

Random Shear BAC Resource & Whole Genome BAC End Sequencing of *Fusarium virguliforme*

Cheng-Cang Wu¹, Bhattacharyya Madan K², Rosa Ye¹, Megan Wagner¹, Svetlana Jasinovica¹, Hargeet K. Brar², Nick Hermersmann¹, Xiaohu Huang², David A. Mead¹

¹Lucigen Corporation, Middleton, WI 53562

²Department of Agronomy, Iowa State University, IA 50011

Abstract

We report the construction and BAC end-sequencing of a Random Shear BAC library of *Fusarium virguliforme* that causes sudden death syndrome (SDS) in soybean. The library consists of 6,144 clones with an average insert size of about 150 kb. The clone coverage of this library is equivalent to 12 × haploid genomes. To facilitate the whole genome sequence and assembly of *Fusarium virguliforme*, we also performed BAC-end sequencing of 3,072 clones from the first 8 plates of the Random Shear BAC library, which represents over six genome coverage. These BACs will be placed against the genome sequence, which means we will be able to get BAC clones for most genomic regions by checking the G-browser. The sequenced Random Shear BAC library of *Fusarium virguliforme* was also distributed to researchers at several institutions and is currently available at Lucigen (www.Lucigen.com).

Background

Unbiased paired-end sequences are critical for whole genome assembly with the short reads of next-gen sequencing. However, the small insert size (average 40 kb) of fosmids or cosmids is less useful for whole genome assembly than larger-insert bacterial artificial chromosomes (BACs). BAC libraries built with conventional vectors and methods are inherently biased, resulting in genome gaps in all complex genomes (Table 1). We have developed techniques to construct unbiased, randomly-sheared BAC libraries with large inserts (>100 kb) as well as a unique transcription free BAC vector. These new tools help accelerate whole genome sequence and assembly using the short reads of high throughput next-gen sequencing data.

The fungal pathogen *Fusarium virguliforme*, the causative agent of soybean sudden death syndrome (SDS), has caused significant yield losses in soybean since it was first noted in 1980. *Fusarium* fungi also include many related pathogens of the world's major food crops. However, the sequence of this economically important fungal genome is not available to the scientific community.

Table 1. Gaps in Whole Genome Physical or Sequencing Maps

Species	Ref.	Genome Size (Mb)	# Libraries (coverage)	Contigs (chr. no.)	Genome Gaps
Plants					
<i>Arabidopsis</i>	1	125	Two (17x)	27 (5)	< 5%
Rice	2	430	Two (26x)	284 (12)	< 10%
Soybean	3	1,115	Three (10x)	2,905 (20)	~ 10%
Maize	4	2,500	Three (15x)	3,488 (10)	unknown
Animals					
Fruit Fly	5	97	One (14x)	9 (2)*	> 2%
Human	6	3,200	Five (15x)	246 (23)	~ 4%
Mouse	7	3,200	Two (33x)	296	~10%

References: *Mozo (1999); *Chan (2001); 2Wu (2004); *www.genome.arizona.edu/*Hoskins (2000); *HAGRI (2001); *Gregory(2002). Grosshigh physical maps of chromosome 2, 3.

Methods and Results

Fusarium virguliforme was grown under standard conditions, collected and stored at -80°C. High molecular weight (HMW) genomic DNA was purified in agarose plugs and the DNA was randomly sheared and cloned in a pSMART-BAC vector (Lucigen). Random Shear BAC library of *F. virguliforme* consists of 6,144 clones with average inserts of 150 kb (Figure 4), which equals ~12x genome coverage. To facilitate the whole genome sequencing and assembly of *F. virguliforme*, we also performed BAC-end sequencing of 3,072 BAC clones from both ends in the first 8 plates of the Random Shear BAC library that represent over six genomes. Of 6,144 BAC end sequence reads, 877 reads are 50 bp or shorter. After removing the short reads and trimming the vector-sequences, we obtained 5,062 paired-end sequences from 2,531 BACs and 205 single BAC end sequences; or 85.7% was successful. The length of total BAC end sequences is 2,991,403bp.

Transcription-free BAC/FOS Vector

An optimized BAC cloning system: Lucigen has developed an optimized BAC cloning system including the transcription-free pSMART BAC vector, sacB gene in the stuffer to select against background (Figure 1), and the CopyRight BAC induction host system.

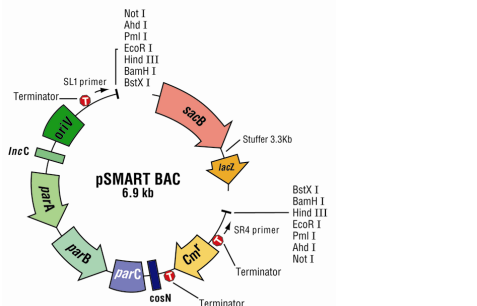


Figure 1. pSMART BAC vector. *ori2*, *repE*, *IncC* - origin of replication (single copy); *oriV* - inducible origin of replication; *parA*, *parB*, *parC* - partition genes; *Cm^r* - chloramphenicol resistance gene; *cosN* - lambda packaging signal; T - CloneSmart transcription terminators; *sacB*, sucrose gene; *lacZ*, alpha peptide portion of the beta galactosidase gene. Approximate positions of sequencing primers and transcription terminators are indicated.

