

ABSTRACT

Next generation sequencing technologies can rapidly and economically produce a draft genome of an organism de novo. However, the draft data is seldom more than 80% complete with >10e5 contigs for large genomes, which is insufficient for many applications. More importantly no hands-on genomic resources (e.g. sequenced BACs, BAC physical maps) are produced for functional genomics after sequencing. Sequence data that is closer to 95% finished with the unambiguous order and placement of genes would have the greatest utility for scientific and commercial research. Tools that bridge the gaps between massively parallel short read sequencing (35-500 bases) and the need for large scaffolds to accurately assemble complex genomes (>100,000 bases) are needed. Lucigen has successfully developed new tools and methods for the construction of large insert random shear BAC libraries. This advance allows the production of whole genome libraries that are unbiased by the non-random distribution of restriction enzyme sites, which significantly reduces the number of clones needed to finish a genome while eliminating gaps due to sequence bias. Lucigen has constructed more than 100 random shear BAC libraries of microbes, plants and animal species for researchers around the world. Combining random shear BAC library capabilities with next generation sequencing technologies should theoretically result in nearly complete coverage, assembly of even daunting genomes and useful genomic information and resources simultaneously. If successful, this approach could allow the rational sequencing and analysis of daunting genomes such as wheat and loblolly pine, enable complete coverage of complex metagenomes and simplify the resequencing of repeat rich regions of the human genome (such as the major histocompatibility complex). We will present a strategy for achieving this aim that seeks to provide a nearly finished genome while reducing computational complexity, maximizing the efficiency of the existing technologies and produce more useful de novo reference genomes.

BACKGROUND

Traditionally BAC libraries are constructed from partial digestions of genomic DNA. However, despite using multiple libraries, many gaps remain in all genomes studied (Table 1). These gaps include, but are not limited to, repetitive DNA and centromeric regions. The number of gaps and contigs is even worse with NGS strategies.

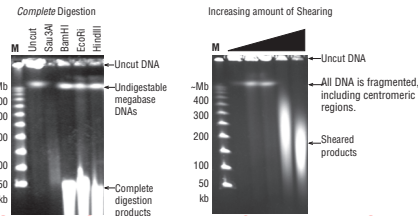
Table 1. Gaps in Whole Genome Physical or Sequencing Maps.

Species	Genome Size (Mb)	Number Libraries (coverage)	Contigs (chr. no.)	Genome Gaps
Plants				
Arabidopsis	125	Two (17X)	27 (5)	~ 5%
Rice	430	Two (26X)	284 (12)	~ 10%
Soybean	1115	Three (10X)	2905 (20)	~ 10%
Maize	2500	Three (15X)	3488 (10)	unknown
Animal				
Fruit Fly	97	One (14X)	9 (2)	~2%
Human	3200	Five (15X)	246 (23)	~4%
Mouse	3200	Two (33X)	296 (20)	~10%

RANDOM SHEARING vs RESTRICTION OF LARGE GENOMIC DNAs

Megabase regions of genomic DNA, such as centromeres, may completely lack recognition sites for common restriction enzymes (e.g., BamHI, EcoRI, HindIII; Figure 1, 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 1, right panel), which allows it to be cloned into BAC vectors or made into clone free NGS libraries.

Figure 1. 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 by random shearing.



UNBIASED CLONING USING RANDOM SHEAR BAC LIBRARIES

The "complete" BAC library of the *Arabidopsis* genome contains numerous regions that are under- or over-represented (Figure 2, 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 2). 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.

Genome gaps and uneven distribution of conventional BAC clones

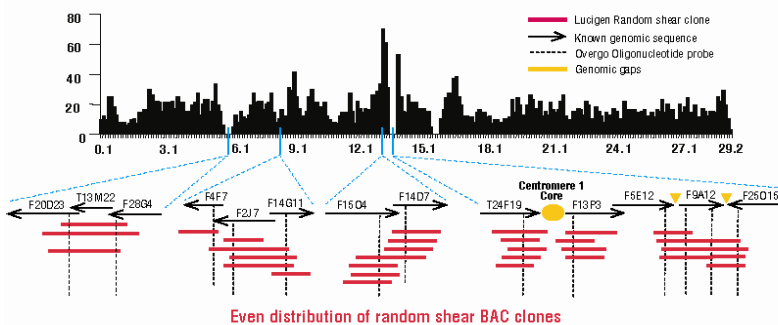


Figure 2. The distribution of BAC clones from Chromosome 1 of the *Arabidopsis* genome project is shown in the bar graph (Mozz, 1999). Overgo oligonucleotide probes were used to screen Lucigen's random shear library. The coverage of Lucigen clones is uniform over all regions tested. Several clone gaps were covered with this library, including centromeric regions.

