NOTE / NOTE

The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.)

Didier Lamoureux, Daniel G. Peterson, Wanlong Li, John P. Fellers, and Bikram S. Gill

Abstract: We report the results of a study on the effectiveness of Cot filtration (CF) in the characterization of the gene space of bread wheat (*Triticum aestivum* L.), a large genome species (1C = 16 700 Mb) of tremendous agronomic importance. Using published Cot data as a guide, 2 genomic libraries for hexaploid wheat were constructed from the single-stranded DNA collected at Cot values > 1188 and 1639 M·s. Compared with sequences from a whole genome shotgun library from *Aegilops tauschii* (the D genome donor of bread wheat), the CF libraries exhibited 13.7-fold enrichment in genes, 5.8-fold enrichment in unknown low-copy sequences, and a 3-fold reduction in repetitive DNA. CF is twice as efficient as methylation filtration at enriching wheat genes. This research suggests that, with improvements, CF will be a highly useful tool in sequencing the gene space of wheat.

Key words: gene enrichment, renaturation kinetics, gene-rich regions, bread wheat.

Résumé : Nous présentons les résultats d'une étude sur l'efficacité de la filtration par Cot (CF) pour caractériser la fraction génique du blé tendre (*Triticum aestivum*), une espèce à grand génome (1C = 16 700 Mb) dont l'importance agronomique est capitale. En utilizant des données Cot publiées, 2 banques génomiques de blé hexaploïde ont été construites à partir de fractions collectées aux Cot > 1188 et 1639 M·s. Comparées aux séquences d'une banque génomique totale d'*Aegilops tauschii* (le donneur du génome D du blé tendre), les banques CF ont montré un enrichissement en gènes d'un facteur 13,7, un enrichissement en séquences inconnues à faible nombre de copies d'un facteur 5,8 tandis que l'ADN répété a été divisé par 3. La CF est deux fois plus efficace que la filtration par méthylation pour l'enrichissement en gènes du blé. Ces travaux suggèrent qu'avec des améliorations, la CF sera un outil très important pour le séquençage des gènes du blé tendre.

Mots clés : enrichissement en gènes, cinétique de renaturation, régions riches en gènes, blé tendre.

Introduction

Wheat is one of the principal sources of calories consumed by humans. More field space is devoted to wheat production than to any other crop (210 million ha in 2002 vs. 147 ha for rice and 139 ha for maize), and the trade value of wheat exceeds that of all other cereal species, including rice and maize ($US 31 billion world trade in 2001 vs. $US 13 billion for rice and $US 19 billion for maize; FAOSTAT 2005). Because of its economic importance, genome sequencing and analysis of the wheat gene space are the next logical steps for wheat improvement. However, bread wheat has one of the largest genomes among crop plants (~16 700 Mb; Bennett and Leitch 2003). Whole-genome shotgun sequencing of wheat is currently cost prohibitive, and furthermore, unambiguous assembly may not be possi-

Received 11 February 2005. Accepted 22 August 2005. Published on the NRC Research Press Web site at http://genome.nrc.ca on 9 December 2005.

Corresponding Editor: J.P. Gustafson.

D. Lamoureux. Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, Kansas, KS 66506, U.S.A.; and INRA Amélioration et Santé des Plantes, 63039 Clermont-Ferrand, CEDEX 2, France.
D.G. Peterson. Department of Plant and Soil Sciences, Mississippi State University, MS 39762, USA.
W. Li and B.S. Gill.[1] Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, Kansas, KS 66506, USA.
J.P. Fellers. USDA-ARS, Department of Plant Pathology, Kansas State University, Manhattan, Kansas, KS 66506, USA.

[1]Corresponding author (e-mail: bsgill@ksu.edu).

 doi: 10.1139/G05-080 © 2005 NRC Canada

ble, since repeat sequences account for at least 85% of the wheat genome (Smith and Flavell 1975). Alternative approaches that allow gene sequencing (normally found in lowcopy numbers) with minimum hindrance by repetitive DNA promise an affordable gateway into the wheat gene space. Presently, there are 3 principal means being employed to enrich the gene space in large, repetitive genomes: expressed sequence tag (EST; Venter 1993) sequencing, methylation filtration (MF; Rabinowicz et al. 1999), and Cot filtration (CF; Peterson et al. 2002*a*, 2002*b*; Yuan et al. 2003).[2]

EST sequences generated by single-pass sequencing of cDNA clones from different tissues represent an economical means to acquire many of the coding regions of genes. However, EST sequencing provides no information on noncoding regions of genes (e.g., promoters, introns, and other functional elements). Additionally, genes expressed at low levels and (or) in response to highly specific stimuli are likely to be missed using an EST approach.

