causes the break of a hit into two, one upstream and another downstream of the gap. Approximately 94% and 96% of the human and baboon sequences, respectively, was included in hits (Table 1).

Acknowledgements
We thank David Harris and John Spouge and two anonymous reviewers for discussions or comments on the manuscript. J.C.S. carried out this work while at the National Center for Biotechnology Information.

References

Joana C. Silva*
The Institute for Genomic Research, Medical Center Drive, Rockville, MD 20850, USA.*e-mail: jsilva@tigr.org

Alexey S. Kondrashov
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bldg. 45, 6AN.24, Bethesda, MD 20892, USA.

Techniques & Applications
Efficient capture of unique sequences from eukaryotic genomes
Daniel G. Peterson, Susan R. Wessler and Andrew H. Paterson

Cot-based cloning and sequencing (CBCS), a synthesis of Cot analysis, DNA cloning and high-throughput sequencing, promises to accelerate the study of eukaryotic genomes. In particular, CBCS will (1) permit efficient gene discovery in species with substantial quantities of repetitive DNA, (2) allow the sequence complexity (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 accelerate comparative genomics research, especially in large genomes such as those of many crops.

Published online: 15 August 2002

In higher eukaryotes, much of the cost of complete genome sequencing is expended on repetitive DNA with no known function. Effective strategies for the ‘capture’ (isolation and sequencing) of an organism’s sequence complexity (SqCx; see Glossary) would facilitate research in species with large genomes, such as those of many major crop plants, by limiting redundant sequencing of repetitive elements.

Cot analysis, an old but powerful biochemical technique, could have an important role in the extraction of unique sequence information from large, repetitive genomes. Developed by Roy Britten and colleagues nearly 35 years ago, Cot analysis is based on the observation that in a solution of heat-denatured, sheared genomic DNA, a specific sequence reassociates at a rate proportional to the number of times it occurs in the genome [1,2]. In a standard Cot study, aliquots of sheared genomic DNA are denatured, and each sample is allowed to renature to a specific Cot value. Hydroxypatite chromatography (HAP chromatography) is then employed to separate single-stranded DNA from double-stranded DNA, and the relative fraction of the genome that has reassociated at each sample’s Cot value is determined. A Cot curve showing reassociation as a function of the log of Cot value (from Cot = 0 until renaturation is complete) provides information about the genome and its various kinetic components (Fig. 1). Resolution of distinct kinetic
components can only be achieved if the DNA fragments used in Cot analysis are relatively short (200–600 bp) [1–3].

Even in its heyday, Cot analysis was only performed in a handful of labs, as it requires considerable technical skill and an extensive knowledge of renaturation theory. With the advent of molecular biology techniques in the late 1970s, even the most successful practitioners of Cot analysis began to abandon it, and by the mid 1980s it was well on its way to becoming a 'lost art' (Fig. 2).

However, Cot/HAP techniques have since been used to construct normalized cDNA libraries [4,5], isolate repetitive genomic DNA for use in chromosomal in situ suppression hybridization [6], clone DNA regions associated with known chromosomal deletions/additions using the phenol emulsion reassociation technique [7,8], and characterize several highly repetitive elements from the ginseng genome [9].

Recently, we developed Cot-based cloning and sequencing (CBCS) as a strategy for efficiently discovering genes with minimum encumbrance by repetitive DNA, (2) a means of capturing the SqCx of large genomes, and (3) a supplemental tool in genome sequencing. Briefly, a Cot analysis is performed for a species of interest, the results of the Cot analysis are used to guide the HAP-based fractionation of the genome into its major kinetic components, each isolated component is cloned separately to create a Cot library, and clones from each library are sequenced in numbers proportional to the kinetic complexity of their respective components.

In an initial study, we generated highly repetitive (HR), moderately repetitive (MR), and single/low-copy (SL) Cot libraries for sorghum, and showed through sequence and blotting analysis that each sorghum Cot library is representative of the Cot component from which it was derived, proving that CBCS is feasible [10].

CBCS in gene discovery
CBCS promises to resolve some of the difficulties that are currently associated with isolation of comprehensive sets of genes from large-genome species.

Expressed sequence tag (EST) or 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 10^5 sequences, many genes expressed only

---

Fig. 1. Onion (Allium cepa) Cot curve. Cot data is from Stack and Comings (1979) and is used with permission of the authors and the publisher. The complete curve consists of foldback (FB) sequences (pink) and highly repetitive (HR, blue), moderately repetitive (MR, green), and single/low-copy (SL, red) components characterized by very fast, fast, intermediate and slow reassociation, respectively. Yellow crosses mark the 'Cot½' values for the HR, MR and SL components. (A component's Cot½ value is the point on the abscissa of the complete Cot curve at which half the DNA in that component has reassociated.) For a Cot component, 80% of the sequences in that component will reassociate in the 'two Cot component region' (TCDR) flanking the component's Cot½ value (see brackets centered at Cot½ markers). For the HR and MR components, double-stranded DNA from the component's TCDR (blue and green shading, respectively) can be isolated and used to construct a corresponding Cot library. For the SL component, single-stranded DNA within the TCDR (red shading) can be used to generate duplexes (using the random-primer method) suitable for Cot library construction.
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 their tissue- or organ-specific patterns of expression.

