

# First-Ever Soil Metagenomic Shuttle BAC Library Constructed with >100kb Inserts

Cheng-Cang Wu<sup>1</sup>, Rui R. Niedfeldt<sup>2</sup>, Rosa Ye<sup>1</sup>, Svetlana Jasinovica<sup>1</sup>, Megan Wagner<sup>1</sup>, Richard H. Ebright<sup>2</sup>, Mark R. Liles<sup>3</sup>, David A. Mead<sup>1</sup>

<sup>1</sup>Lucigen Corporation, Middleton, WI 53562

<sup>2</sup>HHMI, Waksman Institute, Rutgers University, Piscataway, NJ 08854

<sup>3</sup>Department of Biol. Sci., Auburn University, Auburn, AL 36849

## Abstract

Some of the major challenges in early stages of drug discovery by metagenomics are to: develop methods to capture the heterogeneity of complex environmental microbial communities (e.g. from soil and marine); clone much larger genomic fragments that can encompass entire biosynthetic pathways; and express a greater percentage of cloned recombinant DNA for identification of useful bioactive products. We have developed a bacterial artificial chromosome system (BAC) as well as new methods to routinely generate random shear BAC libraries with average insert sizes ~ 100 kb. These novel technologies eliminate the bottlenecks of conventional fosmid libraries, such as the need for expensive packaging reagents and an average insert size of 40 kb or smaller. The BAC system also alleviates problems associated with conventional libraries, such as bias caused by partial restriction digestion as well as small insert sizes. The combination of random shear cloning, a novel shuttle BAC vector, and methods for isolation of highly purified high molecular weight environmental DNA will result in superior metagenomic libraries. In addition, innovative screening methods will significantly enhance the likelihood of identifying new biochemical entities from these libraries.

## Background

Bacterial artificial chromosomes (BAC) vectors can be used to clone intact genetic pathways for the synthesis of secondary metabolites, as well as heterologous expression of novel natural products. However, the small insert-size (average 40 kb) of fosmids or cosmids may not be able to cover entire gene pathways and 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 could help accelerate drug discovery from bacterial and fungal genomes and metagenomes.

Table 1. Gaps in Whole Genome Physical or Sequencing Maps

Species	Ref.	Genome Size (Mb)	# Libraries (coverage)	Contigs (chr. no.)	Genome Gaps
<b>Plants</b>					
<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
<b>Animals</b>					
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: <sup>1</sup>Mazo (1999); <sup>2</sup>Chen (2001); <sup>3</sup>Wu (2004); <sup>4</sup>www.genome.arizona.edu; <sup>5</sup>Holtzman (2000); <sup>6</sup>Gregory (2002); <sup>7</sup>Chromosome physical maps of chromosome 2, 3.

## 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 system.

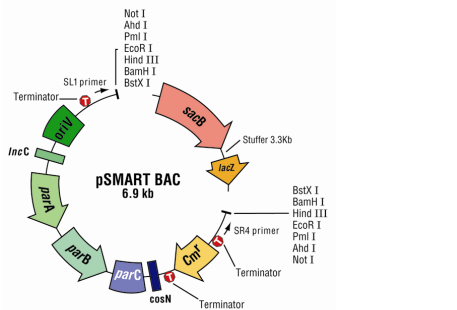


Figure 1. pSMART BAC vector. *ori2*, *repE*, *IncC* - origin of replication (single copy); *oriV* - inducible origin of replication; *parA,B,C* - partition genes; *Cm<sup>r</sup>* - chloramphenicol resistance gene; *cosN* - lambda packaging signal; T - CloneSmart transcription terminators; *sacB*, *sucrose* 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.

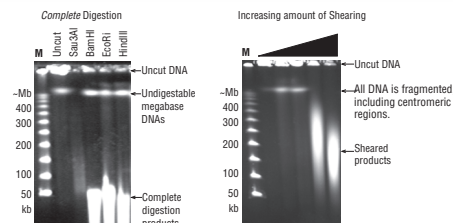


Figure 2. Mouse genomic DNA was over-digested by several restriction enzymes (Left panel) or fragmented by random shearing (Right panel). Lanes: 1, Uncut; 2, Sau3A; 3, BamHI; 4, HindIII; 5, EcoRI. Only Sau3A digested the band at ~1 Mb. In contrast, all the DNA was reduced to 50-500 kb random shearing.