Random Shearing of Genomic DNA

Megabase regions of genomic DNA, such as centromeres, may completely lack recognition sites for common restriction enzymes (e.g., BamHI, EcoRI, HindIII; Figure 2, left panel). Lucigen has developed methods to randomly shear genomic DNA into fragments of 100-400 kb. Significantly, the DNA from all genomic regions is sheared (Figure 2, right panel), which allows it to be cloned into BAC vectors.

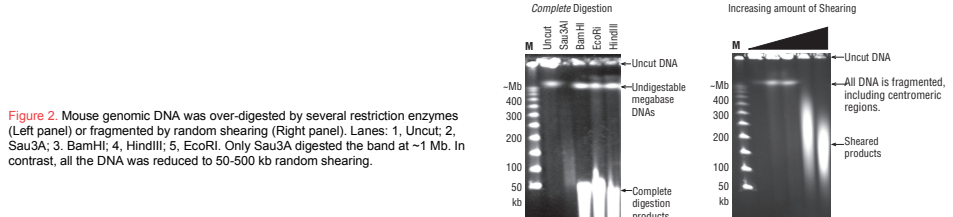


Figure 2. Mouse genomic DNA was over-digested by several restriction enzymes (Left panel) or fragmented by random shearing (Right panel). Lanes: 1, Uncut; 2, Sau3A; 3, BamHI; 4, HindIII; 5, EcoRI. Only Sau3A digested the band at ~1 Mb. In contrast, all the DNA was reduced to 50-500 kb random shearing.

Unbiased Cloning in Random Shear Libraries

The "complete" BAC library of the *Arabidopsis* genome contains numerous regions that are under- or over-represented (Figure 3, black bar graph). To show the unbiased distribution of clones in a random shear BAC library, *Arabidopsis* genomic DNA was randomly sheared, size-selected, and cloned into the pSMART BAC vector. A 5X coverage library was screened with overgo oligonucleotide probes specific for various regions of Chromosome 1. Significantly, clone coverage across all the probed regions, including the centromeric region, was similar in the random shear library (Figure 3). In contrast, these regions show vastly different representation in the *Arabidopsis* genome project (15, 75, or <1 clone per 0.1 Mb, respectively; 17X coverage overall). Most importantly, we have been able to close existing centromeric gaps of this "finished" physical and sequenced genomic map. The same probes also identified clones covering centromeric regions of other chromosomes.

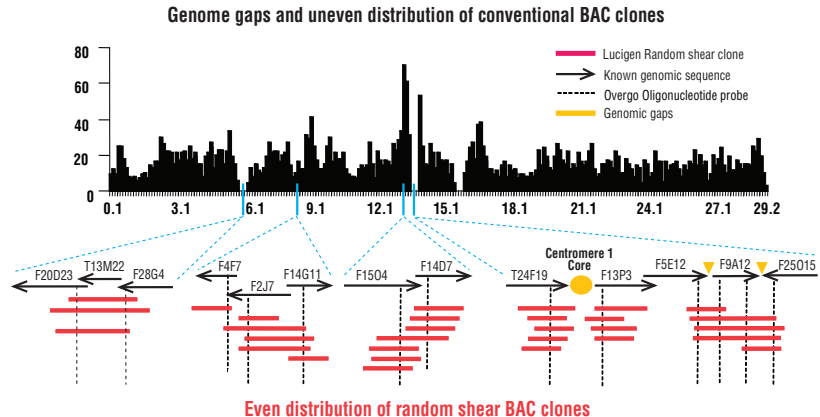


Figure 3. The distribution of BAC clones from Chromosome 1 of the *Arabidopsis* genome project is shown in the bar graph (Mozo, 1999). Overgo oligonucleotide probes were used to screen Lucigen's random shear library. The coverage of Lucigen clones is uniform over all regions tested. Several clone gaps were covered with this library, including centromeric regions.

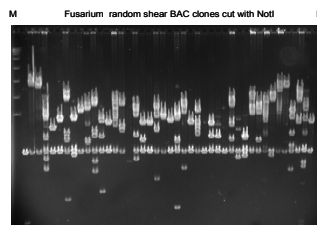


Figure 4. Genomic DNA was isolated from *Fusarium virguliforme*, randomly sheared, size-selected to >100 kb, and cloned into the pSMART-BAC vector. DNA from minipreps was digested with NotI to excise inserts. The pSMART-BAC vector band is visible at 7 kb.

Conclusions

We have successfully constructed an unbiased Random Shear BAC library of *F. virguliforme* with very large average insert size (~150 kb) and completed BAC end sequencing of the library. The BAC end sequences (2,991 Mb) is about 8.3% of the genome size (36 Mb) and about 6x physical genome coverage. These BACs will be placed against the whole shotgun genome sequence by the next-gen sequencing; which means we will be able to get BAC clones for most genomic regions by checking the G-browser. The sequenced Random Shear BAC library of *F. virguliforme* was also distributed to researchers of several institutions and is currently available at Lucigen (www.Lucigen.com). The Future work will characterize the BAC end sequences in more details, study whole genome assembly with next-gen sequencing data, and align the BAC end sequences with other *Fusarium* reference genomes and existing sequence information.

Lucigen®

Advanced Products for Molecular Biology

2120 W. Greenview Drive Middleton, WI 53562 888.575.9695

www.lucigen.com