

Random Shear BAC Library Construction (Lucigen Corporation White Paper)

The quality of genomic BAC libraries depends greatly on the cloning methods and vectors used. The bias introduced in constructing BAC libraries by partial restriction digestion of genomic DNA results in various regions being highly under-represented, over-represented, or missing for all eukaryotic multi-cellular genomes studied, including *Arabidopsis*, *Drosophila*, rice, mouse, and human. This bias results in numerous clone gaps that can be impossible to close, even with multiple complementary libraries and up to 40X coverage, thus dramatically increasing the finishing costs.

To circumvent these problems, Lucigen has successfully developed novel DNA random shear techniques and blunt end ligation of large inserts (>100 kb) to efficiently construct unbiased, full coverage BAC libraries. With these new techniques, we have demonstrated that a Lucigen Random Shear BAC Library of the *Arabidopsis* genome is unbiased. Importantly, with just 5X coverage we were able to close several clone gaps and fill existing centromeric gaps in the *Arabidopsis thaliana* genome. We have also developed a more efficient BAC cloning system including high stability, transcription-free BAC cloning vectors and BAC-optimized competent cells to alleviate instability problems with current BAC cloning systems. These new techniques and tools simplify BAC DNA preparation and sequencing.

Random shearing of all genomic DNA, including undigestible DNA.

Certain regions of genomic DNA, such as centromeres, highly repetitive sequences, and telomeres, often contain megabase-size regions that completely lack recognition sites for common restriction enzymes (e.g., BamHI, EcoRI, and HindIII; Figure 1, left panel). Lucigen has developed methods to randomly shear genomic DNA into fragments of 100-400 kb to successfully clone this "undigestible" DNA. Significantly, the DNA from all genomic regions is sheared (Figure 1, right panel), eliminating this bias inherent in conventional methods.

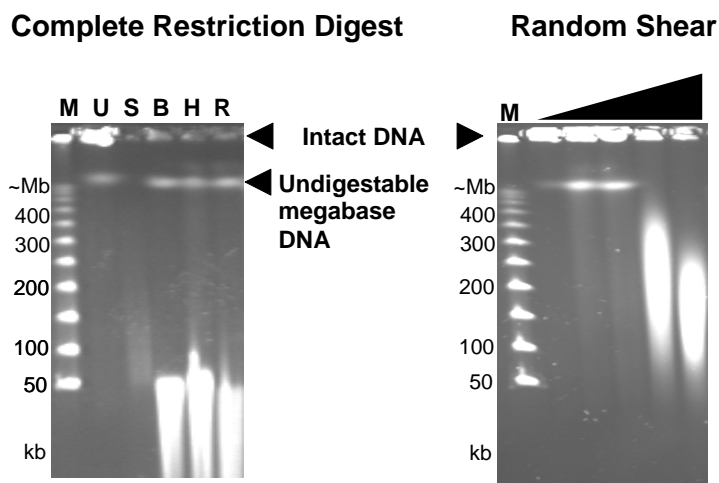


Figure 1. Mouse genomic DNA was digested to completion by several restriction enzymes (Left panel) or fragmented by random shearing (Right panel). Lanes: M, Marker; U, Uncut; S, Sau3A; B, BamHI; H, HindIII; R, EcoRI. Only Sau3A digested the band at ~1 Mb. In contrast, all of the DNA was reduced to 100-400 kb as the degree of random shearing was increased.

A second major obstacle to constructing random shear BAC libraries is the ability to clone large insert blunt end DNA efficiently and reproducibly. The data shown below demonstrates this technical hurdle has been overcome at Lucigen.

>100 Kb Inserts in a Random Shear BAC Library of Arabidopsis

To demonstrate that randomly sheared genomic DNA is suitable for BAC library construction, a Random Shear BAC Library of *Arabidopsis* genomic DNA was constructed and cloned into the transcription-free pSMART[®] BAC vector. The average insert size was > 100 kb (Figure 2), confirming that large fragments were generated by random shearing and that they could be efficiently cloned.

