

Cot-Based Cloning and Sequencing (CBCS) in the advancement of comparative genomics

Daniel G. Peterson,¹ Susan R. Wessler,² and Andrew H. Paterson³

1. Department of Plant & Soil Sciences, Mississippi State University; 2. Plant Biology Department, University of Georgia; 3. Center for Applied Genetic Technologies, University of Georgia

SUMMARY

Cot-Based Cloning and Sequencing (CBCS), a synthesis of Cot analysis, DNA cloning, and high-throughput sequencing, promises to (1) permit efficient gene discovery in species with substantial quantities of repetitive DNA, (2) allow the sequence complexity (SqCx) (*i.e.*, all the unique sequence information) of large genomes to be elucidated at a fraction of the cost of shotgun sequencing, and (3) enhance genome sequencing efforts by facilitating capture of low-copy sequences not secured by EST sequencing. CBCS should make comparative genomics research more cost effective, and expedite detailed study of large genomes such as those of many crops.

PRINCIPLES OF CBCS

In CBCS a Cot (DNA renaturation kinetics) analysis is performed for a species of interest, the results of the Cot analysis are used to guide the hydroxyapatite chromatography-based fractionation of the genome into low-copy and repetitive sequence components, each isolated sequence component is used to construct a corresponding 'Cot library,' and clones from each library are sequenced in numbers proportional to the kinetic complexity (= estimated SqCx) of the component from which they were derived. The utility of CBCS is demonstrated in **FIGURE 1**.

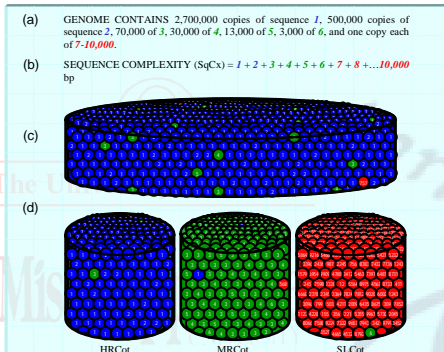


FIGURE 1. SqCx and sequencing. (a) The elements constituting a hypothetical eukaryotic genome. (b) Though repetitive sequences account for the majority of DNA, they contribute very little to SqCx. (c) The net gain in novel sequence information is slow and costly if clones are selected from an unbiased genomic library (shotgun approach). (d) CBCS permits the highly repetitive (HR), moderately repetitive (MR), and single/low-copy (SL) components of the genome to be separately isolated and cloned. Since almost all of the SqCx is contained within the SLcot library, most sequencing resources can be devoted to sequencing SLcot clones.

INITIAL RESEARCH

In an initial test of CBCS, the genome of sorghum (*Sorghum bicolor*) was fractionated into highly repetitive (HR), moderately repetitive (MR), and single/low-copy (SL) sequence components (**FIGURE 2**) that were consequently cloned to produce HRCot, MRCot, and SLcot genomic libraries. Filter hybridization (blotting) and sequence analysis both show that the HRCot library is enriched in sequences traditionally found in high-copy number (retroelements, rDNA, centromeric repeats), the SLcot library

is enriched in low-copy sequences (genes and ESTs), and the MRCot library contains sequences of moderate redundancy (**FIGURE 3**). Foldback (FB) DNA (**FIGURE 2**) was also isolated and cloned, although FB clones have yet to be sequenced (Peterson *et al.* 2002a).

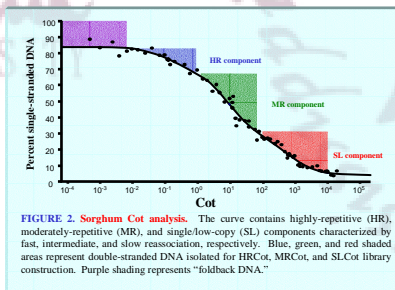


FIGURE 2. Sorghum Cot analysis. The curve contains highly-repetitive (HR), moderately-repetitive (MR), and single/low-copy (SL) components characterized by fast, intermediate, and slow reassociation, respectively. Blue, green, and red shaded areas represent double-stranded DNA isolated for HRCot, MRCot, and SLcot library construction. Purple shading represents "foldback DNA."

