Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat

Mathieu Charles*, Harry Belcram*, Jérémy Just*, Cécile Huneau*, Agnès Viollet†, Arnaud Couloux†, Béatrice Segurens†, Meredith Carter‡, Virginie Huteau§, Olivier Coriton§, Rudi Appels‡, Sylvie Samain†, Boulos Chalhoub*.1

* Organization and evolution of plant genomes (OEPG)
UMR INRA 1165 - CNRS 8114 – UEVE
Unité de Recherche en Génomique Végétale (URGV)
91057 Evry Cedex, France

† CEA: Institut de Génomique
GENOSCOPE
91057 EVRY Cedex, France

‡ State Agricultural Biotechnology Centre and Centre for Comparative Genomics,
Murdoch University, South Road, WA 6150 Australia

§ Unité Mixte de Recherches INRA - Agrocampus Rennes
Amélioration des Plantes & Biotechnologies Végétales
35653 Le Rheu, France

The 10 BAC clone sequences data obtained in this study and their annotation were deposited at EMBL / GenBank Data Libraries under accession nos. AM932680, AM932681, AM932682, AM932683, AM932684, AM932685, AM932686, AM932687, AM932688, AM932689.
Running Title: Differential proliferation of wheat transposons

Key words: Wheat, Transposable-elements, Differential-proliferation, Polyploidization, Genome-evolution

Corresponding author
Boulos Chalhoub:

UMR INRA 1165 - CNRS 8114 – UEVE
Unité de Recherche en Génomique Végétale (URGV)
91057 Evry Cedex, France
2 rue Gaston Crémieux
91057 Evry Cedex
Phone : 0033160874503  Fax : 0033160874549
E-mail : chalhoub@evry.inra.fr
ABSTRACT

Transposable elements (TEs) constitute more than 80% of the wheat genome but their dynamics and contribution to size variation and evolution of wheat genomes (*Triticum* and *Aegilops* species) remains unexplored. In this study, 10 genomic regions have been sequenced from wheat chromosome 3B and used to constitute, along with all publicly available genomic sequences of wheat, 1.98 Mb of sequence (from 13 BAC clones) of the wheat B genome and 3.63 Mb of sequence (from 19 BAC clones) of the wheat A genome. Analysis of TEs sequence proportions (%), ratios of complete to truncated copies and estimation of insertion dates of class I retrotransposons showed that specific types of TEs have undergone waves of differential proliferation in the B and A genomes of wheat. While both genomes show similar rates and relatively ancient proliferation periods for the *Athila* retrotransposons, the *Copia* retrotransposons proliferated more recently in the A genome whereas *Gypsy* retrotransposon proliferation is more recent in the B genome. It was possible to estimate for the first time, proliferation periods of the abundant *CACTA* class II DNA transposons, relative to that of the three main retrotransposon superfamilies. Proliferation of these TEs started prior to and overlaps with that of the *Athila* retrotransposons in both genomes. However, they proliferated also during same periods as *Gypsy* and *Copia* retrotransposons in the A genome, but not in the B genome. As estimated from their insertion dates and confirmed by PCR-based tracing analysis, the majority of differential proliferation of TEs in B and A genomes of wheat (87% and 83% respectively), leading to rapid sequence divergence, occurred prior to the allotetraploidization event that brought them together in *T. turgidum* and *T. aestivum*, less than 0.5 million years ago (MYA). More importantly, the allotetraploidization event appears to have neither enhanced nor repressed
retrotranspositions. We discussed the apparent proliferation of TEs as resulting from their insertion, removal and/or combinations of both evolutionary forces.
INTRODUCTION

Genomes of higher eukaryotes, and particularly those of plants, vary extensively in size (Bennett and Smith 1976, 1991; Bennett and Leitch 1997, 2005). This is observed not only among distantly related organisms, but also between species belonging to the same family or genus (Chooi, 1971; Jones and Brown, 1976). More than 90% of genes are conserved in sequenced plant genomes (Benetzen 2000a; Sasaki et al. 2005; Jallon et al. 2007) and thus differences in gene content explain only a small fraction of the genome size variation. It is widely accepted that whole genome duplication by polyploidization (Adam and Wendel 2005; Blanc et al. 2000; Paterson et al. 2004) and differential proliferation of transposable elements (TEs) are the main driving forces of genome size variation. The differential proliferation of TEs results from their transposition (Sanmiguel et al. 1996; Benetzen 2000b, 2002; Kidwell 2002, Benetzen et al. 2005; Pieg et al. 2006; Hawkins et al. 2006; Zuccolo et al. 2007) as well as differential efficiency of their removal (Petrov et al. 2000; Petrov 2002a, 2002b; Wendel et al. 2002).