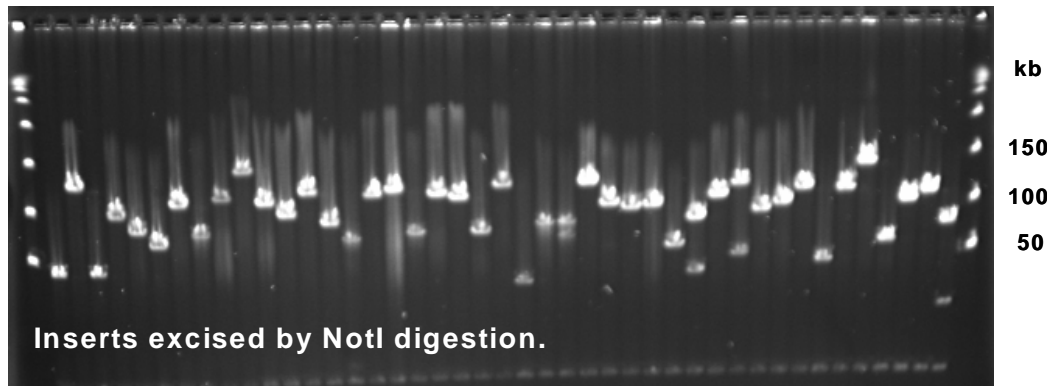


Figure 2. Genomic DNA was isolated from *Arabidopsis* tissue, randomly sheared, size-selected to >100 kb, and cloned into the pSMART BAC vector. DNA from minipreps was digested with NotI to excise inserts. The vector band is visible at 7 kb.

Unbiased Cloning in Random Shear Libraries

The published *Arabidopsis* genome has been declared “finished”, with 17X genomic coverage in the form of BAC clones. However, numerous regions of the genome are under-represented, over-represented, or completely missing (Figure 3, black bar graph; Mozo, 1999). To determine the efficacy of Random Shear Library construction, an *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 was uniform across all the probed regions, including the centromeric region, in the Random Shear Library (Figure 3, red bars). In contrast, these regions show vastly different representation in the *Arabidopsis* genome project (15, 75, or <1 clone per 0.1 Mb, respectively). Most importantly, the random shear technology closed several centromeric gaps that were present in this “finished” physical and sequence genomic map. The same probes also identified clones covering centromeric regions of other chromosomes. These results demonstrate that Random Shear BAC Libraries promise to be a very effective way to finish existing genome sequencing projects quickly and economically. The even representation of genomic fragments in a random shear BAC library will significantly reduce the number of clones that will need to be processed for any type of large scale analysis.

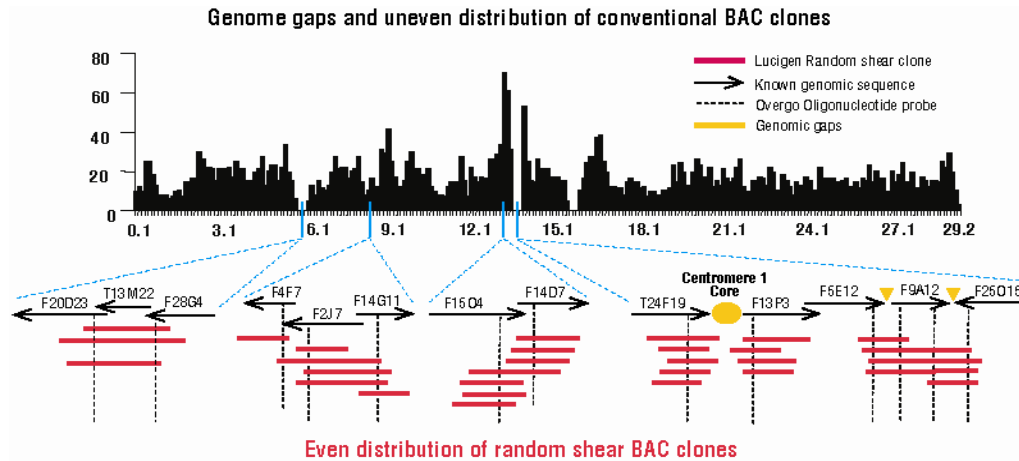


Figure 3. The distribution of BAC clones from Chromosome 1 of the Arabidopsis genome project is shown in the black bar graph (Mozo, 1999). Overgo oligonucleotide probes were used to screen Lucigen’s Random Shear BAC Library. The coverage of Lucigen clones is uniform over all regions tested. Several clone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