MF is another gene-enrichment option based on evidence that genes in some plant genomes tend to be hypomethylated, while retroelements and other repeats are hypermethylated (Bennetzen et al. 1994). In MF, genomic DNA is cloned into a bacterial strain containing a restriction system that preferentially cleaves methylated DNA. Consequently, only clones containing hypomethylated inserts survive on selective medium (Rabinowicz et al. 1999). MF has been successfully applied to sequencing the maize gene space (Palmer et al. 2003; Whitelaw et al. 2003). However, the relationship between methylation and genetic inactivity is not universal among plants. Methylation, and subsequently, transcription patterns within a plant, can change drastically during major developmental transitions (Law and Suttle 2001; Prakash et al. 2003) and (or) in response to abiotic stress (Kovalchuk et al. 2003; Meng et al. 2003; Baurens et al. 2004). Thus, many developmental and stress-activated genes may be lost using methyl filtration.

In CF, genomic DNA is heat denatured and allowed to renature to a Cot value (Cot = DNA concentration × time × a factor based on the cation concentration of the buffer), at which the majority of repetitive elements reassociate, but single- and low-copy elements remain single stranded (see Peterson 2005 for review). Double-stranded, repetitive DNA is separated from single-stranded, low-copy DNA by hydroxyapatite chromatography, and the single-stranded fraction is used to create gene-enriched libraries. Recently, the feasibility of CF as a tool for gene enrichment was demonstrated in sorghum (Peterson et al. 2002*a*) and maize (Yuan et al. 2003; Whitelaw et al. 2003). Unlike EST sequencing and methyl filtration, CF is independent of sequence expression and methylation patterns (see Peterson et al. 2002*b* and Peterson 2005 for review).

In the present study, the ability of CF to enrich genes and low-copy DNA and to filter out repetitive DNA was tested for bread wheat. This study complements recently published research in which the effectiveness of MF was explored in *Aegilops tauschii*, the D genome donor of wheat (Li et al. 2004). Comparative information about various gene enrichment techniques is vital to the international wheat community as they develop a strategy for sequencing the wheat gene space.

Materials and methods

DNA isolation, Cot analysis, and cloning

Nuclear DNA was isolated from 6-day-old seedlings (roots and shoots) of wheat (*Triticum aestivum* 'Chinese Spring'), as described previously (Peterson et al. 1997) with minor modifications (see http://www.mgel.msstate.edu/pdf/nucl_dna.pdf). The DNA was sheared with a VirSonic 50 sonicator (VirTis, Gardiner, N.Y.) with full power settings for 30 s to achieve an average fragment size of about 800 bp. Renaturation of genomic DNA to specific Cot values was performed as described in Peterson et al. (2002*a*) using a Beckman DU640B spectrophotometer (Beckman Coulter, Fullerton, Calif.), except that renaturation and hydroxyapatite chromatography were performed at 60 °C.

Isolated single-stranded fractions were concentrated and transferred into TE buffer (10 mmol Tris/L, 1 mmol EDTA/L, pH 7) using Millipore Amicon Centriplus YM-30 columns (Millipore, Billerica, Mass.). Library construction was performed according to Yuan et al. (2003), except that Qiaquick columns (Qiagen, Valencia, Calif.) were used for intermediary DNA purification. Colonies were handpicked and arrayed in 384-well plates.

Sequencing

CF libraries were replicated and cultured in 96-well plates prior to plasmid isolation with a Qiagen BioRobot 3000 using the Qiaprep 96 Turbo BioRobot kit (Qiagen). Plasmids were then sequenced with an ABI Prism BigDye Terminator Ready Reaction Cycle Sequencing kit (Applied Biosystems, Foster City, Calif.). Reactions were run on an ABI Prism 3700 DNA Analyzer (Applied Biosystems), bases were called by Phred, and vector sequences were masked using CrossMatch (Ewing and Green 1998).

Sequence analysis

The CF sequences were used as queries for BLASTn (Altschul et al. 1997) searches in the National Centre for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) nr nucleotide sequence database, the Institute for Genomic Research (TIGR; http://www.tigr.org) databases for wheat gene index, rice coding sequences (CDS), and Gramineae repeats, and the USDA GrainGenes Triticeae repeat sequence (TREP; http://wheat.pw.usda.gov/ggpages/ITMI/Repeats; Wicker et al. 2002) database. The CF sequences were also used in BLASTx searches in the NCBI protein database and the TREP hypothetical protein sequences database. Based on sequence homology, CF sequences were divided into 3 categories: repeats, genes, and non-hits. A clone was classified as a repeat sequence if it showed similarity to a repeat family or a transposable element (TE) related protein with an E value at or below 10^{-5}. A clone was

[2] Cot filtration is equivalent to the single-/low-copy Cot (SLCot) portion of Cot-based cloning and sequencing (CBCS; Peterson et al. 2002*a*) and (or) high Cot sequencing (HC; Yuan et al. 2003). Because the abbreviation CBCS can also stand for clone-by-clone sequencing and because high Cot is frequently confused with high copy, the term Cot filtration and the abbreviation CF are becoming more widely used.

