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

1 Casane, D. *et al.* (1997) Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* 45, 216–226
2 Matassi, G. *et al.* (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* 9, 786–791
3 Williams, E.J.B. and Hurst, L.D. (2000) The proteins of linked genes evolve at similar rates. *Nature* 401, 900–903
4 Lercher, M.J. *et al.* (2001) Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* 18, 2032–2039
5 Ebersberger, I. *et al.* (2002) Genome-wide comparisons of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* 70, 1490–1497
6 Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* 241, 3–17
7 Bielawski, J.P. *et al.* (2000) Rates of nucleotide substitutions and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. *Genetics* 156, 1299–1308
8 Hurst, L.D. and Williams, E.J.B. (2000) Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene* 261, 107–114
9 Hardison, R.C. *et al.* (1997) Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* 7, 959–966
10 Makalowski, W. and Boguski, M.S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9407–9412
11 Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 16, 369–372
12 Wasserman, W.W. *et al.* (2000) Human–mouse genome comparisons to locate regulatory sites. *Nat. Genet.* 26, 225–228
13 Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304
14 Shabalina, S.A. *et al.* (2001) Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* 17, 373–376
15 Petrov, D.A. and Hartl, D.L.(1999) Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 1475–1479
16 Krawczak, M. *et al.* (2000) Human gene mutation database – a biomedical information and research resource. *Hum. Mutation* 15, 45–51
17 Averof, M. *et al.* (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287, 1283–1286
18 Smith, N.G.C. and Hurst, L.D. (1999) The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* 153, 1395–1402
19 Krawczak, M. *et al.* (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.* 63, 474–488
20 Holmquist, G.P. (1994) Chromatin self-organization by mutation bias. *J. Mol. Evol.* 39, 436–438
21 Boulikas, T. (1992) Evolutionary consequences of nonrandom damage and repair of chromatic domains. *J. Mol. Evol.* 35, 156–180
22 Meijer, M. and Smerdon, M.J. (1999) Accessing DNA damage in chromatin: insights from transciption. *BioEssays* 21, 596–603
23 Ogurtsov, A.Y. *et al.* OWEN: aligning long collinear regions of genomes. *Bioinformatics* (in press)

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. HYDROXYAPATITE 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

 0168-9525/02/$ – see front matter PII: S0168-9525(02)02764-6

Glossary

Chromosomal *in situ* suppression hybridization: A method in which excess unlabeled repetitive genomic DNA isolated using Cot/HAP techniques is used to block repetitive elements in large genomic clones that have been selected as probes for *in situ* hybridization.
Cot value: In the context of this paper, a DNA sample's Cot value (in M.s) is defined as the product of its nucleotide concentration in moles per liter, its renaturation time (t) in seconds, and, if applicable, a buffer factor that accounts for the effect of positive ions on the speed of renaturation. However, readers should note the following:
(1) The scientific term 'Cot' originated as a way to easily pronounce the formula C_0t where C_0 is nucleotide concentration at time zero and t is renaturation time. 'Cot' and 'C_0t' (both pronounced 'kot') are used interchangeably in the literature, although the former is more common.
(2) Some authors use the term 'Cot' to describe only those DNA reassociation reactions that occur in 'neutral' (0.12 M) sodium phosphate buffer – in such instances, the term 'equivalent Cot' (Ecot) is used to describe renaturation in buffers other than 0.12 M sodium phosphate buffer.
Foldback (FB): The fraction of the genome exhibiting reassociation at the smallest Cot values attainable. FB DNA contains duplexes resulting from intrastrand pairing of complementary sequences.
Hydroxyapatite chromatography: The use of a hydroxyapatite column to separate single-stranded DNA and double-stranded DNA from mixtures containing both types of DNA molecule.
Kinetic complexity: An estimate of the SqCx of a particular kinetic component as determined in a Cot analysis.
Kinetic component: A group of genomic DNA sequences that exhibit similar reassociation properties and consequently appear as a mathematically distinct sigmoidal region of a complete Cot curve. The similarity in reassociation characteristics between different sequences in a kinetic component indicates that those sequences possess similar sequence complexities (i.e. they are found in similar copy numbers in the genome).
Normalized cDNA libraries: cDNA libraries from which the extremely common (i.e. highly expressed) sequences have been partially extricated. In general, Cot/HAP techniques are used to remove fast-reassociating (highly repetitive) sequences from cDNA populations.
Phenol emulsion reassociation technique (PERT): PERT is the reassociation of single-stranded DNA in an emulsion of phenol and aqueous buffer. PERT substantially increases the rate of DNA reassociation allowing Cot values of 200 000 or greater to be attained.
Sequence complexity (SqCx): The minimal group of sequences (expressed in terms of bp) that define a genome. For a eukaryote, SqCx is theoretically the combined length of all of the single-copy DNA sequences plus one copy of each repetitive sequence (Fig. 3).
Shotgun sequencing: Sequencing of clones randomly selected from a DNA library.