Dramatic Cost Savings with Random Shear Libraries

Random Shear BAC Libraries are the ideal choice for initial library construction in new genome projects. Due to the unbiased coverage, only a single Random Shear Library needs to be constructed, instead of the multiple restriction-digestion libraries now needed. Importantly, with a Random Shear BAC Library the finishing process is greatly simplified because of the dramatically reduced number of gaps. Random Shear BAC Libraries provide superior data with less time, effort, and expense (Table 1).

Table 1. Comparison of Random Shear and partial digestion BAC libraries

BAC library	Character	Number of Libraries Needed (coverage)	Cost of Library Construction (\$US) *	Finishing cost (\$US) **
Random Shear	Unbiased; No gaps	1 (10X)	Up to \$30,000	Up to \$1 million
Partial Digestion	Biased; gaps	≥ 2 (>20X)	≥ \$50,000	≥ \$3 million

*Construction cost is based on a genome of ~500 Mb, requiring one Random Shear BAC Library (10X) and at least two complementary partial-digest BAC libraries (20X total), to minimize restriction site bias.

**Finishing cost includes BAC end sequencing, whole genome physical mapping, BAC library screening, and integrating the physical map with about 1000 genetic makers.

CopyRight® v2.0 BAC Cloning System for constructing unbiased BAC libraries

The CopyRight v2.0 BAC cloning system includes the pSMART® BAC vector, incorporating transcription-free technology for the highest stability of BAC clones possible (1). The pSMART® BAC vector also has a lacZ/sacB stuffer region, allowing uncut vector to be detected by blue/white screening and selected against by plating on sucrose (Figure 4). This feature completely eliminates the background of uncut vector, improving the efficiency of colony picking. However, unlike other cloning vectors, the promoter as well as the coding sequence of the stuffer fragment is completely removed during processing. This design prevents active expression of the insert DNA by the lacZ or sacB promoter, contributing greatly to plasmid stability, especially for inserts

containing toxic coding sequences, strong secondary structure, or other deleterious features.

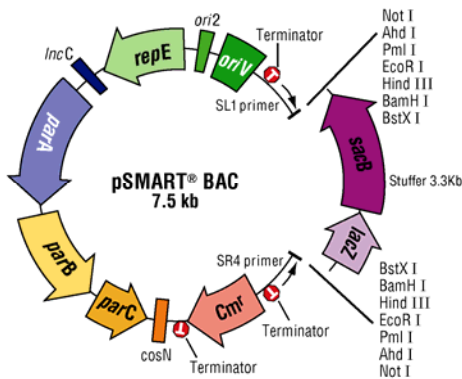


Figure 4. pSMART BAC vector. ori2, repE, IncC - origin of replication (single copy); oriV - inducible origin of replication; par A,B,C- partition genes; Cmr- chloramphenicol resistance gene; cosN - lambda packaging signal; T – CloneSmart transcription terminators; sacB, sucrose gene; lacZ, alpha peptide portion of the beta galactosidase gene. Approximate positions of sequencing primers and transcription terminators (T) are indicated.

CopyRight® v2.0 BAC Cloning System simplifies BAC preparation and sequencing

Lucigen’s CopyRight v2.0 BAC Cloning System consists of the pSMART® BAC vector and BAC-Optimized Replicator™ v2.0 Electrocompetent Cells (2). The pSMART BAC vector contains both the single-copy origin of replication from the *E. coli* F-factor and the inducible high-copy oriV origin of replication. BAC-Optimized Replicator™ v2.0 Cells are an *E. coli* strain that contains the *trfA* gene for initiation of replication from oriV. The *trfA* gene is under tight control of the inducible araBAD promoter. Upon addition of the induction solution arabinose, CopyRight BAC clones replicate to ~10-20 copies/cell, providing high yields of high purity DNA. Clone stability is not affected by induction (Figure 5). This feature simplifies BAC DNA preparation and increases the success rate of BAC end sequencing to >95%.

