*The Plant Journal* (2004) **40**, 500–511 doi: 10.1111/j.1365-313X.2004.02228.x

# Sequence composition, organization, and evolution of the core Triticeae genome

**Wanlong Li[1], Peng Zhang[1], John P. Fellers[2], Bernd Friebe[1] and Bikram S. Gill[1,*]**
[1]*Department of Plant Pathology and Wheat Genetics Resource Center, Kansas State University, Manhattan, KS 66506, USA, and*
[2]*USDA-ARS, Kansas State University, Manhattan, KS 66506, USA*

Received 20 April 2004; revised 28 July 2004; accepted 17 August 2004.
[*]For correspondence (fax +1 785 532 5692; e-mail bsg@ksu.edu).

**Summary**

**We investigated the composition and the basis of genome expansion in the core Triticeae genome using *Aegilops tauschii*, the D-genome donor of bread wheat. We sequenced an unfiltered genomic shotgun (trs) and a methylation–filtration (tmf) library of *A. tauschii*, and analyzed wheat expressed sequence tags (ESTs) to estimate the expression of genes and transposable elements (TEs). The sampled D-genome sequences consisted of 91.6% repetitive elements, 2.5% known genes, and 5.9% low-copy sequences of unknown function. TEs constituted 68.2% of the D-genome compared with 50% in maize and 14% in rice. The DNA transposons constituted 13% of the D-genome compared with 2% in maize. TEs were methylated unevenly within and among elements and families, and most were transcribed which contributed to genome expansion in the core Triticeae genome. The copy number of a majority of repeat families increased gradually following polyploidization. Certain TE families occupied discrete chromosome territories. Nested insertions and illegitimate recombination occurred extensively between the TE families, and a majority of the TEs contained internal deletions. The GC content varied significantly among the three sequence sets examined ranging from 42% in tmf to 46% in trs and 52% in the EST. Based on enrichment of genic sequences, methylation–filtration offers one option, although not as efficient as in maize, for isolating gene-rich regions from the large genome of wheat.**

**Keywords: Triticeae, sequence composition, genome organization, genome expansion, DNA methylation, wheat.**

## Introduction

Unlike animals, genome size may vary a 1000-fold in higher plants, ranging from 125 Megabases (Mb) in *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) to 123 000 Mb in *Fritillaria assyriaca* (Bennett and Smith, 1976). Most of these large plant genomes consist of repeated sequences with a small fraction belonging to single or low-copy sequences (i.e. genes). The transposable elements (TEs) dispersed in the intergenic regions form dominant components of repeated sequences (reviewed by Bennetzen, 2000). At the biological level, the activities of TEs drive the structure, function, and evolution of genes and genomes. However, the large number and high-sequence identity of TEs in some plant genomes pose major obstacles for gene isolation, global sequencing, and functional studies.

Based on their mode of dispersion, retroelements and DNA transposons constitute two basic TE classes. In Class I TEs, the retroelements transpose via an RNA intermediate through a copy-and-paste mechanism. Class II TEs, the DNA transposons move from site-to-site through a cut-and-paste mechanism (Bennetzen, 2000). Class I TEs can be further subdivided into long-terminal repeat (LTR) and non-LTR type of TEs. LTR retrotransposons are the most redundant component of plant genomes, consisting of two groups, Ty1-*copia* and Ty3-*gypsy* (reviewed by Kumar and Bennetzen, 1999). High-copy retroelements nest in the intergenic regions (SanMiguel *et al.*, 1996), while low-copy retroelements associate with genes (Bennetzen, 2000; Kumar and Bennetzen, 1999). Unequal recombination between the two LTRs of a retrotransposon may cause the deletion of internal domains to form solo LTRs (Shirasu *et al.*, 2000) or terminal repeat retrotransposons in miniature (TRIM) (Witte *et al.*, 2001). Recombination

 

between two retrotransposons of the same family (Shirasu *et al.*, 2000) or illegitimate recombination between different elements (Devos *et al.*, 2002) may result in large deletions. These mechanisms are the counterforce against genome expansion.

The miniature inverted transposable elements (MITEs) are derivatives of DNA transposons. MITEs have no coding capacity yet retain terminal inverted repeats and a minimal subterminal portion (reviewed by Feschotte *et al.*, 2002). In the rice and maize genomes, MITEs are present at high copy numbers (1000s) and have been associated with genic regions (Bureau and Wessler, 1992, 1994a,b; Bureau *et al.*, 1996; Mao *et al.*, 2000; Zhang *et al.*, 2000).

In contrast to TEs, other repeats are organized tandemly and typically account for a small portion of a genome yet may have high copy numbers. Three major types of tandem repeat include ribosomal DNA, satellite DNA, and microsatellite or simple sequence repeats (SSRs), which have been investigated extensively.

Differential methylation is a characteristic feature of plant nuclear genomes in which the carbon 5 of cytosine residues, preferentially in CG dinucleotides and CNG trinucleotides (Gruenbaum *et al.*, 1981), is methylated. Approximately 30% of all cytosines are methylated (Adams and Burdon, 1985). Two outcomes of cytosine methylation are transcription inactivity and conversion of C to T. Typically, repeated DNA sequences are highly methylated whereas genes and low-copy sequences are hypomethylated. Methylation-sensitive restriction enzymes have been used for the efficient isolation of low-copy sequences. Recently, a new strategy was proposed to selectively clone and sequence genes and low-copy sequences using the *Escherichia coli* cytosine methylation restriction (Mcr) system (Rabinowicz *et al.*, 1999).

Among cereal crops, common, bread or hexaploid wheat (*Triticum aestivum* L., $2n = 6 \times = 42$, AABBDD), a member of the tribe Triticeae that includes approximately 300 species, and has the largest genome in this tribe at 16 000 Mb (Arumuganathan and Earle, 1991). Of the three genomes of polyploid wheat, the D-genome of diploid donor *Aegilops tauschii* Coss. ($2n = 2 \times = 14$, DD) has a relatively small genome at approximately 4000 Mb (Arumuganathan and Earle, 1991). Among the analyzed grass genomes, the rice genome (430 Mb) consists of 16% TEs in a proportion of 88% retroelements and 12% DNA transposons (Feng *et al.*, 2002; Mao *et al.*, 2000; Sasaki *et al.*, 2002). The maize genome at approximately 2500 Mb consists of 50% TEs, in a proportion of 98% retrotransposons and only 2% DNA transposons (Meyers *et al.*, 2001). Although there have been many studies on wheat genome organization, our overall understanding of the wheat genome is still incomplete. In this study, we report on the sequence composition, organization, and evolution of a core Triticeae genome.

## Results and discussion

To determine the sequence composition, methylation, and expression of major DNA elements within a core Triticeae genome, we used *A. tauschii* as a surrogate of a basic Triticeae genome. *A. tauschii* trs and tmf clones were sample sequenced and analyzed *in silico* using an EST database. Various DNA elements were further analyzed through hybridization on membranes and *in situ* on cytological preparations (see Experimental procedures).

### *Basic Triticeae genome composition*

The wheat D-genome is approximately 10-fold larger than that of rice and twice that of the maize genome (Arumuganathan and Earle, 1991). Previous studies on renaturation kinetics showed that over 80% of the wheat genome consists of repeated sequences with less than 12% as low-copy sequences (Smith and Flavell, 1975). To determine the composition of various DNA elements in more detail, we generated 3938 end sequences from random clones from a genomic shotgun trs library. Of which 15 sequences (0.4%) were mitochondrial and 94 (2.4%) contained chloroplast DNA. The remaining 3830 sequences were of nuclear origin and represented 2.911 Mb of sequence and 0.075% of the D-genome (Figure 1a and Table S1). Hereafter, the analyses are based on this sequence sample and we assume that they reflect the composition of the D-genome.