© 2005 NRC Canada

considered genic if it did not match any repeat but showed similarity either to a non-TE protein, cloned genes, cDNAs, rice CDS, or wheat ESTs (E value at or below 10^{-5}). The sequences in the non-hit class did not match any sequences in the aforementioned public (i.e., NCBI, TIGR, and TREP) databases. Statistical analyses were conducted using the Institute of Phonetic Sciences' Binomial Proportions program (http://fonsg3.let.uva.nl/Service/Statistics.html).

Results and discussion

Determination of Cot cloning value

More than 30 years ago, Mitra and Bhatia (1973) prepared a Cot curve for bread wheat. However, they did not perform a statistical analysis of the curve, and thus statistically relevant information on the features of individual Cot components was not available. To counter this deficiency, the *x* and *y* coordinates of each point on Mitra and Bhatia's Cot curve of wheat were determined and the resulting data set was analyzed using the reassociation kinetics computer program of Pearson et al. (1977). As shown in Fig. 1, the analysis providing the lowest RMS (root mean square deviation) and goodness-of-fit values (0.0217769 and 0.0249798, respectively) is a 3-component fit, with the fraction of DNA remaining single stranded at the highest Cot value fixed at 1% to prevent the generation of a negative value for this parameter. The wheat Cot curve consists of fast, intermediate, and slow reassociating components, with rate constants (*k* values) of 12.5 $M^{-1}\cdot s^{-1}$, 0.326 $M^{-1}\cdot s^{-1}$, and 8.42×10^{-5} $M^{-1}\cdot s^{-1}$, respectively (Fig. 1).

In diploid organisms, the slowest reassociating component of a Cot curve often represents single-copy DNA sequences and can thus be used to estimate genome size by comparison with the *Escherichia coli* rate constant (k_{coli} = 0.22 $M^{-1}\cdot s^{-1}$; Zimmerman and Goldberg 1977) and genome size G_{coli} = 4 639 221 bp; Blattner et al. 1997) as in Eq. 1:

$$[1] \qquad G_{coli} \times k_{coli} = G_{wheat} \times k_{wheat}$$

If the slow reassociating wheat component (k = 0.0000842 $M^{-1}\cdot s^{-1}$) is composed of single-copy DNA, the estimated 1C genome size of wheat would be G = (4 639 221 bp × 0.22 $M^{-1}\cdot s^{-1}$) ÷ 0.0000842 $M^{-1}\cdot s^{-1}$) = 1.212×10^{10} bp or 12 120 Mb. The 1C DNA content for wheat determined by Feulgen densitometry is 16 700 Mb, or roughly 1.38 times that predicted if the slow reassociating component is truly single-copy DNA. Although the DNA content estimated from the Cot curve is within an acceptable range of error, wheat's fairly recent hexaploid status makes it probable that its slow reassociating component is a combination of single- and low-copy sequences. Consequently, the smaller genome size predicted from the Cot curve may be a reflection of polyploidization. Since k is the inverse of Cot½ (Britten et al. 1974), the Cot½ for the single-/low-copy component in this experiment is 11 876 M·s. According to the Two-Cot Decade rule, 90% of the slow reassociating component should renature after a Cot value of 1188 M·s (see Peterson 2005 for review). Thus, a small CF library was prepared from wheat DNA with a Cot > 1188 M·s.

Based on a genome size for wheat of 16 700 Mb, the predicted k value for a pure single-copy component would be k = (4 639 221 bp × 0.22 $M^{-1}\cdot s^{-1}$) ÷ 16 723 Mb = 6.103×10^{-5} $M^{-1}\cdot s^{-1}$. Consequently, the predicted Cot½ for a pure single-copy component of the wheat genome is 16,385 M·s, and 90% of the single-copy DNA of wheat should remain single stranded at a Cot value of 1639 M·s. For comparison with the CF1188 library, a CF library was prepared from wheat DNA with a Cot > 1639 M·s.