NotI digestion of BAC clones

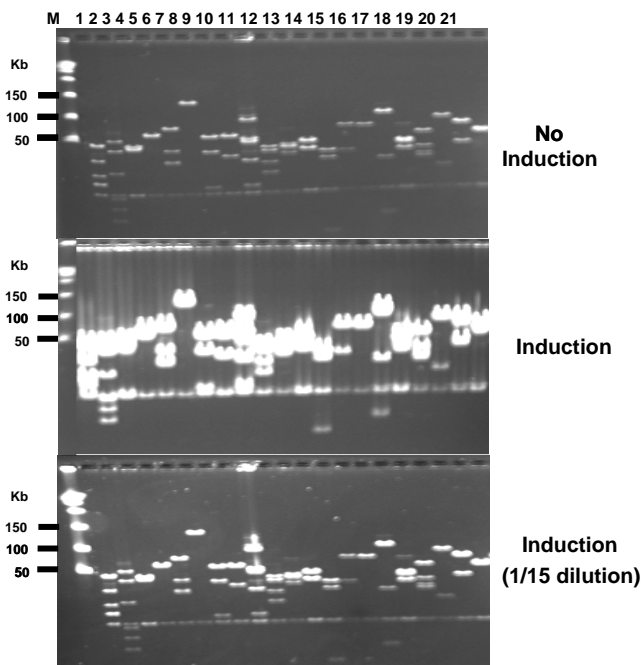


Figure 5. Copy-number amplification of large insert CopyRight BACs. BAC DNA was prepared according to a standard protocol, and 10 µl of each clone was digested with NotI. Insert sizes are 50 to ~200 kb.

M, lambda ladder DNA marker. Upper panel: BAC samples without induction; middle panel: BAC samples with induction; lower panel: BAC samples with induction, diluted 15-fold.

Unrivalled BAC expertise

Lucigen's BAC library construction team is led by Dr. Chengcang Wu, who has more than ten years of experience in large-insert DNA and BAC cloning in industrial and academic labs, including: Pioneer Hi-Bred International, Inc.; Texas A&M University; and the National Institute of Sericultural and Entomological Sciences (NISES, now the National Institute of Agrobiological Sciences, NIAS), Japan. He successfully constructed the first plant artificial chromosomes *de novo* in maize at Pioneer Hi-Bred International, Inc. The patent on this work issued recently (3). Dr. Wu has published peer-reviewed papers in large-insert DNA cloning and co-authored several reviews on BAC cloning and related research (4-11).

Dr. Wu has constructed more than 100 BAC libraries during his career, and more than 35 Random Shear BAC Libraries at Lucigen in the last 20 months (Figure 6 and 7).