Repeated sequences constituted 91.6% of the genome, 83.4% of which consisted of known or unknown repeat families. Another 540 sequences were singletons with no match in the public databases or within the trs library. Dot–blot hybridization with labeled genomic DNA revealed that 315 singletons (8.2% of the genome) contained unknown repeated sequences. A total of 94 sequences (2.5%) represented genes encoding for known proteins, four encoded tRNAs and U3 RNA, and two encoded putative pseudogenes (Table S1). The remaining 225 sequences (5.9%) represented low-copy DNA of unknown function.

Retroelements constituted 53.5% of the repeated fraction representing 39 (12 Ty1-*copia*, 20 Ty3-*gypsy*, five LINE, and two TRIM) families. LTR retrotransposons were predominant, accounting for 95.7% of retroelements and 51.2% of the genome. Overall, Ty3-*gypsy* made up to 35.2% and Ty1-*copia* 15.8% of the genome. DNA transposons and related sequences constituted 13.3% of the genome, the second largest repeated fraction. The CACTA superfamily alone accounted for 12.3% of the genome. Four long inverted transposable element (LITE) families and 11 MITE families were encountered. As a result of their small size (90–240 bp), MITEs did not constitute a major fraction of the genome, but their estimated copy number was estimated at approximately 84 000 MITEs per genome.

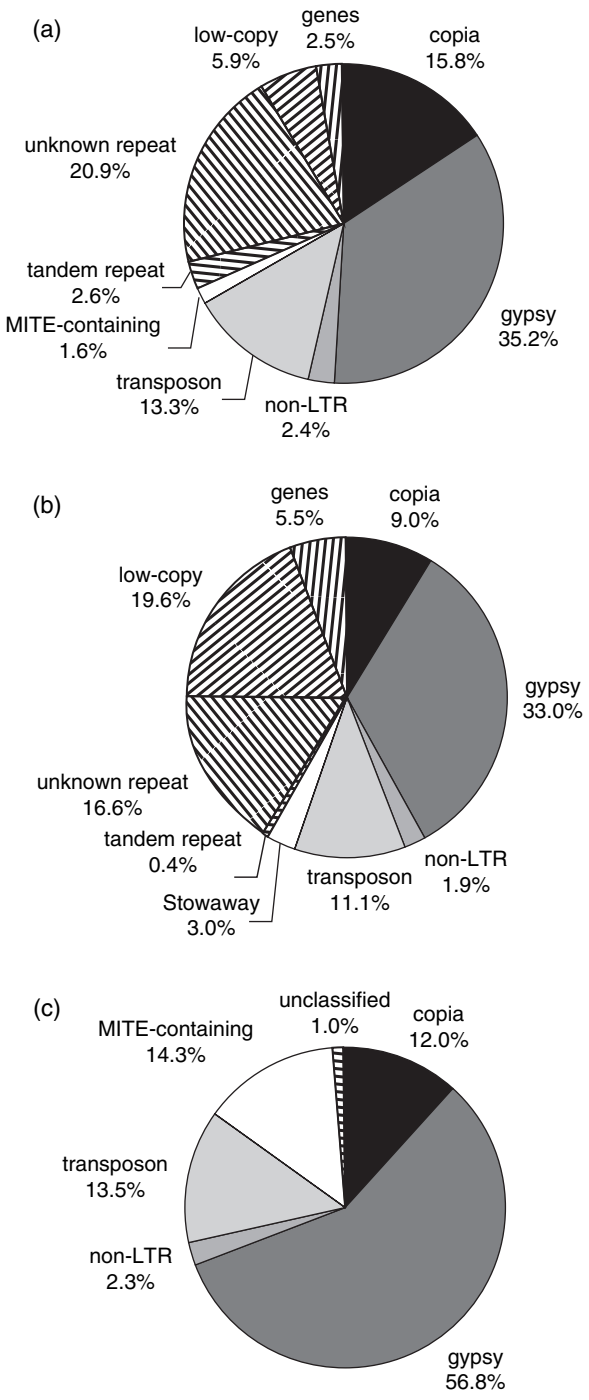


**Figure 1.** Diagrams showing sequence composition of genomic libraries and wheat expressed sequence tags (ESTs).
(a) Unfiltered *Aegilops tauschii* shotgun library trs.
(b) Methylation-filtered *A. tauschii* library tmf.
(c) Transposable elements (TEs) from the wheat EST database TAGI, of which protein-coding genes account for 96.7% and TEs account 3.3%.

Tandem repeats constituted less than 3% of the genome. Two major families, the 18S-26S ribosomal and the satellite *Afa*, a part of CACTA-type transposon Caspar (Wicker *et al.*, 2003), each accounted for approximately 1% of the genome. Other families, such as Type 5, Type 7, and one with weak similarity to *Tail*, constituted less than 1% of the genome. In addition, 17 sequences (0.4% of genome) showed similarities to rye hypervariable sequences (GenBank accession numbers AF153212, AF153213, and AF153326).

### *Repeated families show variable rates of methylation and transcription*

A second, small-insert methylation–filtration library, tmf, was constructed using *A. tauschii* nuclear DNA and represented the hypomethylated portion of the genome. We generated 1070 sequences, of which 13.8% (148 sequences) were of chloroplast and 0.84% (nine sequences) of mitochondrial DNA origin. The remaining 913 sequences contained DNA from the nuclear genome and represented 556 kb of sequence. The fraction of repeated sequences (74.8%) in the tmf library was significantly lower ($P < 10^{-45}$) than the trs library (91.6%), indicating that repeated sequences are largely methylated (Figure 1b and Table S2). Ty1-*copia* and tandem repeats decreased most significantly in the tmf library. Among different families, three categories of repeats were recognized (Figure 2 and Table S2). Families of the first category were present in significantly lower frequency in the tmf library than in the trs library. This category included 18 repeat families. The second category (15 repeat families) maintained similar frequencies in both libraries. The third category of repeats (Caspar, Sabrina, Deimos, and MITE superfamily Stowaway) showed an increase in the tmf library. The frequency of genes increased by twofold and low-copy sequences by threefold in the tmf library.

Although few ESTs are available for *A. tauschii*, a large body of ESTs have been developed in bread wheat. Considering that *A. tauschii* is the D-genome donor and its close phylogenetic relationship to the A and B genomes of bread wheat, we investigated the TE transcription in wheat by alignments of TE sequences with wheat ESTs. TE transcripts were identified in the TIGR (The Institute for Genomic Research) wheat gene index (TAGI) database by a BLASTN search against cereal TE sequences. We found 2369 singletons or tentative contigs (TCs) assembled from 3267 ESTs at an *E*-value of $10^{-20}$ or less (Figure 1c and Table S3). Additionally, 245 singletons/TCs, grouped from 298 ESTs, were identified by a TBLAST search using TE protein sequences as queries. Overall, 2.4% of singletons/TCs or 0.9% of ESTs (2614 singletons/TCs representing 3565 ESTs in the TAGI) were similar to TEs. The frequency of retrotransposons and DNA transposons in ESTs were positively correlated with their relative abundance in the trs ($r = 0.765$, $P < 4 \times 10^{-8}$) and the tmf ($r = 0.717$, $P < 4 \times 10^{-6}$) libraries and their copy number in D-genome ($r = 0.747$, $P < 2 \times 10^{-7}$). The TE expression level differed as seen from the depth of TCs, from 2 to 27 ESTs. The average expression rate was