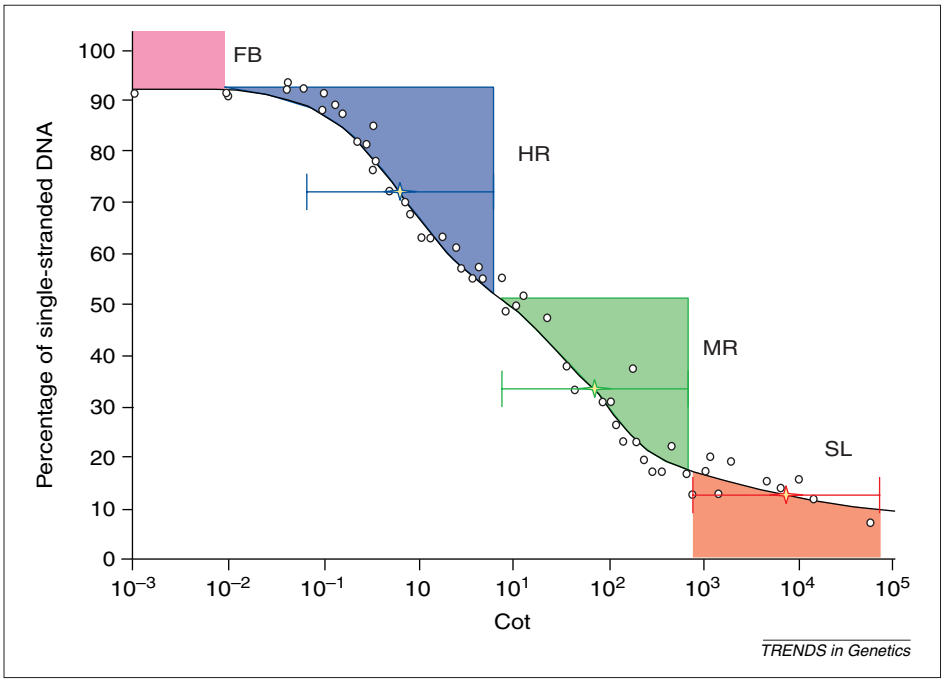

Fig. 1. Onion (*Allium cepa*) Cot curve. Cot data is from Stack and Comings (1979) and is used with permission of the authors and the publisher. The complete curve consists of foldback (FB) sequences (pink) and highly repetitive (HR, blue), moderately repetitive (MR, green), and single/low-copy (SL, 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 renature in the 'two Cot decade 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.

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 (1) 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

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.

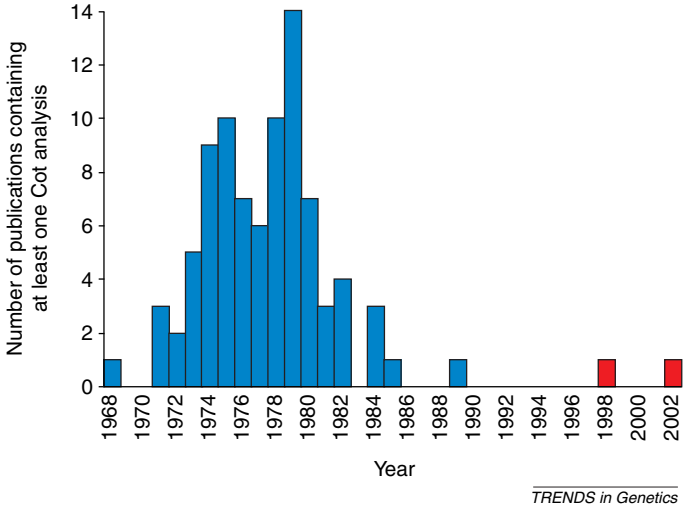

Fig. 2. The rise and fall of Cot analysis: the number of publications including at least one Cot analysis for each year. Red bars represent Cot analyses performed by us. An extensive listing of Cot analysis publications can be found at http://www.plantgenome.uga.edu/CBCS or http://www.msstate.edu/research/mgel/cbcs/cbcs.htm.