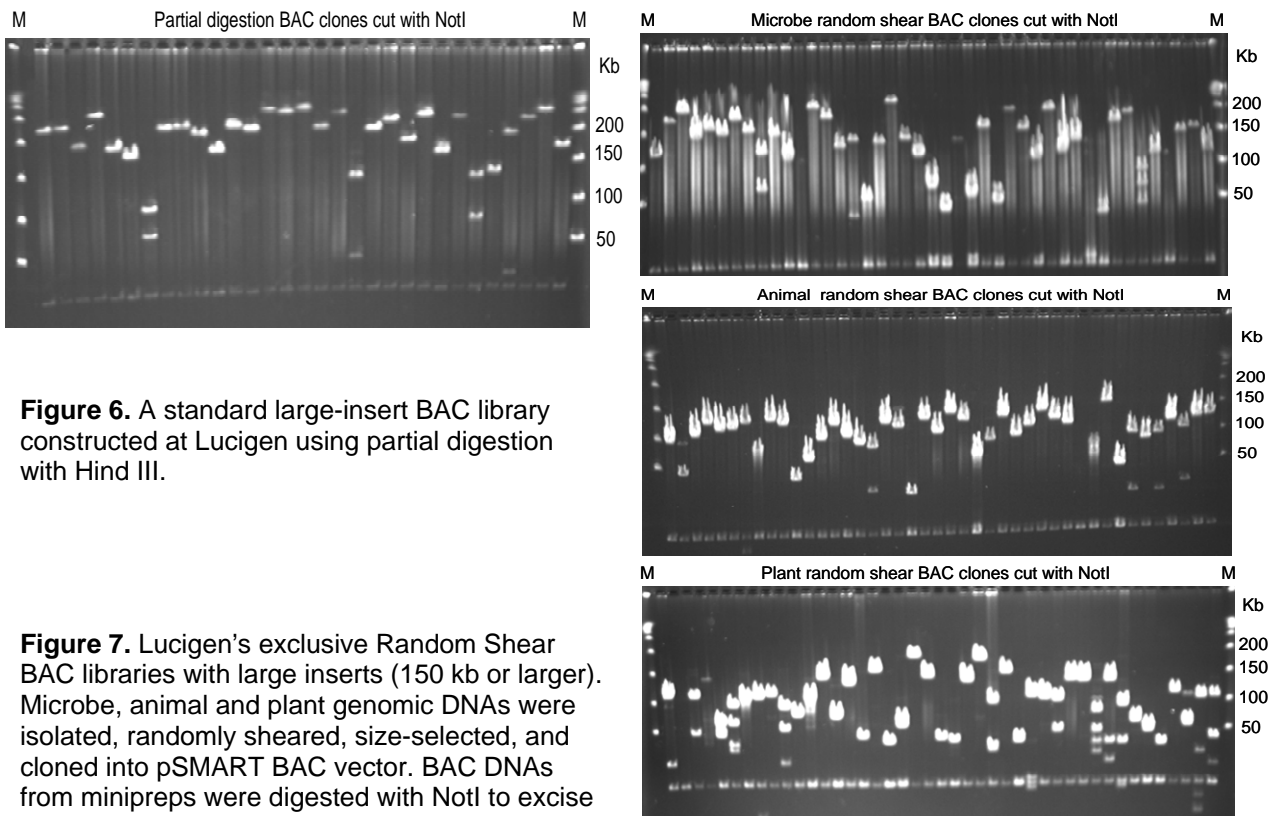


Figure 6. A standard large-insert BAC library constructed at Lucigen using partial digestion with Hind III.

Figure 7. Lucigen's exclusive Random Shear BAC libraries with large inserts (150 kb or larger). Microbe, animal and plant genomic DNAs were isolated, randomly sheared, size-selected, and cloned into pSMART BAC vector. BAC DNAs from minipreps were digested with NotI to excise inserts. The vector band is visible at 7kb.

Random shear BAC Libraries completed recently:

The Lucigen BAC team successfully completed more than 35 Random Shear BAC Libraries, including the most important models: **mouse**, ***Xenopus Tropicalis***, and ***Medicago truncatula***, as well as a library from a **Crustacean species** that was impossible to construct using partial digestion. See the example Random Shear BAC projects listed at Table 2.

Table 2. Random Shear BAC projects done recently at Lucigen (07/08).

Organisms	Average insert size (kb)	Institute
mouse	170~200	Harvard Univ.
Duckweed	100	private firm
Arabidopsis	105	Lucigen
Peanut	120	UC Davis
Potato	118	Univ. Wisc. at Madison
Pineapple	80	UIUC
<i>Cajanus cajan</i>	90	UCdavis
<i>Medicago truncatula</i>	105	LIPM INRA-CNRS, Fr.
Apricot	130	Bologna Univ. Italy
mouse	150	privet firm
Halibut (fish)	103	NRC-IMB, Canada
<i>Xenopus tropicalis</i>	115	JGI-Stanford-Berkeley
<i>Triops longicaudatus</i>	100	UNLV
Chinese Cabeagge	110	cnu.ac.kr
Channel catfish	150	USDA-ARS, MS
Palm	130	private firm
Barley	150	ipk-gatersleben.de
Rice1	125	CAAS
Fungi1	150	private firm
Fungi2	155	private firm
Grape 1	110	MissouriSt. Univ
Grape 2	100	Inst. Gen. Appl. Italy
Human	165	private firm
Fungi (<i>Bremia lactucae</i>)	150	UC Davis
Cotton1	110	UT Austin
Rice2	120	USDA, CA