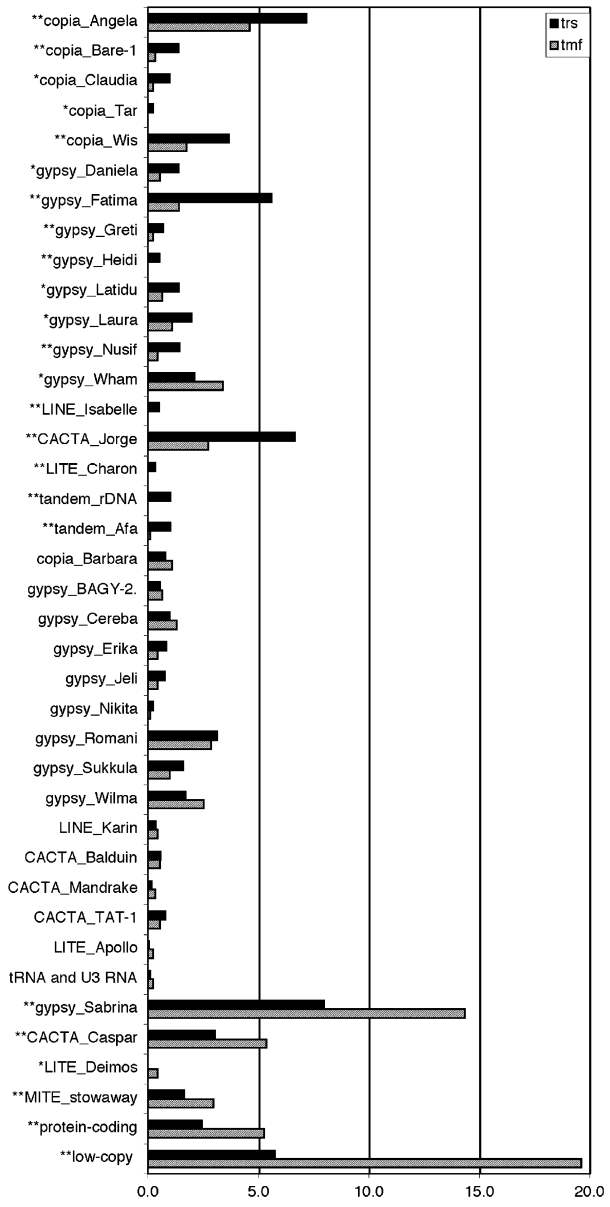


**Figure 2.** Comparison of sequence fractions of transposable element (TE) families in the trs and tmf libraries. The symbols '*' and '**' preceding the TE family names indicate significance levels of $P < 0.05$ and $P < 0.01$, respectively.

evaluated using the ratio of EST number in a TE family to the copy number of the TE family in the D-genome. Transcription levels were the highest for Ty3-*gypsy* elements (1.7%), lowest for Ty1-*copia* elements (0.6%), and moderate for non-LTR (1.4%) and CACTA elements (1.2%). The highly transcribed were the Ty3-*gypsy* element Latidu and the CACTA element TAT-1 (Table S3).

Of 2080 MITEs recovered from the TE search of the TAGI, 97% belonged to the Stowaway superfamily with only 3% Tourist-like in origin. Considering their small size (80–300 bp), if any MITEs were transcribed as an independent transcription unit, they would most likely have been filtered out by size fractionation during cDNA library construction. Therefore, the MITE-containing expressed sequences most likely represent host genes with MITEs inserted in their exons.

### *Nested Insertions and deletions of TEs*

During BLAST searches against the Triticeae repeat (TREP) database (Wicker *et al.*, 2002), we found many bipartite, and some tripartite or tetrapartite chimeric clones in which portions of each clone shared homology with different repeat families. A chimera was unlikely to be formed during ligation as the inserts were treated with calf intestinal phosphatase to remove the phosphate group from their 5′ termini and topoisomerase I was bound to the ends of the linearized vector. Thus, the mosaic sequences probably arose either from nested insertions or illegitimate recombination among the TEs. We identified 169 junctions in 163 chimeric clones from the trs and tmf libraries; 108 junctions arose from nested insertions because one of the two elements had an intact 3′ or 5′ terminus and 33 originated by illegitimate recombination because broken ends were closely connected. In four junctions, two intact ends were close, perhaps because of physical proximity. The structure was not clear for the remaining 24 junctions because of the presence of 100–200 bp of intervening DNA. Of the 169 junctions, 148 (100 from nested insertion, 24 from illegitimate recombination, three because of close connection, and 21 with unclear structures) were found in the 2.9 Mb sequence from the trs library. The estimated frequency of approximately 50 TE rearrangements/Mb indicated that nested insertions and illegitimate recombination has occurred extensively in the D-genome.

The distribution of insertion junctions among the repeat groups was not random ($P < 2.4 \times 10^{-8}$). As the ratios of Observe/Expect indicated, the Preferred insertions of CACTA into Ty1-*copia* and Ty1-*copia* into CACTA were detected while the preferred targets of a Ty3-*gypsy* family were CACTA elements (Table 1).

Long-terminal repeats are essential for retrotransposition. Analysis of insertion-target junctions can provide important clues as to whether the LTRs were targeted for insertion. Seventy-nine independent insertions occurred in known retrotransposons, of which 27 (34.2%) were within LTRs. LTRs constitute approximately 40% of the element length indicating that targeting of insertions was random between the LTR and other parts of a retrotransposon ($P = 0.45$). We found an identical insertion-target junction in two independent clones, in which a Fatima element inserted in the LTR of a Sabrina element, 297 bp from the 5′ terminus, indicating that the entire length of an LTR is not required for retrotransposition.

**Table 1** Nested insertion of transposable elements and target preference in *Aegilops tauschii*

| Insertion | Target | Observe | Expect | Observe/Expect |
|---|---|---|---|---|
| CACTA | CACTA | 0 | 5.53 | 0.00 |
| CACTA | Copia | 13 | 3.43 | 3.79 |
| CACTA | Gypsy | 7 | 7.95 | 0.88 |
| Copia | CACTA | 9 | 3.43 | 2.62 |
| Copia | Copia | 1 | 8.49 | 0.12 |
| Copia | Gypsy | 11 | 9.86 | 1.12 |
| Gypsy | CACTA | 16 | 7.95 | 2.01 |
| Gypsy | Copia | 10 | 9.86 | 1.01 |
| Gypsy | Gypsy | 35 | 45.51 | 0.77 |

Defective TE elements can arise during transposition or retrotransposition by internal deletion (see reviews by Bennetzen, 2000; Feschotte *et al.*, 2002; Kumar and Bennetzen, 1999) or unequal intrastrand recombination between the 5′ and 3′ LTRs of a single element (Shirasu *et al.*, 2000). To measure the frequency of internal deletions, we compared the ratio of the coding region to the non-coding region of each transposable element family present in the trs library to the ratio of the complete elements from the TREP database. A chi-square test was made on 21 TE families for which intact elements were available in the TREP database or could be assembled using truncated fragments. Twenty families tended to have a deletion in the coding region, of which 13 high-copy TE families were significant at $P = 0.005$ (Table 2). Inga was the only family in which the coding region was longer than expected as the complete copy of Inga in the TREP database was identified from barley (Rostoks *et al.*, 2002).

In the trs library, nine Nikita, 61 Sukkula, and 255 Jorge sequences appeared to be defective and non-autonomous as they did not show sequence similarity to any retrotransposon- or transposon-related proteins by BLASTX search against the TREP and NCBI protein databases. Intact members of these families have been identified in BAC sequences but did not encode any proteins (Shirasu *et al.*, 2000; Wicker *et al.*, 2003).