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:

$$n = \ln(1-0.99) \div \ln\left(1-\frac{Z}{G}\right) \qquad \text{[Eqn 1]}$$

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 base pairs it contains – this is likely to prove a very conservative estimate

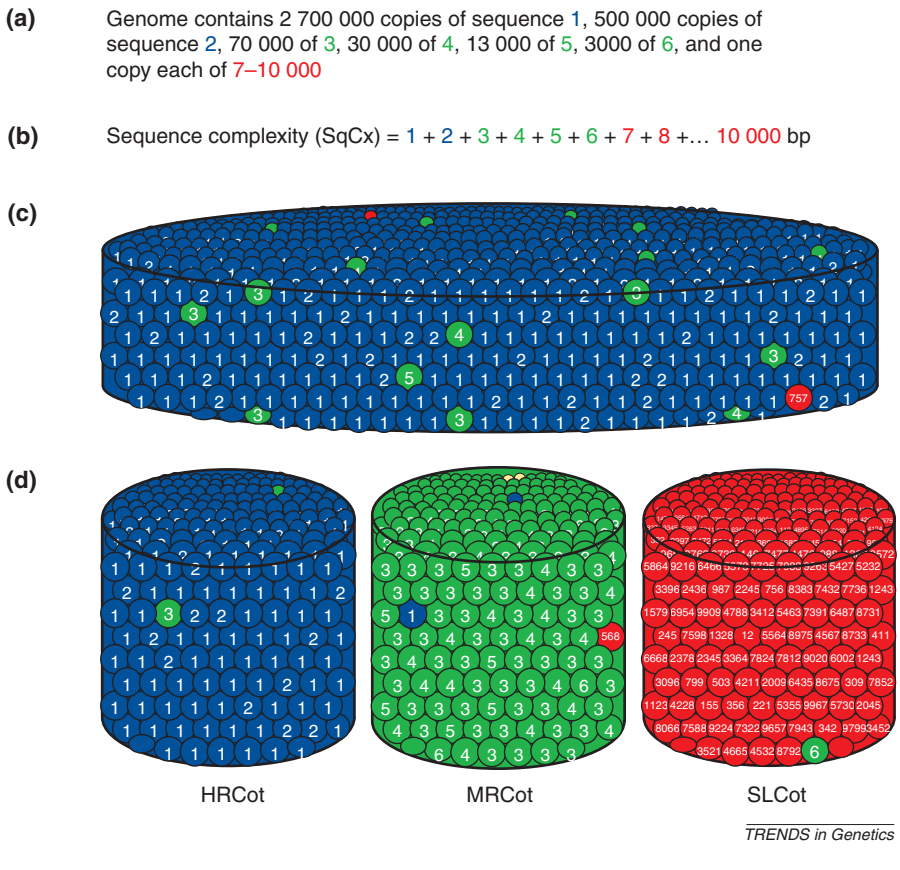

Fig. 3. SqCx and sequencing. (a) The elements constituting a hypothetical eukaryotic genome. (b) Though repetitive sequences account for the majority of DNA, they contribute very little to SqCx. (c) The net gain in novel sequence information is slow and costly if clones are selected from an unbiased genomic library (shotgun approach). (d) CBCS permits the highly repeptitive (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 SLCot library, most sequencing resources can be devoted to sequencing SLCot clones.