BAC CLONES AS REDUCED REPRESENTATION LIBRARIES

The challenge of de novo sequencing with larger genomes is that assembly becomes difficult as repeat content increases, and many larger genomes, particularly those of crop plants, have significant repetitive content. These assembly challenges sometimes impact even traditional sequencing approaches, but are particularly problematic with short read technologies. Additionally, de novo assembly from WGS short-read sequencing currently requires large computational resources, on the order of hundreds of gigabytes of RAM, when scaled to larger genomes. BAC clones accurately partition genomes into more manageable subregions for sequencing and assembly, and the use of random shear BAC libraries significantly reduces the number of clones needed to cover a genome. In silico studies have investigated the cost-saving potential of using a randomized BAC clone library with short-read sequencing (Sundquist et al. 2007). Clones are randomly picked, with a 10X coverage more than sufficient for this process. A key part of the assembly methodology is where clone contigs are identified and ordered to form scaffolds. The sequencing protocol is a variant on the well-known hierarchical sequencing technique, but removes the time-consuming and manual selection of a tiling path in favor of a parallelizable, random selection strategy.

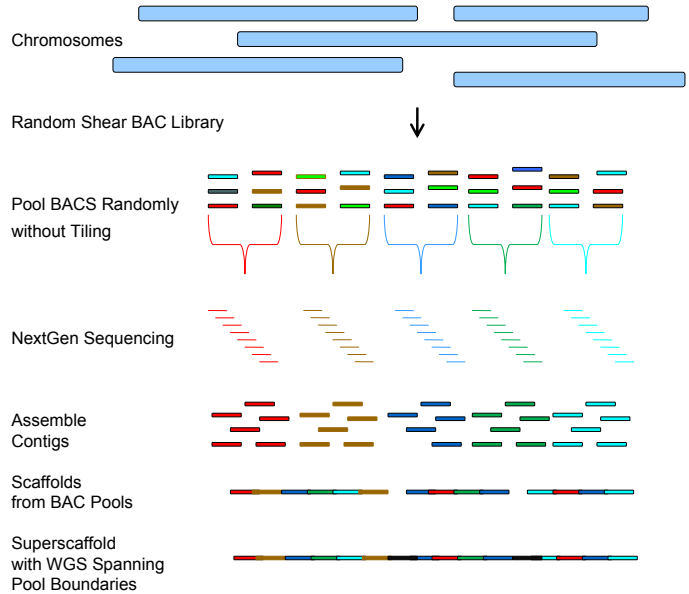


Figure 3. Random shear BAC library construction and pooled sequencing using next gen technology allows for the complete assembly of large genomes. With other strategies the sequence of each individual BAC can be deciphered.

FEATURES & BENEFITS OF RANDOM SHEAR DE NOVO GENOME SEQUENCING

BAC size inserts simplify scaffolding and increase accuracy of genome assembly.

"Finished-grade" whole genome assembly.

A single random shear 10X coverage sequenced BAC resource.

Sequence-based whole genome physical and sequence map.

Integration with clone-free approaches.

Individual BAC sequences can be deciphered with other strategies.

Deep and accurate sequence from selected genomic subregions that exhibit extensive segmental duplication or repeats.

SUMMARY

Lucigen's Genome Resource Center has constructed more than 100 Random Shear BAC libraries with large inserts for researchers around the world, arrayed a total of more than 2,500,000 clones from bacteria, fungi, algae, plants, and animal species, and provided affordable public access to this second-generation high-quality BAC library technology. The international plant and animal societies have started cloning genomic gaps by utilizing Random Shear BAC libraries made from *Arabidopsis*, rice, cotton, barley, grape, Chinese cabbage, *medicago*, soybean, potato, peanut, duckweed, pineapple, tomato, mouse, *Xenopus Tropicalis*, and catfish. Efficient whole genome sequencing of biomass/bioenergy plants from a single Random Shear BAC library has been also demonstrated recently (e.g., for oil palm and *Jatropha curcas*). These new technologies provide unparalleled opportunities for whole genome sequencing, complex traits/gene/pathway discovery, metabolic engineering and genetic improvement of microbes, plants and animals.

References

Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S. Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS One. 2007 May 30;2(5):e484.