### *TE families dominate or share chromosomal territories*

We visualized the chromosomal location of the repeated sequences by fluorescence *in situ* hybridization (FISH) in *A. tauschii* and hexaploid wheat (Figure 3). Four distribution patterns were observed: localized, dispersed, localized-and-dispersed, and polar-dispersed. Selected examples are given in Figure 3 and data on all others are summarized in Table S4.

Our results demonstrate that TE families either have their own chromosomal territories or share them with other families. The distribution pattern of a TE along a chromosome seems to be related to its copy number and reflects the insertion spectrum. Interestingly, all the localized TEs discovered to date have relatively low redundancy and belong to the Ty3-*gypsy* group (Table S4). These localized TEs may have high insertion specificity or alternatively, the wheat genome is under selection pressure for localized insertion. The highly redundant TEs were dispersed along the chromosomes except for the centromeres and nucleolar organizer regions while dispersed TEs appear to have lower insertion specificity.

**Table 2** Variation in extent of internal deletion of wheat TEs

| | Coding region (kb) | | Non-coding region (kb) | | | |
|---|---|---|---|---|---|---|
| Family | Observe | Expect | Observe | Expect | Chi-sq | Significance |
| Angela | 31.999 | 79.642 | 149.098 | 101.455 | 49.81 | $P < 1.7 \times 10^{-12}$ |
| Barbara | 1.138 | 9.263 | 14.354 | 6.229 | 15.61 | $P < 7.8 \times 10^{-5}$ |
| Bare | 13.115 | 14.033 | 19.855 | 18.937 | 0.02 | $P < 0.88754$ |
| Inga | 1.517 | 1.181 | 2.716 | 3.052 | 0.03 | $P < 0.86249$ |
| Tar | 1.172 | 2.262 | 2.805 | 1.714 | 0.36 | $P < 0.54851$ |
| Wis | 13.054 | 37.011 | 71.286 | 47.329 | 26.49 | $P < 2.7 \times 10^{-7}$ |
| BAGY-2 | 1.501 | 3.102 | 6.366 | 4.765 | 0.65 | $P < 0.42011$ |
| Cereba | 3.919 | 12.215 | 13.836 | 5.540 | 15.95 | $P < 6.5 \times 10^{-5}$ |
| Daniela | 6.095 | 12.905 | 15.460 | 8.650 | 7.69 | $P < 0.005553$ |
| Erika | 2.608 | 5.341 | 14.220 | 11.487 | 1.37 | $P < 0.24181$ |
| Fatima | 21.274 | 35.203 | 64.005 | 50.076 | 8.72 | $P < 0.00315$ |
| Jeli | 3.491 | 3.788 | 6.216 | 5.919 | 0.02 | $P < 0.88754$ |
| Latidu | 0.867 | 7.962 | 17.084 | 9.989 | 9.82 | $P < 0.00173$ |
| Laura | 2.046 | 12.893 | 40.762 | 29.914 | 11.88 | $P < 0.00057$ |
| Romani | 10.391 | 24.935 | 54.125 | 39.58 | 12.89 | $P < 0.00034$ |
| Sabrina | 3.337 | 39.078 | 151.357 | 115.615 | 42.52 | $P < 7.0 \times 10^{-11}$ |
| Wham | 3.200 | 11.653 | 38.334 | 29.881 | 7.54 | $P < 0.00604$ |
| Wilma | 4.808 | 10.49 | 28.889 | 23.207 | 3.72 | $P < 0.05377$ |
| Balduin | 2.399 | 8.774 | 12.323 | 5.948 | 9.74 | $P < 0.00181$ |
| Caspar | 9.341 | 22.043 | 41.108 | 28.406 | 12.00 | $P < 0.00054$ |
| TAT | 2.129 | 3.484 | 7.260 | 5.905 | 0.33 | $P < 0.56566$ |