Sequence analysis

CF libraries were constructed for wheat based on the Cot curve of Mitra and Bhatia (1973), (i.e., Cot > 1188, or CF1188) and based on the theoretical Cot½ of single-copy wheat DNA, assuming a genome size of 16 700 Mb (i.e., Cot > 1639, or CF1639). We picked 258 clones for CF1188 and 3840 clones for CF1639 with insert sizes ranging from 800 to 2000 bp. Single-pass sequencing of these clones produced 252 and 2282 sequence reads, respectively. With an average length of 635 bp, these 2534 reads represent the first 1433 kb of the wheat genomic sequence in GenBank (accession Nos. CL900626–CL902992 and CW991694–CW991860). We observed only 1 mitochondrial sequence in the CF1188 library and 1 chloroplast sequence in the CF1639 library, which suggests less than 0.4% and 0.07% organellar DNA contamination, respectively.

No significant difference in the sequence composition was found in either the CF1188 or the CF1639 libraries (P = 0.779 for repeat content, P = 0.121 for gene content, and P = 0.067 for the non-hit percentage). Therefore, we combined sequences from the 2 CF libraries for all subsequent analyses.

We identified 868 genic sequences based on the alignments with gene, EST, and protein sequences deposited in the public databases, which accounts for 34.3% of the CF libraries. Of the 868 genic sequences, 425 showed sequence similarity to database entries at an E value cut-off < 10^{-20}, 265 showed similarity to entries with an E value between 10^{-20} and 10^{-10}, and the remaining 178 showed similarity to entries with an E value between 10^{-10} and 10^{-5} (Table 1). In total, 220 CF sequences matched protein sequences from the NBCI nr database (see Supplementary data[3]). Of the 220 CF sequences, 108 matched wheat ESTs with an E value < 10^{-20}, 143 matched with an E value < 10^{-10}, and 151 matched with an E value of 10^{-5}, suggesting that the current coverage of wheat EST collection ranges between 49% and 69%. Furthermore, 42 of the 220 sequences recognized only protein database entries and did not show significant homology to any DNA database entries.

Alignments with repeats from the cereal genomes identified 792 (31.4%) CF clones containing repeated sequences (Table 1). Retrotransposons contributed 81% to the repeated sequences. Seventy-two sequences showed similarities to miniature inverted-repeat transposable elements (MITEs), most of which belong to the *stowaway* superfamily. Half of

[3] Supplementary data for this article are available on the Web site or may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Building M-55, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada. DUD 4057. For more information on obtaining material refer to http://cisti-icist.nrc-cnrc.gc.ca/irm/unpub_e.shtml.

© 2005 NRC Canada

Fig. 1. Analysis of the bread wheat Cot data of Mitra and Bhatis (1973) using the program of Pearson et al. (1977). Fraction refers to the percentage of the Cot curve occupied by a particular component. Cot½, the value on the *x* axis at which 50% of the DNA in a component has reassociated; *k*, the rate constant of a given component (the inverse of its Cot½ value), and KnCx, kinetic complexity, the estimated sequence complexity of the component. The curve has 3 distinct parts: a fast reassociating (highly repetitive) component, a middle reassociating (moderately repetitive) component, and a slow reassociating (single- and low-copy) component. The Cot½ value of each component is marked by a star, and the Two Cot Decade region flanking each Cot½ value is indicated by a horizontal bar centered at the Cot½ value and enclosed with small vertical bars. The portion of the curve showing Cot > 1190 DNA used in cloning is shaded. Because the data of Mitra and Bhatia (1973) was apparently normalized so that 100% of the DNA was found in the 3 kinetic components (i.e., foldback and unannealable portions of the genome were left out of the analysis), the curve generated by the program of Pearson et al. (1977) starts at a value slightly greater than 100% single-stranded DNA.