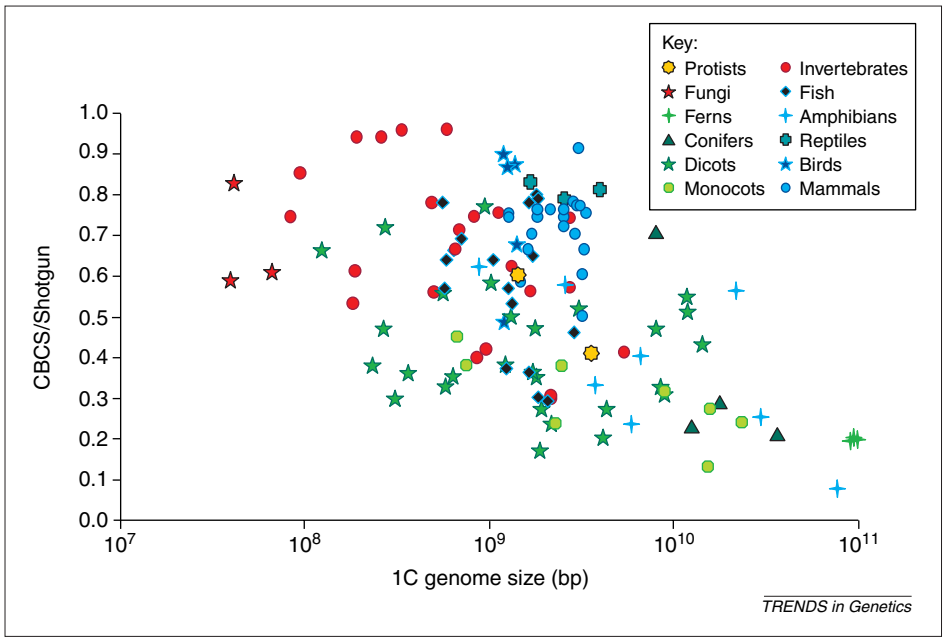

Fig. 4. Cot-based cloning and sequencing (CBCS) versus shotgun sequencing. For each species, the number of Cot clones that would need to be sequenced to attain a specific level of SqCx coverage has been divided by the number of 'shotgun clones' that would have to be sequenced to attain the same level of coverage. Resulting values have been plotted against genome size. See http://www.plantgenome.uga.edu/CBCS or http://www.msstate.edu/research/mgel/cbcs/cbcs.htm for reference data.

in most genomes. For a genome composed of the components *a*, *b* and *c* with *f* bp of foldback DNA:

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

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 FB fraction of 1.12×10^9 bp and HR, MR, and SL components with kinetic complexities of 2.86×10^5, 2.43×10^7 and 9.25×10^8 bp, respectively [19]. Using Eqns 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-NRICGP award 99-35300-7819 to D.G.P.

References

1 Britten, R.J. and Kohne, D.E. (1968) Repeated sequences in DNA. *Science* 161, 529–540
2 Britten, R.J. *et al.* (1974) Analysis of repeating DNA sequences by reassociation. *Methods Enzymol.* 29, 363–405
3 Davidson, E.H. *et al.* (1973) General interspersion of repetitive with non-repetitive sequence elements in the DNA of *Xenopus*. *J. Mol. Biol.* 77, 1–23
4 Ko, M.S.H. (1990) An 'equalized cDNA library' by reassociation of short double-stranded cDNAs. *Nucleic Acids Res.* 18, 5705–5711
5 Soares, M.B. *et al.* (1994) Construction and characterization of the normalized cDNA library. *Proc. Natl. Acad. Sci. U. S. A.* 91, 9228–9232
6 Landegent, J.E. *et al.* (1987) Use of whole cosmid cloned genomic sequences for chromosomal localization by non-radioactive *in situ* hybridization. *Hum. Genet.* 77, 366–370
7 Kunkel, L.M. *et al.* (1985) Specific cloning of DNA fragments absent from the DNA of a male patient with an X chromosome deletion. *Proc. Natl. Acad. Sci. U. S. A.* 82, 4778–4782
8 Clarke, B. *et al.* (1992) Targeting deletion (homeologous chromosome pairing locus) or addition line single copy sequences from cereal genomes. *Nucleic Acids Res.* 20, 1289–1292
9 Ho, I.S.H. and Leung, F.C. (2002) Isolation and characterization of repetitive DNA sequences from *Panax ginseng*. *Mol. Genet. Genomics* 266, 951–961
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
11 Rabinowicz, P.D. *et al.* (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* 23, 305–308
12 Simmen, M.W. *et al.* (1999) Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* 283, 1164–1167
13 Lois, R. *et al.* (1990) Active β-globin gene transcription occurs in methylated, DNase I-resistant chromatin of nonerythroid chicken cells. *Mol. Cell. Biol.* 10, 16–27
14 Wölfl, S. *et al.* (1991) Lack of correlation between DNA methylation and transcriptional inactivation: the chicken lysozyme gene. *Proc. Natl. Acad. Sci. U. S. A.* 88, 271–275
15 Heslop-Harrison, J.S. (2000) Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. *Plant Cell* 12, 617–635
16 Li, E. *et al.* (1993) Role for DNA methylation in genomic imprinting. *Nature* 366, 362–365
17 Riesewijk, A.M. *et al.* (1996) Maternal-specific methylation of the human *IGF2R* gene is not accompanied by allele-specific transcription. *Genomics* 31, 158–166
18 Burtseva, N.N. *et al.* (1979) Intragenome distribution of 5-methylcytosine and kinetics of the reassociation of cow blood lymphocyte DNA in the normal state and in chronic lympholeukemia. *Biochemistry* (*Mosc.*) 44, 1636–1641
19 Stack, S.M. and Comings, D.E. (1979) The chromosomes and DNA of *Allium cepa*. *Chromosoma* 70, 161–181
20 Benos, P.V. *et al.* (2001) From first base: the sequence of the tip of the X chromosome of *Drosophila 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