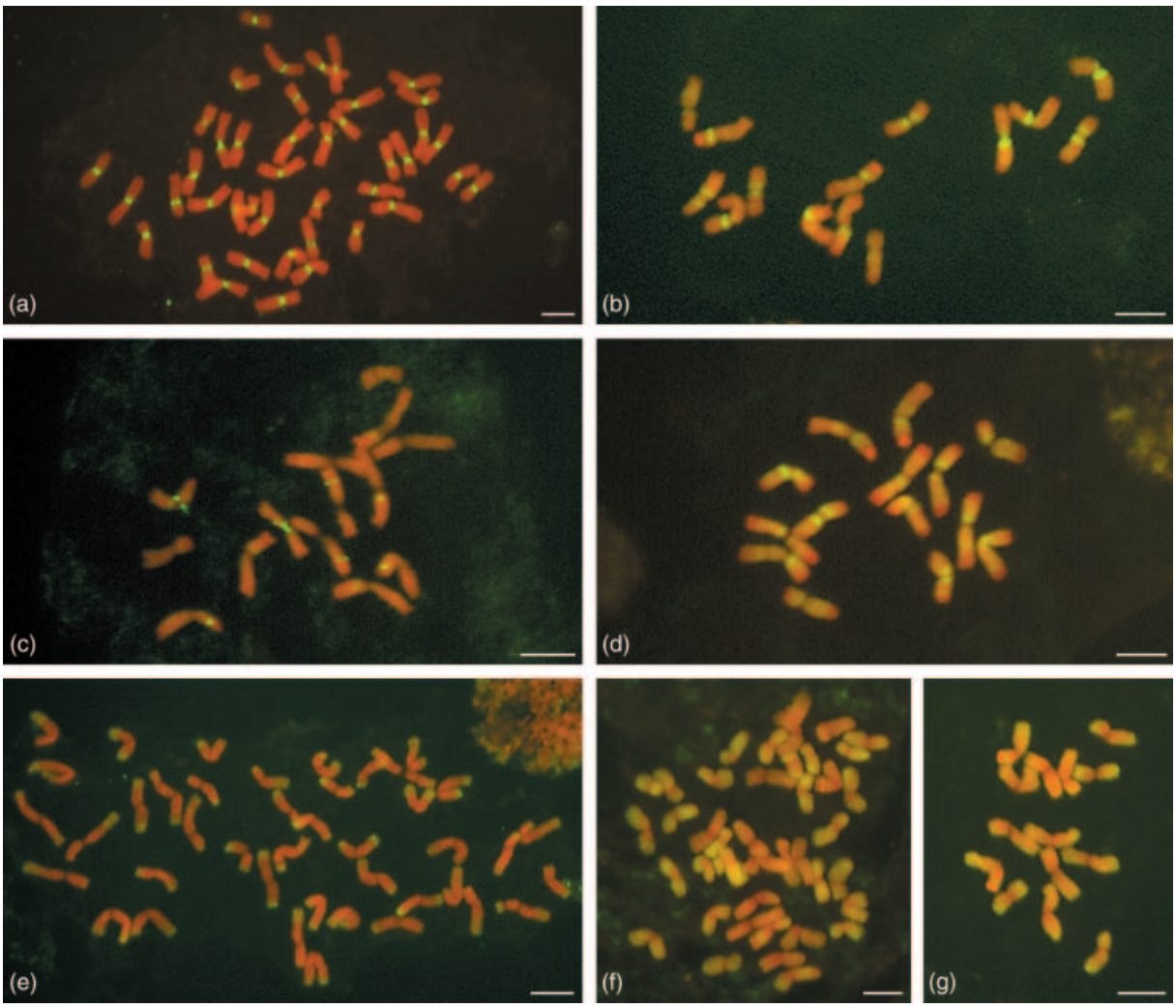


**Figure 3.** Fluorescence *in situ* hybridization (FISH) pattern of seven transposable elements.
FISH pattern of seven transposable elements, Cereba (a), Laura (b), Romani (c), Erika (d), Caspar (e), Angela (f), and Wis (g), on mitotic metaphase chromosomes of *Triticum aestivum* cv. Chinese Spring (CS) (a, e, and f) and *Aegilops tauschii* (b, c, d, and g). Plasmid clones containing Romani, Erika, Caspar, Angela, and Wis were labeled with biotin-14-dATP and detected with fluorescein-avidin DN, which is visualized by yellowish-green fluorescence. Chromosomes were counterstained with propidium iodide and fluorescence red.
(a) Cereba hybridized to the centromeres of all the CS chromosomes.
(b) Laura hybridized mainly to the region around the centromeres of *A. tauschii* chromosomes.
(c) Romani localized to sites on five chromosome pairs of *A. tauschii*.
(d) Erika hybridized stronger in the proximal region of *A. tauschii* than in the distal region.
(e) Caspar hybridized to the subtelomeric regions of all the CS chromosomes.
(f) The long-terminal repeat (LTR) portion of Angela hybridized strongly to all the CS chromosomes except for the centromeric regions.
(g) The LTR portion of Wis hybridized to all the chromosomes of *A. tauschii* and detected much stronger signals in distal than in the proximal region. Scale bars equal 10 μm.

The genome appears to control the retroelement invasion at the level of transcription and/or integration. A high correlation between the level of transcription and amplification was observed in Tos17 (Hirochika *et al.*, 1996). However, Bare-1 was highly transcribed in barley, but its retrotransposition was rarely detected (Suoniemi *et al.*, 1997). In our results, Latidu was highly expressed but only accounted <1% of the D-genome. Therefore, integration, rather than transcription, is likely the major restriction factor for retrotransposition in the large genomes of wheat and barley. We can further infer that the insertion spectrum of a retroelement is a determinant of its copy number in a large genome. The narrower the spectrum, the lower the copy number, and vice versa. Studies in yeast showed that insertion specificity was determined by an interaction between the integration complex and the proteins associated with the target sites (Kirchner *et al.*, 1995). Latidu and Romani detected both localized and dispersed signals,

implying that this family has a broad insertion spectrum and higher affinity to some specific genome sites.

Caspar, Erika and Laura displayed gradient signals along chromosome arms suggesting that these TEs were driven by a polar genome force, most likely recombination. The increased copy number of Caspar elements toward the telomeres appears to be related to a high rate of recombination in the distal regions. Gene conversion is a major means to repair gaps left by excision of CACTA transposons (Bennetzen, 2000). The stronger signals of Erika and Laura elements towards the centromere may be related to low recombination in the centromeric regions, suggesting that they are potential recombination suppressors, excluded from the distal regions by negative selection and may play an important role in stabilizing centromere function.

### *Copy-number variation and polyploidization*

We estimated the copy number of TE families using dot–blot hybridization for a range of genotypes of diploid and polyploid wheat, rye, barley, oats, and rice. Wheat TEs did not hybridize to oats and rice. Seven of the 17 TE families did not hybridize or detected a much lower copy number in barley than in *Triticum* or *Aegilops* species (Table S5). Bare-1 is the largest TE family in barley (approximately 50 000 copies) (Suoniemi *et al.*, 1997), but only 100 copies were detected by wheat Bare-1 probe on barley indicating that TEs are a fast-evolving component of a genome. For most TE families, the copy numbers estimated for rye are within the range of diploid wheat and *Aegilops* species, and only the Fatima element was not detected in rye. The D-genome tandem repeat *Afa*, a component of CACTA transposon Caspar (Wicker *et al.*, 2003), was detected a higher copy number in barley than in rye.

Polyploid wheat had a higher copy number of TEs than do diploid wheat and *Aegilops* species. Among the tetraploid wheat, *Triticum turgidum* L. had higher copy numbers for most TE families than *Triticum timopheevii* Zhuk (Zhuk). Comparing the copy numbers in polyploid wheat and the sums of their donor species showed variation of copy number after polyploidization. *Triticum turgidum* and *T. timopheevii* had higher copy numbers for most repeat families than the sums of their parental species *Triticum urartu* Tumanian ex Gandilyan and *A. speltoides* Tausch. *T. aestivum* showed an increase in copy number for five TE families and decrease in seven families. No significant change in copy number was detected in the synthetic amphiploids AASS (*A. speltoides* × *T. urartu*), $A^m$ $A^m$ $S^{sh}$ $S^{sh}$ (*A. sharonensis* × *Triticum monococcum*), SSDD (*A. speltoides* × *A. tauschii*) and AABBDD (*T. turgidum* × *A. tauschii*) (data not shown). These results suggest that copy-number changes do not occur immediately and most probably arise slowly and progressively after polyploidization.

**Table 3** GC (%) content of random shotgun (trs) and methylation–filtration (tmf) libraries of *Aegilops tauschii* and wheat expressed sequence tag sequences from TAGI database sequences

| | trs | tmf | TAGI |
|---|---|---|---|
| Overall | 46.03 | 41.62 | 52.09 |
| Genes | 48.23 | 47.94 | 52.21 |
| Low copy | 45.96 | 41.94 | – |
| *Copia* | 44.14 | 42.59 | 45.75 |
| *Gypsy* | 46.54 | 39.85 | 48.43 |
| Non-LTR | 44.62 | 42.70 | 50.72 |
| CACTA | 46.34 | 40.51 | 47.85 |
| MITE | 41.77 | 42.10 | 45.48 |
| Satellites | 40.74 | – | – |

### *Nucleotide composition varies significantly among the genome components*

The tmf sequences had the lowest GC content (41.6%), while the trs sequences were intermediate (46.0%) and the TAGI sequences had the highest GC content (52.1%) (Table 3). The GC content for genes and MITEs did not change significantly among trs, tmf, and TAGI libraries. The GC content of low-copy sequences and TEs in the tmf library was significantly lower than in the trs library. The decrease of GC content in the tmf library was caused by the filtration of the cytosine-methylated DNA. In the trs library, GC content was highly correlated with the frequency of dinucleotides CG ($r = 0.8589$, $P = 0$) and trinucleotides CNG ($r = 0.8681$, $P = 0$), which are the targets of cytosine methylation.

Normalized frequencies of dinucleotides and trinucleotides were estimated for collective sequences from the trs, tmf, and the TAGI datasets (Figure 4a). Significant depletion was observed for TA, CG, and AC/GT from all the three-sequence collections as well as the trinucleotide CNG from the trs and tmf libraries. In contrast, an obvious increase was observed for CA/TG, AA/TT, and CC/GG in all the three-sequence collections, and for GC from the tmf and the TAGI libraries. The di- and trinucleotide frequencies in CACTA and Ty3-*gypsy* were similar to the collective sequences but GA/TC increased and GC decreased in Ty1-*copia* (Figure 4b–d). For Ty1-*copia*, both CG and CNG frequencies were lower by approximately 5% in the tmf than in the trs library (Figure 4b). The difference was twofold greater for CACTA and Ty3-*gypsy* (Figure 4c,d), suggesting that CACTA and Ty3-*gypsy* elements had twice the level of DNA methylation than Ty1-*copia*. The increased frequency of trinucleotide CNA/TNG for Ty1-*copia* and Ty3-gypsy groups in the tmf library suggested a higher level of C to T transition.