## 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.

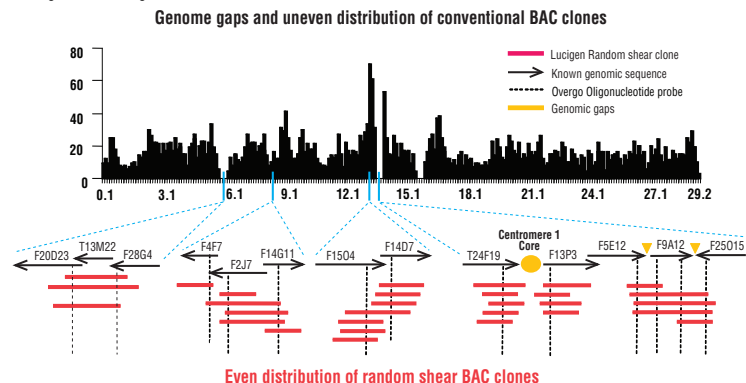


Figure 3. The distribution of BAC clones from Chromosome 1 of the *Arabidopsis* genome project is shown in the bar graph (Mozz, 1999). Overgo oligonucleotide probes were used to screen Lucigen's random shear library. The coverage of Lucigen clones is uniform over all regions tested. Several clone gaps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

## Methods and Results

Two antibiotic-producing bacteria, *Actinoplanes deccanensis* ATCC 21983 and *Dactylosporangium aurantiacum* NRRL 18085, and one antibiotic-producing soil sample were used. The bacteria were grown under standard conditions. The bacterial cells and soil sample were collected and stored at -80°C. High molecular weight (HMW) genomic DNA was purified in agarose DNA plugs and the DNA was randomly sheared and cloned in a pSMART-BAC vector (Lucigen). Random Shear BAC libraries of *A. deccanensis* and *D. aurantiacum* were successfully constructed. Both BAC libraries consist of 1,920 clones with average inserts of 100 kb, which equals ~50x genome coverage. HMW genomic DNA from the soil sample was also successfully cloned into a shuttle pSMART-BACs vector (Lucigen). The soil Random Shear Shuttle BAC library consists of 19,200 clones with average insert size at least 100kb.

## Conclusions

The results show that large genomic regions (average 100 kb) including complete gene clusters/pathways can be randomly cloned without restriction partial digestion in both common and shuttle BAC vectors for the discovery of novel natural products. Future work will characterize the BAC and shuttle BAC clones containing complete gene pathways and their functional studies. The preparation of high quality and quantity of HMW genomic DNA from soil samples is extremely challenging. Previous published attempts to construct soil BAC libraries with average insert size 100kb or larger have not been successful. Random Shear BAC cloning of soil metagenomic libraries with average 100 kb or larger is presented here for the first. The novel Random Shear BAC techniques could be applied to many other environmental samples, such as algae, fungi, marine sediments, sponges, etc. This opens a brand new possibility of drug discovery and production.

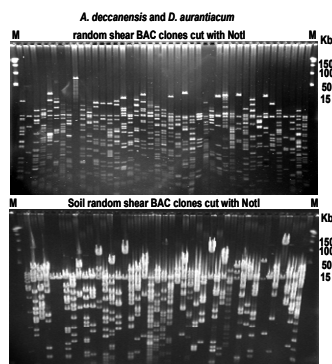


Figure 4. Genomic DNA was isolated from bacteria or soil, randomly sheared, size-selected to >100 kb, and cloned into the pSMART-BAC or shuttle pSMART-BACs vector. DNA from minipreps was digested with NotI to excise inserts. The vector band is visible at 7 kb of pSMART-BAC or 13kb of pSMART\_BACs.

Funding : NIH SBIR Phase I  
PI: CC. Wu  
Grant Number: 1R43AI085840-01

**Lucigen**<sup>®</sup>  
Advanced Products for Molecular Biology  
2120 W. Greenview Drive Middleton, WI 53562 888.575.9695  
[www.lucigen.com](http://www.lucigen.com)