CBCS possesses a significant advantage over methyl-filtration, a technique that has been suggested as an intermediate step between EST and genomic shotgun sequencing [11]. Briefly, methyl-filtration results in the production of genomic libraries enriched in hypomethylated (presumably gene) sequences. Although 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 [12–17]. Consequently, exclusion of hypermethylated DNA will probably result in the loss of important or interesting genes. Initial comparisons of genomic sequences from bacterial artificial chromosome (BAC) clones with sequences from libraries enriched in hypomethylated sequences suggest that as few as 50% of genes are recovered by methylation-based gene enrichment techniques (see abstract of M. Vaudin et al., 44th Maize Genetics Conference [2002]; www.agron.missouri.edu). Because HAP-based fractionation of genomic DNA is independent of sequence methylation [18], 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, highly 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 (Fig. 3). Using a shotgun approach, the number of different clones (n) that must be sequenced to have 99% confidence that all genomic elements have been sequenced at least once is estimated using the formula:

$$n = \ln(1 - 0.99) + \ln\left(1 - \frac{Z}{G}\right)$$

[Eqn 1]

where Z is the mean insert size in bp and G is 1c genome size in bp [10]. In CBCS, sequencing resources are allocated on the basis of 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 [2], 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 basepairs it contains – this is likely to prove a very conservative estimate.

Fig. 2. The rise and fall of Cot analysis: the number of publications including at least one Cot analysis for each year. Red bars represent Cot analyses performed by us. An extensive listing of Cot analysis publications can be found at http://www.plantgenomics.uga.edu/CBCS or http://www.msstate.edu/research/mgel/cbcs/cbcs.htm.

Fig. 3. 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 isolated and cloned separately. Because almost all of the SqCx is contained within the SL Cot library, most sequencing resources can be devoted to sequencing SL Cot clones.
in most genomes. For a genome composed of the components a, b and c with f bp of foldback DNA:

\[ n = \ln \left( 1 - 0.99^{\frac{Z}{\gamma_a + \gamma_b + \gamma_c + f}} \right) \]  

[Eqn 2]

CBCS reduces by two-thirds or more the number of clones that need to be sequenced to capture the SqCx of many eukaryotic genomes (Fig. 4). For example, the onion (Allium cepa, \( 1C = 15,544 \) Mb) genome is composed of a F B fraction of 1.12 \( \times 10^9 \) bp and HR, MR, and SL components with kinetic complexities of 2.86 \( \times 10^9 \), 2.43 \( \times 10^9 \) and 9.25 \( \times 10^8 \) bp, respectively [19]. Using Eqs 1 and 2, and assuming an average insert size of 600 bp, capture of 99% of onion’s SqCx would require 119 million shotgun sequences but only 16 million CBCS sequences – an 87% saving.

As with shotgun sequencing, standard assembly and finishing techniques [20] would be required to generate full-length gene and genome sequences from CBCS data.

Conclusions
CBCS is a powerful means of discovering genes. Because it is independent of expression and methylation patterns, CBCS is well suited for isolating key regulatory sequences and genes expressed at low levels, during short developmental timeframes and in response to subtle environmental stimuli. Although CBCS-based capture of SqCx does not provide information on the exact chromosomal locations of all sequences or information on small variations in individual members of repetitive DNA families, it should permit elucidation of the unique elements of even the largest genomes at a fraction of the cost of genomic shotgun sequencing. CBCS promises to accelerate greatly the timetable for genome-wide study of many of the world’s biota including large-genome agricultural plants and animals that sustain humanity.

Acknowledgements
This project was supported in part by USDA-NCICGP award 99-35300-7819 to D.G.P.

References
8 Clarke, B. et al. (1992) Targeting deletion (homologous chromosome pairing locus) or addition line single copy sequences from cereal genomes. Nucl. Acids Res. 20, 1289–1292
10 Peterson, D.G. et al. (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. 12, 795–807
18 Burtseva, N.N. et al. (1979) Intron gene distribution of 5-methylcytosine and kinetics of the reassociation of cow blood lymphocyte DNA in the normal state and in chronic lymphocytic leukemia. Biochemistry (Moscow) 41, 1636–1641
20 Benos, P.V. et al. (2001) From first base: the sequence of the tip of the X chromosome of Drassophila melanogaster, a comparison of two sequencing strategies. Genome Res. 11, 710–730

D.G. Peterson
Dept of Plant and Soil Sciences, Mississippi State University, 117 Dorman Hall, Box 9555, Mississippi State, MS 39762, USA.

S.R. Wessler
Plant Biology Dept, University of Georgia, Miller Plant Sciences Building, Athens, GA 30602, USA.

A.H. Paterson*
Center for Applied Genetic Technologies, University of Georgia, Riverbend Research Bldg., Room 162, 110 Riverbend Road, Athens, GA 30602, USA.

*e-mail: paterson@dogwood.botany.uga.edu