Although CACTA and Ty3-*gypsy* were methylated at a higher level than Ty1-*copia*, the transcription levels as estimated from EST data were higher for Ty3-*gypsy* and CACTA elements. Transcription level for a TE family is a function of the transcription rate of an individual element

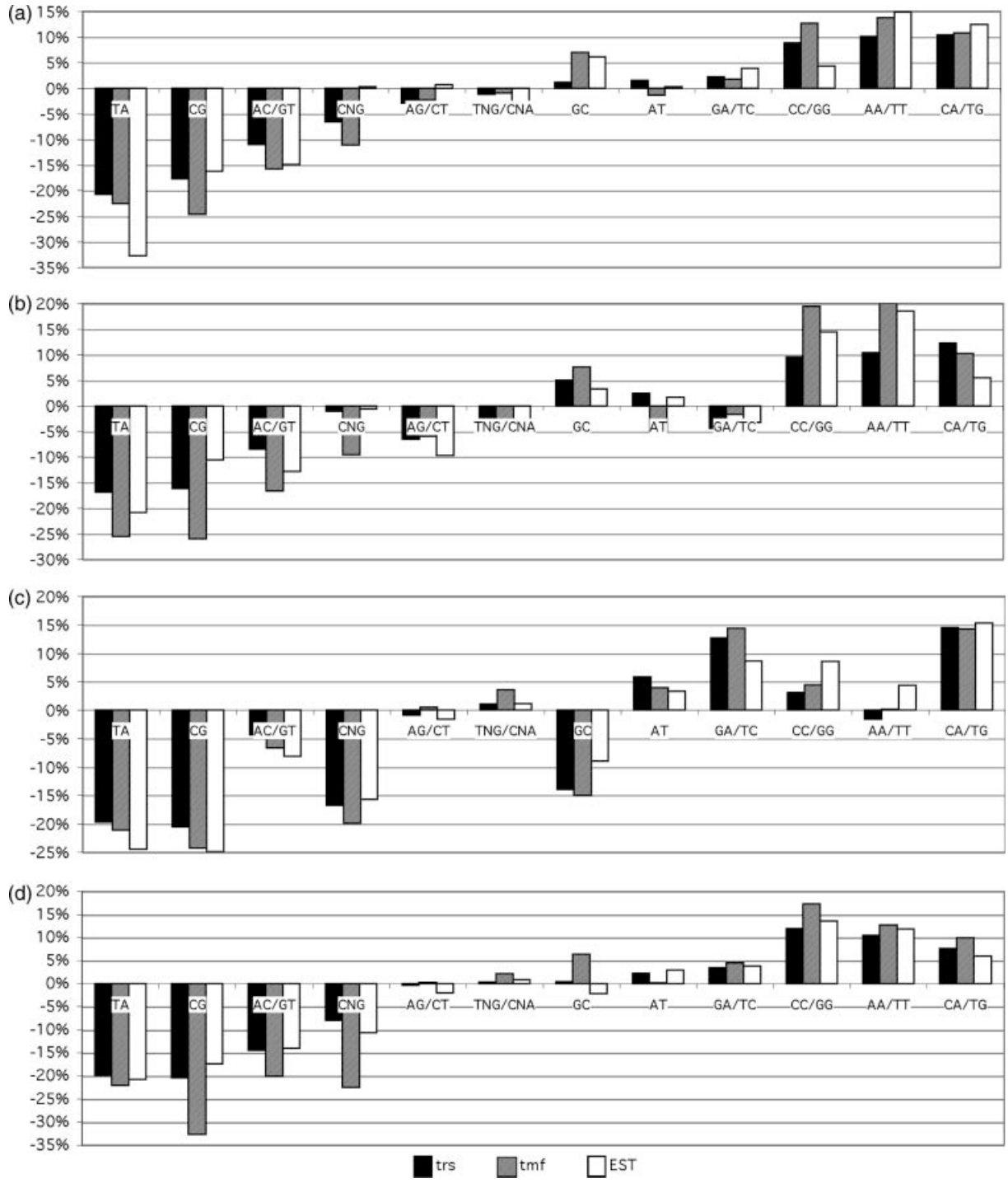


**Figure 4.** Normalized frequencies of dinucleotides and trinucleotides.
(a) For all sequences of the trs and tmf genomic libraries and wheat expressed sequence tags.
(b) For CACTA elements.
(c) For Ty1-*copia* elements.
(d) for Ty3-*gypsy* elements.

and the number of active (hypomethylated) elements. Inspection of the CG and CNG frequencies in the tmf library showed that variation in frequencies was approximately fourfold higher for Ty3-*gypsy* and CACTA than for Ty1-*copia* elements, suggesting that methylation occurred more homogeneously among Ty1-*copia* elements than among

Ty3-*gypsy* and CACTA elements. Therefore, Ty1-*copia* may have fewer hypomethylated copies than expected.

### *Gene content*

BLASTX (translated alignment) revealed that 94 sequences (2.5%) from the trs library showed sequence similarities to various non-TE proteins (Table S6). Forty-eight sequences from the tmf library were found homologous to protein-coding genes at the amino acid sequence level (Table S6). Of the 142 genes identified in the trs and tmf sequences, 50 encode for proteins with unknown function, designated as hypothetical, putative, and unknown proteins. Twelve genes encode disease-resistance proteins, with four containing NBS-LRR domains, two similar to *Xa21*, two similar to *Cf2*/*Cf5*, one similar to potato blight resistance protein RGA4, and the others similar to receptor-like kinases or LRR transmembrane protein kinases.

In addition to the protein-coding sequences identified, 225 low-copy sequences (5.9%) were revealed as they did not show signal in hybridization with genomic DNA. These low-copy sequences are potentially unknown genes or genic components, promotors, introns, and other regulatory elements.

### *Why is basic Triticeae genome so large?*

Comparison of the genome sequence composition of wheat and maize may provide some insight into the genome inflation. First, DNA transposons fueled wheat genome obesity. DNA transposons account for just 1.22% of the maize genome (Meyers *et al.*, 2001), but account for 13.3% of the wheat genome, with an increase of approximately 500 Mb in genome size. Among DNA transposons, the CACTA superfamily contributed 12.3% to genome size. This superfamily is very diverse in sequence composition (Wicker *et al.*, 2003) and is highly redundant in other temperate grasses and cereals (Langdon *et al.*, 2003). Transposition of DNA transposons occurs by excision and insertion. The gap left at the site of excision is repaired by simple ligation or gene conversion using a homologue or sister chromatid as template (reviewed by Bennetzen, 2000). Repair by gene conversion will lead to increase in copy number. The high copy number of CACTA elements implies that gene-conversion was the major strategy for gap repair in wheat and the Triticeae.

Second, TEs are transcribed at a higher level in the wheat than in the maize genome. Meyers *et al.* (2001) recovered 56 (0.014%) retrotransposons from 407 000 maize ESTs and found that only low-copy elements were expressed. Echenique *et al.* (2002) found only 0.15% of wheat ESTs were significantly similar to the retrotransposons. In our data, 0.9% of the wheat transcriptome is similar to retroelements and transposons. Unlike maize, the amount of expressed TEs in wheat is highly correlated with their redundancy in the genome. Almost all of the known TEs were found in the wheat EST collection. High TE expression directly and/or indirectly accelerated the pace of wheat genome inflation. The high TE expression may be associated with overall low level of DNA methylation in wheat. The repeated sequences accounted for three-fourths of the tmf library when compared with approximately one quarter of methylation–filtration library of maize (Meyers *et al.*, 2001; Rabinowicz *et al.*, 1999).