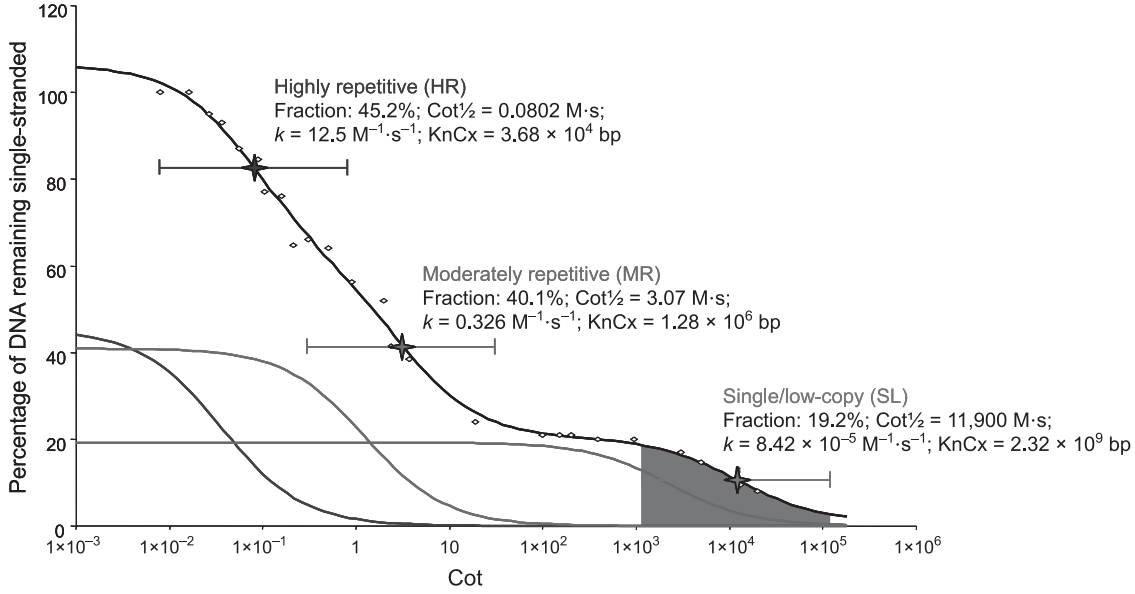


Table 1. Categories and frequencies of wheat Cot filtration sequences.

Category	Class	No.	%
Genes	E<10^{-20}	425	16.8
	E<10^{-10}	265	10.5
	E<10^{-5}	178	7.0
Total genes		868	34.3
Retroelements	Ty1-*copia*	94	3.7
	Ty3-*gypsy*	364	14.4
	non-LTR	73	2.9
	unknown	113	4.5
Transposons	CACTA	42	1.7
	MITEs	51	2.0
	others	13	0.5
Tandem repeats		2	0.1
Unclassified repeats		40	1.6
Total repeats		792	31.4
Non-hit		872	34.4

the MITEs were found to be associated with genes in wheat CF sequences. Because MITEs tend to be preferentially associated with genes (Zhang et al. 2000), it is possible that relatively low-copy MITEs and genes were enriched together.

The remaining 872 (34.4%) sequences did not match any sequences in the public databases and were classified as non-hit. These non-hit sequences are potentially new, uncharacterized coding sequences, noncoding regions of genes (e.g., regulatory elements and introns), or other low-copy elements.