Application examples of Random Shear BAC projects:

- I. **Genome gap closing:** The international plant and animal societies have started trying to close these genomic gaps by utilizing additional Lucigen Random Shear BAC libraries, such as: Arabidopsis, rice, barley, grape, Chinese cabbage, *medicago*, soybean, potato, and tomato; mouse, *Xenopus Tropicalis*, and catfish.

II. Efficient genomic studies and applications with a single Random Shear BAC library: For example,

- 1) whole genome sequencing of 2 energy-plant species, we have made one single Random Shear BAC library for both projects from our collaborators at Synthetic Genomics Inc;
- 2) Studying of the interaction of pathogens and animal/plant hosts;
- 3) Discoveries of pharmaceutical compounds and drugs;
- 4) Positional cloning of QTLs, transgene loci in animals and plants, etc;
- 5) Enzyme discoveries for biofuels;
- 6) Analysis of metabolic pathways;
- 7) Multiple gene stacking and transformation applications.

References:

1. R. Godiska, et al. (2004) Beyond pUC: Vectors for cloning unstable DNAs. In Optimization of the DNA sequencing process, Jones and Bartlett Publishers, Boston.
2. J., Wild, Hradecna Z, Szybalski W. (2002) Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Res.* 12:1434-1444.
3. CC. Wu, as one of the co-inventors (2007) Artificial Plant Minichromosomes. US20070271629
4. X. Zhang, Scheuring C, Tripathy S, Xu Z, Wu CC, Ko A, Tian SK, Arredond F, Lee M-K, Santos AF, Zhang H-B, Tyler BM. (2006) An integrated BAC and genome sequence physical map of *Phytophthora sojae*. *Molecular Plant-Microbe Interactions.* 12:1302-1310.
5. G.A.L. Broggin, B. Le Cam, L. Parisi, Wu, CC., H.-B. Zhang, C. Gessler and A. Patocchi (2007) Construction of a contig of BAC clones spanning the region of the apple scab avirulence gene AvrVg. *Fungal Genetics and Biology.* 44: 44-51
6. RM. Stupar, Beaubien, K.A., Jin, W., Song, J., Lee, M-K., Wu, CC., Zhang, H-B., Han, B., Jiang, J. (2006) Structural diversity and differential transcription of the patatin multicopy gene family during potato tuber development. *Genetics.* 172:1263-75.
7. C. Ren, Sun, S., Lee, M-K., Xu, Z., Wu, CC. and Zhang, H-B. (2005) Chapter 4. Genomic libraries and physical mapping (YACs and BACs). In, *Handbook of Plant Genome Mapping*, pp173-214. Edited by Meksem K. and Gresshoff P., Wiley-VCH Verlag GmbH: Weinheim, Germany.
8. CC. Wu, Xu, Z. and Zhang, HB (2004) DNA libraries. In, *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, Volume 3 (2nd Edition), pp385-425. Edited by Meyers R.A., Wiley-VCH Verlag GmbH: Weinheim, Germany
9. CC. Wu, Santos, FA., Nimmakayala, P., Springman, R., Meksem, K., Lightfoot, D., and Zhang, H-B. (2004) Construction of a complementary BAC library for soybean genome-wide physical mapping and whole genome sequencing. *Theor. Appl. Genet.* 109, 1041-1050.

10. H-B. Zhang and Wu CC. (2001) BAC as tools for genome sequencing. *Plant Physiol. Biochem.* 39, 1-15
11. CC. Wu, Asakawa S, Shimizu N, Kawasaki S, and Yasukochi Y.(1999) Construction and characterization of bacterial artificial chromosome libraries from the silkworm, *Bombyx mori*. *Mol Gen Genet.* 261:698-706.