Third, genome inflation may be related to the mechanism of nested insertions. In maize, over 50% of nested insertions target LTRs of retrotransposons and inactivate them (Bennetzen, 2000; SanMiguel *et al.*, 1996). In wheat, LTRs are not a hot spot for nested insertions and not all the insertions in LTRs impair further retrotransposition. Amplification of targeted retrotransposons was observed. As a result, retrotransposition frequency is much higher in wheat when compared with maize and has contributed to the genome inflation.

Other factors affecting wheat genome size were also noted in the present study. Internal deletions and illegitimate recombination act as counter forces to contract wheat genome inflation, but polyploidization seemingly provoked the amplification of TEs, although at a reduced rate.

### *Wheat genome sequencing*

The large amount of repetitive sequences poses a big challenge for sequencing the wheat genome. To sequence a majority of the genes yet minimize repeat contamination at a reasonable cost, two methods have been proposed: methylation–filtration (Rabinowicz *et al.*, 1999) and $C_0t$-based cloning and sequencing (CBCS) (Peterson *et al.*, 2002). In our study, the fraction of genes, low-copy sequences and MITEs were enriched by over twofold in the tmf library when compared with the trs library. Although tandem repeats such as rDNA and *Afa* satellite DNA were efficiently filtered out from the tmf library, it still contained 74.8% TEs and unknown repeats, which is comparable to the total repeat content of the maize genome. Furthermore, the low-copy sequences recovered from the tmf library were strongly biased in GC content. Compared with the filtration efficiency in maize (Palmer *et al.*, 2003; Whitelaw *et al.*, 2003), methylation–filtration does not work as efficiently in wheat.

$C_0t$-based cloning and sequencing takes advantage of the differential renaturation rate of sequences of different redundancy in a genome and is independent of transcription and methylation patterns. CBCS was successful in isolating new low-copy sequences from the sorghum genome (Peterson *et al.*, 2002). The wheat genome has been thoroughly analyzed by reassociation kinetics (Britten and Kohne, 1968; Smith and Flavell, 1975). Recently, we sample

sequenced a small-insert genomic library of bread wheat generated from the slow-association fraction collected at $C_0t$ >1600. The preliminary result indicated that the repeat content was reduced by approximately fourfold and the gene content was enriched by approximately 10-fold (D. Lamoureux, D. Peterson, W. Li, J. Fellers, and B. S. Gill, unpubl. data). As current coverage of the wheat EST collection is approximately 60%, low-copy sequences from CBCS would be helpful for wheat gene discovery. However, some inherent disadvantages exist, such as the under-representation of gene families and small insert size that may cause chimeras during assembly of sequences from the three homoeologous genomes of hexaploid wheat.

## Experimental procedures

### *Plant materials*

The plant materials used in this research are listed in Table S7.

### *Library construction*

The trs and tmf genomic libraries were constructed from accession AL8/78 of *A. tauschii*. For the trs library, total genomic DNA was isolated from leaf tissue as described by Faris *et al.* (2000). After digestion with RNase, DNA was extracted once with phenol–chloroform, twice with chloroform, precipitated with alcohol and resuspended in water. The DNA was nebulized and fragments were size fractionated on an agarose gel. Fragments ranging in size from 0.8 to 1.2 kb were recovered from agarose gel and cloned into PCR®4Blunt-TOP® vector (Invitrogen Life Technologies, Carlsbad, CA, USA). The ligation mix was transformed into *E. coli* strain DH10B using electroporation with a Cell-Porator® following the manufacturer's instructions (Invitrogen). White colonies were selected from LB agar plates containing X-gal, IPTG, and carbenicilin and transferred to 384-well plates containing LB freezer medium with carbenicillin. After overnight growth, the plates were stored in a –80°C freezer. For the tmf library, nuclei were isolated from young leaves following the method described by Zhang *et al.* (1995). Subsequent library construction was the same as described above, with the exception that the ligation was transformed into a Mcr+ strain of *E. coli* DH5α (McrA, McrBC, and Mrr) which restricts methylated DNA and thus, the recovered clones should contain hypomethylated DNA.

### *Sequence manipulation*

All clones were single-pass sequenced with an ABI 3700 sequencer (Applied Biosystems, Foster City, CA, USA). Raw sequence data were analyzed with the base-calling program 'PHRED' (Ewing and Green, 1998). The average length called by PHRED was 787 bases and the average quality score was 37.3 per base. Vector sequences were trimmed using cross match with settings of minmatch 12 and minscore 20. The cleaned, single-pass sequences were subjected to a batch search using BLASTN and BLASTX (Altschul *et al.*, 1997) against the non-redundant and protein databases at NCBI. BLASTN was performed against complete set of cereal repeat sequence and BLASTX against hypothetical proteins of TEs at the TREP web site http://wheat.pw.usda.gov/ITMI/Repeats/blastrepeats3.html (Wicker *et al.*, 2002). Clones were categorized based on their similarities with known sequences with the expectation score of $10^{-5}$ or better. Unknown sequences were grouped using program CAP ASSEMBLER (Huang and Madan, 1999) at the web site http://bio.ifom-firc.it/ASSEMBLY/assemble.html. All analyses were performed using default settings. In total, 109 782 unique sequences (mainly ESTs) were downloaded from wheat gene index database TAGI of TIGR (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=wheat, release 6.0). WU-BLASTN and TBLAST searches (http://blast.wustl.edu) were performed against the TAGI using the complete set of cereal repeated sequences and hypothetical protein sequences of TEs from the TREP database as queries. To transform the score and *E*-value into corresponding parameters of NCBI BLAST, the retrieved sequences from WU-BLASTN and TBLAST were used as queries to search the TREP database.

Computer programs were written to total matched length in BLASTN and BLASTX, for calculating base composition and di- and trinucleotide frequencies and search for microsatellites. The normalized frequency of a dinucleotide was estimated as [(observed number) – (expected number)]/(expect number). Expected number = (frequency of base 1) × (frequency of base 2) × (total nucleotides). To test for insertion specificity for the 108 insertion-target junctions, the expected numbers and the frequencies of junctions were estimated from the frequencies of the 39 known repeat families in the mosaic clones based on their genome fractions: expected frequency = (frequency of insertion family) × (frequency of target family). All the statistical analyses were conducted online (http://fonsg3.let.uva.nl/Service/Statistics.html).

### *FISH analysis*

Pretreatment of root tips, preparation of chromosome spread, slide pretreatment, and denaturation were performed as described in Zhang *et al.* (2001). Plasmid DNA was isolated using a QIAprep Spin Miniprep Kit (Qiagen, Valencia, CA, USA) and 1 μg of plasmid DNA was labeled with biotin-14-dATP using BioNick Labeling System (Invitrogen). FISH was carried out according Zhang *et al.* (2004).

### *Copy number*

Copy number was estimated by two methods. The first method was based on the fraction of the individual repeat family in the available sequence datasets followed by determination of copy number in *A. tauschii* via [(genome fraction) × 4024 000/(size of complete element in kb)]. The second method for wheat and related species, copy number was estimated using dot blot hybridization. For each accession, 1000 C of genomic DNA in 5 μl was blotted onto $N^+$ membrane (Amersham Bioscience, Piscataway, NJ, USA). Plasmid DNA was diluted in three series of $10^5$–$10^6$, $10^6$–$10^7$, and $10^7$–$10^8$ copies, and blotted onto a membrane. The membrane was allowed to dry overnight and treated with 0.4 M NaOH for 5 min and 2 × SSC for 5 min. Inserts were released with *Eco*RI and separated by agarose gel electrophoresis and purified using a NucleoTrap® Gel Extraction Kit (BD Science, Palo Alto, CA). Probe labeling, filter hybridization and washing were according to Faris *et al.* (2000).