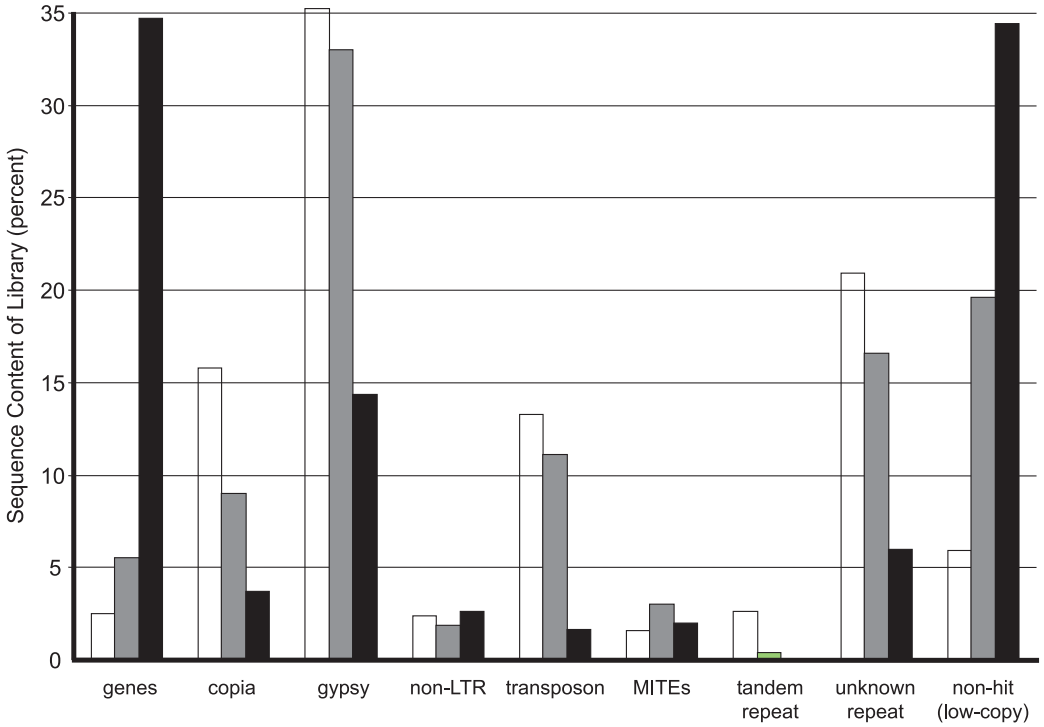
Filtration efficiency

Smith and Flavell (1975) estimated that 85% of the wheat genome is repetitive based on DNA reassociation kinetics experiments (note: they did not perform a complete Cot analysis). Based on analysis of a whole genome shotgun (WGS) library of *Ae. tauschii* (the D genome donor of wheat), Li et al. (2004) estimated a repeat content of 91% and a gene content of 2.5%. Assuming that bread wheat has a similar gene content to *Ae. tauschii*, CF technology has reduced wheat repeated sequences by 3-fold and enriched wheat genes by 14-fold (Table 1 and Fig. 2). Compared with the maize high-Cot (>466 M·s) library, which exhibits 26% repeats and 23% genes (Yuan et al. 2003), our result showed slightly higher efficiency of filtration for genes and against repeated sequences. However, it is important to remember that the maize genome is about 7 times smaller than the hexaploid wheat genome. Consequently, the savings to be expected from CF for genome sequencing should be relatively larger in wheat. In reality, the gene fraction in the wheat CF library may be higher than 34.6%. Indeed, the non-hit sequences may contain unidentified genes.

In the wheat CF libraries, not all repeat types were filtered evenly (Table 1 and Fig. 2). For example, the ratio of Ty3-*gypsy* to Ty1-*copia* is 2.2 in the D genome, but the ratio is 3.8 in our CF libraries. This may suggest a higher degree of

© 2005 NRC Canada

Fig. 2. Comparison of the sequence contents of *Ae. tauschii* unfiltered (UF,) and methylation-filtered (MF,) libraries (Li et al. 2004) with the *T. aestivum* Cot-filtered libraries (CF,). The CF unknown repeat bar in the figure represents the sum of the unknown retroelement and unclassified repeat categories in Table 1. The CF gene category contains all sequences showing homology to non-TE ESTs, rice CDS, or known genes at an E value $\leq 10^{-5}$ (Table 1).



sequence divergence of Ty3-*gypsy* retrotransposons than the Ty1-*copia* type in wheat. In pine, diverged retroelements significantly contribute to the amount of low-copy sequences (Elsik and Williams 2000).

Based on 1.02-Mb WGS sample sequences, it is estimated that wheat has a GC content of 45.7% (Lagudah et al. 2001), which is very close to the 46.0% GC content of the D genome (Li et al. 2004). The GC content for wheat ESTs is 52% (Li et al. 2004). Compared with the *Ae. tauschii* WGS sequence, the GC content increased from 1.1% to 46.8% in our CF libraries. The highest GC content was estimated at 47.9% for the non-hit sequences, and the lowest was 43.5% for the MITEs. Compared with the D genome WGS library, GC content was elevated for the retroelements (*copia*, *gypsy*, and non-LTR elements) and MITEs, but surprisingly, it decreased for genic sequences and CACTA transposons (Table 2). Considering that GC content is a potential factor influencing reassociation kinetics, the reduction of GC content for genic sequences suggests that CF tends to enrich genes with lower GC content more efficiently than those with high GC content. This bias could presumably be remedied by constructing several CF libraries that differ with regard to their Cot cutoff values and (or) the reassociation temperature.

A comparison of the filtration abilities of CF vs. MF is shown in Fig. 2. For *Ae. tauschii*, MF results in only a 2.2-fold enrichment in genes (Li et al. 2004). In contrast, a comparison of *T. aestivum* CF sequences with *Ae. tauschii* unfiltered sequences indicates that CF results in a 14-fold gene enrichment. Likewise, MF produces only a 1.2-fold reduction in repeat content, while CF reduces repeats 3-fold.

Table 2. Comparison of GC content (%) for different categories of sequences from hexaploid wheat Cot-filtration (CF) libraries and a D genome whole-genome shotgun (WGS) library. (Li et al. 2004).

Category	Wheat CF	D genome WGS	Significance
Total average	46.8	46	$P < 2.1\times10^{-51}$
copia retrotransposons	47.1	44.1	$P < 6.5\times10^{-34}$
gypsy retrotransposons	47.4	46.5	$P < 2.0\times10^{-10}$
non-LTR retrotransposons	46.5	44.6	$P < 4.9\times10^{-06}$
CACTA transposons	43.6	46.3	$P < 1.9\times10^{-13}$
MITEs	43.5	41.8	$P < 7.7\times10^{-05}$
Genes	46.4	48.2	$P < 0.002470$
No hits	47.9	—	n/a

However, it should be noted that MF data for *T. aestivum* (Rabinowicz et al. 2005) suggests an MF gene enrichment of roughly 4.7-fold.