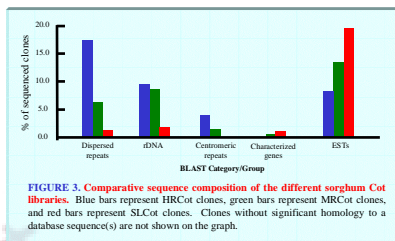


FIGURE 3. Comparative sequence composition of the different sorghum Cot libraries. Blue bars represent HRCot clones, green bars represent MRCot clones, and red bars represent SLcot clones. Clones without significant homology to a database sequence(s) are not shown on the graph.

DISCUSSION

CBCS in gene discovery

CBCS promises to resolve some of the difficulties that are presently associated with isolation of comprehensive sets of genes from large-genome species. EST (cDNA) sequencing is an economical first step in gene discovery, but only a fraction of the transcriptome is expressed in any single source tissue. Even by studying cDNA libraries from multiple tissues, diminishing returns typically accrue after about 10⁵ sequences, many genes expressed only rarely or at low levels are likely to be missed, and no information is obtained on regulatory sequences or other important low-copy elements. Unlike EST sequencing, CBCS provides access to regulatory sequences and also secures genes independently of their levels or (tissue or organ-specific) patterns of expression.

CBCS possesses a significant advantage over methylfiltration, a technique that has been suggested as an intermediate step between EST and genomic shotgun sequencing. Briefly, methylfiltration results in the production of genomic libraries enriched in hypomethylated (presumably genic) sequences. While this approach has merit, the pattern and significance of DNA methylation differs markedly between species, developmental stages, genes within an organism, and regions of a gene.

Consequently, exclusion of hypermethylated DNA is expected to result in the loss of important/interesting genes. Because hydroxyapatite-based fractionation of genomic DNA is independent of sequence methylation, CBCS should not result in the loss of any genes based upon their methylation status.

CBCS as a means to capture sequence complexity

For species with large, repetitive genomes, capture of SqCx should provide many of the benefits of complete genome sequencing at substantially reduced costs. At present, genomic shotgun sequencing is the main tool used to capture SqCx (usually within the context of a genome sequencing project), but CBCS offers a much more efficient method of sequence discovery (**FIGURE 1**). Using a shotgun approach, the number of different clones (n) that must be sequenced in order to have 99% confidence that all genomic elements have been sequenced at least once is estimated using the formula

$$[\text{Equation 1}] \quad n = \ln(1 - 0.99) \div [\ln(1 - (Z \div G))]$$

where Z = mean insert size in bp and G = 1C genome size in bp. In CBCS, sequencing resources are allocated based on the contribution of each kinetic component library to genomic SqCx. The probability of sequencing 99% of DNA elements using CBCS is therefore a function of the sum of the kinetic complexities (γ) of the different components. Because the kinetic complexity of the foldback (FB) fraction is unknown, the most conservative means to assure capture of all cloned FB sequences is to assign the FB fraction a 'kinetic complexity' equal to the number of base pairs it contains. For a genome composed of the components a, b, and c with f/b bp of foldback DNA

$$[\text{Equation 2}] \quad n = \ln(1 - 0.99) \div [\ln(1 - (Z \div (\gamma_a + \gamma_b + \gamma_c + f)))]$$

CBCS reduces by two-thirds or more the number of clones that need to be sequenced to capture the SqCx of many eukaryotic genomes (Peterson *et al.* 2002b). For some plant species, CBCS could save tens to hundreds of millions of dollars in sequencing expenses (**TABLE 1**).

TABLE 1. Estimated savings if CBCS rather than "shotgun sequencing" is used to capture a genome's SqCx.*				
SPECIES	METHOD	CLONES	COST	
Sugarcane	Shotgun	280 million	\$955,000,000	
	CBCS	59 million	\$281,000,000	\$674,000,000 SAVINGS
Oat	Shotgun	119 million	\$485,000,000	
	CBCS	15 million	\$51,000,000	\$434,000,000 SAVINGS
Pea	Shotgun	32 million	\$108,800,000	
	CBCS	6.4 million	\$21,760,000	\$87,040,000 SAVINGS
Maize	Shotgun	19 million	\$65,000,000	
	CBCS	7 million	\$24,000,000	\$41,000,000 SAVINGS
Sorghum	Shotgun	5.8 million	\$20,000,000	
	CBCS	2.2 million	\$7,750,000	\$12,250,000 SAVINGS

*600 bp insert size, 99% probability of capturing entire sequence complexity of genome, \$3.40 per clone

ACKNOWLEDGEMENTS

Research funded in part by USDA-NRICGP award 99-35300-7819 to D.G. Peterson.