td>
<td>9</td>
<td>&gt;2%</td>
</tr>
<tr>
<td>Human</td>
<td>3,200</td>
<td>Five -15X</td>
<td>23</td>
<td>1,246</td>
<td>&lt;4%</td>
</tr>
<tr>
<td>Mouse</td>
<td>3,200</td>
<td>Two -33X</td>
<td>20</td>
<td>296</td>
<td>&lt;10%</td>
</tr>
</tbody>
</table>

Table 1. Published Genomes

Listed are some published BAC-based whole genome physical maps. All of these genomes contain numerous gaps that represent a significant fraction of the genomes, even the best “finished” genomes. The gaps exist despite using as many as 5 BAC libraries and up to 33-fold redundancy.

Figure 1. Shearing comparison

Restriction digestion has significant limitations when making BAC libraries. The red arrow indicates undigestable megabase-size genomic DNA, which are from genomic regions, such as centromeres, highly repetitive sequences, telomeres, etc., that may completely lack recognition sites for common restriction enzymes (e.g., BamHI, EcoRI, and HindIII). These regions will not be represented in the BAC libraries and will result in gaps in the assembly. The solution for this is random shear BAC cloning. The same HMW genomic DNA can be randomly sheared into large fragments of 100-400 kb. Significantly, there is no megabase-size DNA left behind. In the other words, the DNA from all genomic regions is randomly and equally sheared, which allows it to be cloned into BACs.

Figure 2. Closing centromeric gaps with Random Shear BAC library

Lucigen created a random shear A. thaliana BAC library consisting of 7680 clones with an average insert size of 100kb. This is ~5x coverage of the genome. They screened this library using Overgo probes for some of the regions flanking known gaps to isolate individual randomly sheared BAC clones that span or reach into these gaps. The dashed lines represent the probe locations and the red lines represent the clones that were isolated using the probes. Now that we can isolate and span these gaps, the challenge becomes sequencing and assembling through these previously difficult regions.

Figure 3. Assemblies including PacBio long read data were more contiguous

Three assemblies were performed using the data generated for Haibao Tang etc./JCVI. a PacBio only assembly using ALLORA, an Illumina only assembly using CLC, and a hybrid assembly using Celera (PacBio reads were error corrected using raw Illumina reads). The ALLORA assembly generated 4 total contigs, one of which was 99kb. The Celera assembly also produced one contig of 99kb and a total of 5 contigs. The CLC assembly generated >7000 contigs, but only 3 were >1kb with one >70kb.

The longest ALLORA and Celera contigs and the 3 longest CLC contigs were then mapped to the reference using cross match. The figure above depicts the alignment of each of them. Both the ALLORA assembly and the Celera assembly appear to have assembled the BAC into a single contig. The CLC assembly also performed well, but could not create a single contig from the data: the three contigs have <1kb gaps between them. The known gap size in all 3 assemblies was ~20kb.

Figure 4. Mapping raw BAC sequence data reveals BAC location

PacBio data was filtered and mapped against the plasmid backbone using BLASTR to determine end sequences. The blue bars represent the coverage when all reads were mapped against the TAIR v9 reference genome using BLASTR. The top two BAC maps very well to locations with a single gap (1329bp and 235bp gaps, respectively). The bottom BAC maps to several locations, likely indicating that this BAC is from a repetitive region of the genome. The middle BAC only has ~70kb mapping to the reference: this gap is likely much larger than the 235bp in the annotation and was used for further analysis.

Figure 5. This gap sequence is real and appears conserved

This region appears to be conserved in several A. thaliana lines. Raw Illumina data from 5 lines was mapped to the longest contig from each of the assemblies. The green graphs represent the coverage on the Celera contig (5-25kb), the blue lines represent the ALLORA assembly (21-41kb), and the red graphs represent the CLC assembly (28-48kb). The boxes on each plot represent the insertion sequence in each contig. As you can see, the gap has relatively even coverage when the raw whole genome sequencing data was mapped against it. 3 of the lines (1 shown, ecotype ID 9596) are part of the ~500 lines of A. thaliana that Monsanto is sequencing as part of the 1001 Arabidopsis Genome Project. This suggests that this region was accurately assembled, exists, and is conserved in the A. thaliana pangenome.

Summary and Discussion

Efficient genomic studies and applications with a single Random Shear BAC library

Before the availability of the Lucigen Random Shear BAC techniques, all plant and animal whole genome sequences were built on the conventional BAC libraries and their physical maps. Many genomic gaps remain in all of the sequenced plant and animal genomes. Random Shear cloning is a powerful tool for avoiding the bias of restriction digestion or partial digestion and minimizing genomic gaps. Lucigen has provided >150 Random Shear BAC libraries for our international customers and partners in order to efficiently and effectively study plant and animal genomes and their genomic applications. The new technologies provide unparalleled opportunities for de novo next-gen sequencing too. A single Random Shear BAC library (10x) is better than multiple (3–5) partial digestion BAC libraries (30–50x). It will save your money, time and efforts.

Medicago Genome Sequencing also Shows only Random Shear BAC Library is Unbiased by Large-Scale BAC End Sequencing

Medicago Genome Sequencing (published Nature 2012) showed: Partial digestion BAC libraries are extremely biased, and genomic gaps exist even with 5 partial digestion BAC libraries with about 50x coverage in total. Random Shear BAC library is unbiased and it covers the genomic gaps of the conventional BAC libraries (Figure 6).

Figure 6. A: Significant variations of BAC coverage (BAC clone count per 100 kb genomic region) were observed in all partial digestion BAC libraries including met1, EcoRI, mth2 and mth4, both HindIII libraries (showed); in contrast random-shear BAC library (mtr) shows the least coverage variations across this Medicago chromosome 5 particularly and the entire Medicago genome in general.

B: Two examples of genomic gaps covered by a random shear BAC library (<5x). A lot of contig ends are dead ends due to absence of clones in the very large collection of partial digestion BAC libraries (49 5x coverage and 5 libraries) .25 x HindIII BAC library, a 5x EcoRI BAC library, a 4 x BamHI library, a 14x large insert HindIII library, and a 3.5 x random shear found library in addition.

Acknowledgements

This work was partially supported by a NIH SBR grant to CCW.