Potential improvements and future applications

At present, the construction of Cot curves is labor-intensive work requiring a considerable understanding of biochemistry. In the present research, we compared CF based on the results of a complete Cot curve with CF for which the Cot cutoff value was derived from genome size alone. Perhaps not surprisingly, we found only negligible differences between the sequence content of the 2 libraries. This suggests that, for the isolation of single- and low-copy sequences from large genome species, the construction of a

© 2005 NRC Canada

complete Cot curve to determine fractionation parameters may be unnecessary. Rather, the use of a Cot value derived from a known genome size, as shown in eq. 1, may be sufficient to get quality fractionation of single- and low-copy sequences, provided the genome size is reasonably accurate.

An obvious means by which to increase the level of gene enrichment would be to increase the Cot value used for cloning. Unfortunately, the lack of a significant difference between the 2 assayed Cot values, 1188 and 1639, suggests that cloning at a higher Cot value may not dramatically enhance the results. In fact, increasing the Cot value could lead to the loss of genes present in low-copy gene families. Because bread wheat is hexaploid, most genes are presumably triplicated in the genome. Using CF cutoffs based on the assumption that the slowest reassociating component of the genome is composed of duplicated genes (tetraploid, CF > 1639 ÷ 2 = 819.5 M·s) or triplicated genes (hexaploid, CF > 1639 ÷ 3 = 546.3 M·s) rather than single-copy genes (diploid, CF > 1639 M·s) could be useful in capturing gene families that might have been eliminated in our initial experiments.

Another potential way of increasing the level of gene enrichment would be to use a larger fragment size to obtain longer stretches of single- and low-copy DNA; this might also increase the probability of securing complete genes. However, changing this parameter should be considered with care, since increasing fragment length may result in the elimination of low-copy sequences near (or flanked by) repeats. Indeed, increasing the fragment length also increases the statistical percentage of obtaining fragments that contain both low-copy and medium- or high-copy elements. The behavior of such fragments is not predictable and is likely to be very complex (Flavell and Smith 1976). Despite a better level of enrichment, the loss of unique sequence information may not be worth the change. A possible compromise would be to make CF libraries from DNA fragments of several different lengths and sequence from each of the resulting libraries. Considering that current sequencing technology routinely allows about 700 bp to be read in a single pass, sequencing clones from the present library from both ends to get the full length of inserts of 1 or 2 kb would undoubtedly increase the amount of gene space captured by CF.

An expanded version of the present CF library should greatly facilitate sequencing of the wheat gene space, and indeed, CF libraries have been used in the sequencing and assembly of the maize gene space (Whitelaw et al. 2003). Because the coverage of wheat genes from ESTs is estimated to be about 60%, CF libraries are expected to help in the discovery and characterization of the 40% of remaining wheat genes that are absent from the wheat EST collection.

Acknowledgements

We thank Katie Gleason for technical help with the sequencing of the Cot clones, and Zhigang Xie for computer assistance in sequence analysis. This research was funded in part by NSF Awards 0115903 and 0421717 to D.G.P and NSF award 0077766 to B.S.G. This paper is contribution No. 04-299-J from the Kansas Agricultural Experiment station (Manhattan, Kans.).

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

Baurens, F.C., Nicolleau, J., Legavre, T., Verdeil, J.L., and Monteuuis, O. 2004. Genomic DNA methylation of juvenile and mature Acacia mangium micropropagated in vitro with reference to leaf morphology as a phase change marker. Tree Physiol. **24**: 401–407.

Bennett, M.D., and Leitch, I.J. 2003. Angiosperm DNA C-values database [online]. Available from http://www.rbgkew.org.uk/cval/homepage.html. [release 4.0, January 2003].

Bennetzen, J.L., Schrick, K., Springer, P.S., Brown, W.E., and SanMiguel, P. 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome, **37**: 565–576.

Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shao, Y. 1997. The complete genome sequence of *Escherichia coli* K-12. Science (Washington, D.C.), **277**: 1453–1474.

Britten, R.J., Graham, D.E., and Neufeld, B.R. 1974. Analysis of repeating DNA sequences by reassociation. Methods Enzymol. **29**: 363–405.

Elsik, C.G., and Williams, C.G. 2000. Retroelements contribute to the excess low-copy-number DNA in pine. Mol. Gen. Genet. **264**: 47–55.

Ewing, B., and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. Genome Res. **8**: 186–194.

FAOSTAT 2005. Food and Agriculture Organization of the United Nations (FAOSTAT) database [online]. Agriculture section. Available from http://faostat.fao.org [updated February 2005].

Flavell, R.B., and Smith, D.B. 1976. Nucleotide sequence organization in the wheat genome. Heredity, **37**: 231–252.

Kovalchuk, O., Burke, P., Arkhipov, A., Kuchma, N., James, S.J., Kovalchuk, I., and Pogribny, I. 2003. Genome hypermethylation in *Pinus silvestris* of Chernobyl—a mechanism for radiation adaptation? Mutat. Res. **529**: 13–20.