## Acknowledgements

We thank Angie Matthews for technical help in sequencing the genomic clones, Zhigang Xie for computer assistance in sequence analyses, Dr Moshe Feldman for providing seeds of parents and

amphiploid between *T. monococcum* and *A. sharonensis*, and Dr Jan Dvorak for supplying seeds of accession AL8/78 of *A. tauschii*, Dr R. C. Buell and anonymous reviewers for their critical reading and suggestion for improvement of the manuscript. Research supported by NSF contract no. 0077766. This paper is contribution number 04-108-J from the Kansas Agricultural Experiment Station.

## Supplementary Materials

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ2228/TPJ2228sm.htm

**Table S1** Summary of analysis of sequences from the random shotgun (trs) library of *Aegilops tauschii*

**Table S2** Relative enrichment or depletion of sequences in the methylation–filtration library of *Aegilops tauschii* (tmf) when compared with the random shotgun library (trs) (see Table S1)

**Table S3** Summary of repeated sequences from wheat EST collection TAGI

**Table S4** The estimated copy numbers of major TE (transposable element) families in the D-genome and their chromosomal distribution of in *Aegilops tauschii* and Chinese Spring (CS) wheat

**Table S5** Copy-number variation of repeated sequences in the Triticeae

**Table S6** Genes identified from trs and tmf libraries of *Aegilops tauschii* by BLASTX searches against the NR database of NCBI and their similarities to wheat ESTs by BLASTN searches against TAGI database

**Table S7** Species, subspecies, accession numbers, and nuclear DNA contents of plant material used

## References

**Adams, R.L.P. and Burdon, R.H.** (1985) *Molecular Biology of DNA Methylation*. New York: Springer Verlag. pp. 182–185.

**Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., 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.

**Arumuganathan, K. and Earle, E.D.** (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 211–215.

**Bennett, M.D. and Smith, J.B.** (1976) Nuclear DNA amounts in angiosperms. *Trans. R. Soc. Lond. B*, **274**, 227–274.

**Bennetzen, J.L.** (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42**, 251–269.

**Britten, R.J. and Kohne, D.E.** (1968) Repeated sequences in DNA. *Science*, **161**, 529–540.

**Bureau, T.E. and Wessler, S.R.** (1992) Tourist – a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*, **4**, 1283–1294.

**Bureau, T.E. and Wessler, S.R.** (1994a) Stowaway – a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*, **6**, 907–916.

**Bureau, T.E. and Wessler, S.R** (1994b) Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. *Proc. Natl Acad. Sci. USA*, **91**, 1411–1415.

**Bureau, T.E., Ronald, P.C. and Wessler, S.R.** (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc. Natl Acad. Sci. USA*, **93**, 8524–8529.

**Devos, K.M., Brown, J.K.M. and Bennetzen, J.L.** (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12**, 1075–1079.

**Echenique, V., Stamova, B., Wolters, P., Lazo, G., Carollo, V. and Dubcovsky, J.** (2002) Frequencies of Ty1-copia and Ty3-gypsy retroelements within the Triticeae EST databases. *Theor. Appl. Genet.* **104**, 840–844.

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

**Faris, J.D., Haen, K.M. and Gill, B.S.** (2000) Saturation mapping of a gene-rich recombination hot spot region in wheat. *Genetics*, **154**, 823–835.

**Feng, Q., Zhang, Y., Wang, S. *et al.*** (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.

**Feschotte, C., Jiang, N. and Wessler, S.R.** (2002) Plant transposable elements: where genetics meets gemomics. *Nat. Rev. Genet.* **3**, 329–341.

**Gruenbaum, Y., Naveh-Many, T., Cedar, H. and Razin, A.** (1981) Sequence specificity of methylation in higher plant DNA. *Nature*, **292**, 860–862.

**Hirochika, H, Sugimoto, K., Otsuki, Y., Tsugawa, H. and Kanda, M.** (1996) Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl Acad. Sci. USA*, **93**, 7783–7788.

**Huang, X. and Madan, A.** (1999) CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877.

**Kirchner, J., Connolly, C.M. and Sandmeyer, S.B.** (1995) Requirement of RNA-polymerase-III transcription factors for *in vitro* position-specific integration of a retrovirus-like element. *Science*, **267**, 1488–1491.

**Kumar, A. and Bennetzen, J.L.** (1999) Plant retrotransposons. Search result. *Annu. Rev. Genet.* **33**, 479–532.

**Langdon, T., Jenkins, G., Hasterok, R., Jones, R.N. and King, I.P.** (2003) A high-copy-number CACTA family transposon in temperate grasses and cereals. *Genetics*, **163**, 1097–1108.

**Mao, L., Wood, T.C., Yu, Y. *et al.*** (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. *Genome Res.* **10**, 982–990.

**Meyers, B.C., Tingey, S.V. and Morgante, M.** (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11**, 1660–1676.

**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*, **302**, 2115–2117.

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

**Rabinowicz, P.D., Schitz, 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 maize genome. *Nat. Genet.* **23**, 305–308.

**Rostoks, N., Park, Y.-J., Ramakrishna, W. *et al.*** (2002) Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. *Funt. Integr. Genomics*, **2**, 51–59.

**SanMiguel, P., Tikhonov, A., Jin, Y.K. *et al.*** (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science*, **274**, 765–768.

**Sasaki, T., Matsumoto, T., Yamamoto, K. *et al.*** (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.

**Shirasu, K., Schulman, A.H., Lahaye, T. and Schulze-Lefert, P.** (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res.* **7**, 908–915.

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

**Suoniemi, A., Schmidt, D. and Schulman, A.H.** (1997) BARE-1 insertion site preferences and evolutionary conservation of RNA and cDNA processing sites. *Genetica*, **100**, 219–230.

**The Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

**Whitelaw, C.A., Barbazuk, W.B., Pertea, G. *et al.*** (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science*, **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.

**Wicker, T., Guyot, R., Yahiaoui, N. and Keller, B.** (2003) CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol.* **132**, 52–63.

**Witte, C.-P., Le, Q.H., Bureau, T. and Kumar, A.** (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl Acad. Sci. USA*, **98**, 13778–13783.

**Zhang, H.-B., Zhao, X.P., Ding, D.L., Paterson, H.A. and Wing, R.A.** (1995) Preparation of megabase-size DNA from plant nuclei. *Plant J.* **7**, 175–184.

**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. USA*, **97**, 1160–1165.

**Zhang, P., Friebe, B., Lukaszewski, A.J. and Gill, B.S.** (2001) The centromere structure in Robertsonian wheat-rye translocation chromosomes indicates that centric breakage-fusion can occur at different positions within the primary constriction. *Chromosoma*, **100**, 335–344.

**Zhang, P., Li, W., Fellers, F.P. and Gill, B.S.** (2004) BAC-FISH in wheat identifies chromosome landmarkers consisting of different types of transposable elements. *Chromosoma*, **112**, 288–299.

Accession numbers: The sequence data from this study have been deposited in GenBank under accession numbers CG672285–CG677323.

**Web site references**

TAGI database at TIGR, URL: ftp://ftp.tigr.org/private/NHGI_tagi_87j7ik.
The CAP EST Assembler at IFOM, URL: http://bio.ifom-firc.it/ASSEMBLY/assemble.html.
WU-BLAST archive at the Washington University, URL: http://blast.wustl.edu.
BLAST server of Triticeae repeat (TREP) database, URL: http://wheat.pw.usda.gov/ITMI/Repeats/blastrepeats3.html.
Statistical page of Institute of Phonetic Sciences, URL: http://fonsg3.let.uva.nl/Service/Statistics.html.