Lagudah, E.S., Dubcovsky, J., and Powell W. 2001. Wheat genomics. Plant Physiol. Biochem. **39**: 335–344.

Law, D.R., and Suttle, J.C. 2001. Transient decreases in methylation at 5′-CCGG-3′ sequences in potato (*Solanum tuberosum* L.) meristem DNA during progression of tubers through dormancy precede the resumption of sprout growth. Plant Mol. Biol. **51**: 437–447.

Li, W., Zhang, P., Fellers, J.P., Friebe, B., and Gill, B.S. 2004. Sequence composition, organization and evolution of a basic Triticeae genome of the grass family. Plant J. **40**: 500–511.

Meng, L., Bregitzer, P., Zhang, S.B., and Lemaux, P.G. 2003. Methylation of the exon/intron region in the Ubi1 promoter complex correlates with transgene silencing in barley. Plant Mol. Biol. **53**: 327–340.

Mitra, R., and Bhatia, C.R. 1973. Repeated and non-repeated nucleotide sequences in diploid and polyploid wheat species. Heredity, **31**: 251–262.

Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., Balija, V.S., Nascimento, L.U., Dike, S., de la Bastide, M., Martienssen, R.A., and McCombie, W.R. 2003. Maize genome sequencing by methylation filtration. Science (Washington, D.C.), **302**: 2115–2117.

© 2005 NRC Canada

Pearson, W.R., Davidson, E.H., and Britten, R.J. 1977. A program for least squares analysis of reassociation and hybridization data. Nucleic Acids Res. **4**: 1727–1737.

Peterson, D.G. 2005. Reduced representation strategies and their application to plant genomes. *In* The handbook of plant genome mapping: genetic and physical mapping. *Edited by* K. Meksem and G. Kahl. John Wiley & Sons, Inc., Hoboken, N.J. pp. 307–335.

Peterson, D.G., Boehm, K.S., and Stack, S.M. 1997. Isolation of milligram quantities of DNA from tomato (*Lycopersicon esculentum*), a plant containing high levels of polyphenolic compounds. Plant Mol. Biol. Rep. **15**: 148–153.

Peterson, D.G., Schulze, S.R., Sciara, E.B., Lee, S.A., Bowers, J.E., Nagel, A., Jiang, N., Tibbits, D.C., Wessler, S.R., and Paterson, A.H. 2002*a*. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. **12**: 795–807.

Peterson, D.G., Wessler, S.R., and Paterson, A.H. 2002*b*. Efficient capture of unique sequences from eukaryotic genomes. Trends Genet. **18**: 547–550.

Prakash, A.P., Kush, A., Lakshmanan, P., and Kumar, P.P. 2003. Cytosine methylation occurs in a CDC48 homologue and a *MADS-box* gene during adventitious shoot induction in Petunia leaf explants. J. Exp. Bot. **54**: 1361–1371.

Rabinowicz, P.D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L.D., Stein, L., McCombie, W.R., and Martienssen, R.A. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat. Genet. **23**: 305–308.

Rabinowicz, P.D., Citek, R., Budman, M.A., Nunberg, A., Bedell, J.A., Lakey, N., et al. 2005. Differential methylation of genes and repeats in land plants. Genome Res. **15**: 1431–1440.

Smith, D.B., and Flavell, R.B. 1975. Characterization of the wheat genome by renaturation kinetics. Chromosoma, **50**: 223–242.

Venter, J.C. 1993. Identification of new human receptor and transporter genes by high throughput cDNA (EST) sequencing. J. Pharm. Pharmacol. **45**(Suppl. 1): 355–60.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., SanMiguel, P., Lakey, N., Bedell, J., Yuan, Y., Budiman, M.A., Resnick, A., Van Aken, S., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C.M., and Quackenbush, J. 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science (Washington, D.C.), **302**: 2118–2120.

Wicker, T., Matthews, D.E., and Keller, B. 2002. TREP: a database for Triticeae repetitive elements. Trends Plant Sci. **7**: 561–562.

Yuan, Y., SanMiguel, P.J., and Bennetzen, J.L. 2003. High-Cot sequence analysis of the maize genome. Plant J. **34**: 249–255.

Zhang, Q., Arbuckle, J., and Wessler, S.R. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family *Heartbreaker* into genic regions of maize. Proc. Natl. Acad. Sci. U.S.A. **97**: 1160–1165.

Zimmerman, J.L., and Goldberg, R.B. 1977. DNA sequence organization in the genome of *Nicotiana tabacum*. Chromosoma, **59**: 227–252.

© 2005 NRC Canada