Polyploidization and differential proliferation of TEs are particularly obvious in the case of wheat species belonging to the closely related Triticum and Aegilops genera. Rice (Oryza sativa), Brachypodium and diploid Triticum or Aegilops species underwent the same whole genome duplications (Adams and Wendel 2005; Salse et al. 2008), but Triticum or Aegilops genomes are more than 10 times larger (Bennett and Smith 1991), mainly due to proliferation of repetitive DNA, which represents more than 80% of the genome size (Smith and Flavell 1975; Vedel and Delseny 1987). Diploid wheat species can differ in their genome sizes by hundreds or even thousands of megabases (Mb) (Bennett and Smith 1976, 1991; http://data.kew.org/cvalues/homepage.html). For example, the genome size of T.
monococcum (6.23 pg) is 1.3 pg greater than that of T. urartu (4.93 pg) (Bennett and Smith 1976, 1991), although these species diverged less than 1.5 million years ago (MYA) (Dvorak et al. 1993; Huang et al. 2002; Wicker et al. 2003b). Similarly the calculated size of the B genome of polyploid what species (7 pg) is higher than that of any diploid wheat species (http://data.kew.org/cvalues/homepage.html).

The genome size variation within wheat is also accentuated by frequent allopolyploidization events, among which two successive events have led to the formation of the allohexaploid bread wheat T. aestivum (2n=6x=42, AABBDD). The first event led to the formation of the allotetraploid T. turgidum (2n=4x==28, AABB) and occurred less than 0.5-0.6 MYA between the diploid species T. urartu (2n=2x=14, AA), donor of the A genome and an unidentified diploid species of the Sitopsis section, donor of the B genome (Feldman et al. 1995; Blake et al. 1999; Huang et al. 2002; Dvorak et al. 2006). The second allopolyploidization event occurred 7,000 to 12,000 years ago, between the early-domesticated tetraploid T. turgidum ssp. dicoccum and the diploid species Ae. tauschii (2n=14), donor of the D genome, resulting in hexaploid wheat (Feldman et al. 1995).

The amount of available wheat genomic sequences is very limited, compared to other organisms (reviewed by Sabot et al. 2005; Stein 2007; http://genome.jouy.inra.fr/triannot/index.php and http://www.ncbi.nlm.nih.gov/). Individual bacterial artificial chromosome (BAC) clones, selected primarily because they contained genes of agronomic interest, have been sequenced. Analyses of randomly chosen BAC clones from wheat have been also performed (Devos et al. 2005), and 2.9 Mb of sequences from a whole-genome shotgun library of Ae. tauschii were analyzed by Li et al. (2004). More recently, a detailed analysis of 19,400 BAC-end sequences of chromosome 3B, representing
a cumulative sequence length of nearly 11 Mb (1.1% of the estimated chromosome length) was reported (PAUX et al. 2006). Altogether, these sequencing efforts have confirmed previous estimates of the amount of repetitive DNA in the wheat genome (around 80%) (SMITH and FLAVELL 1975; VEDEL and DELSENY 1987) and identified the major types of TEs (WICKER et al. 2002; SABOT et al. 2005).

Because of the limited genomic sequence information, the extent to which various TEs contribute to the wheat genome and affect its size variation, or how they are distributed among different genomes, remain unexplored. Little is known about the dynamics of TEs, their proliferation processes, and whether they proliferated gradually or in waves of sudden bursts of insertions. In this study, 10 genomic regions from wheat chromosome 3B were sequenced and used to constitute, along with three other genomic sequences, 1.98 Mb of sequence from the wheat B genome. Transposable element dynamics and proliferation in these B genome sequences were analyzed and compared to those in 3.63 Mb of sequence from 19 genomic regions of the wheat A genome. Our study provides novel insights into the dynamics and differential proliferation of TEs as well as their important role in the evolution and divergence of the wheat B and A genomes.
MATERIALS AND METHODS

Plant material and genomic DNA isolation

Hexaploid wheat deletion lines used to map the 10 BAC clones on different deletion BINs of chromosome 3B (see results) were originally described by Qi et al. (2003) and kindly provided by Dr. Catherine Feuillet (INRA-Clermont-Ferrand, France). Hexaploid wheat genotypes were kindly provided by Dr. Joseph Jahier (INRA-Rennes, France). Tetraploid wheat genotypes were kindly provided by Dr. Moshe Feldman (Weizemann Institute, Israel). Genomic DNA was extracted from leaves as described by Graner et al. (1990).

Primer design and PCR-based tracing of retrotransposon insertions

The program Primer3 (Rozen et al. 2000) was used to design oligonucleotide primers based on TE-TE or TE-unassigned DNA junctions. We often designed and used multiple couples (including nested) of PCR primers. Internal controls (PCR primers designed within the TE) were also used. Primer sequences are given in Supplemental Table 1. PCR reactions were carried out in a final volume of 10 µl with 200 µM of each dNTP, 500 nM each of forward and reverse primers, 0.2 U Taq polymerase (Perkin Elmer, USA). PCR amplification was conducted using the following “touchdown” procedure: 14 cycles (30 sec 95°C, 30 sec 72°C minus 1 °C each cycle, 30 sec 72 °C), 30 cycles (30 sec 95°C, 30 sec 55°C, 30 sec 72°C) and one additional cycle of 10 min 72°C. Amplification products were visualized using standard 2% agarose gels.
BAC sequencing, sequence assembly and annotation

BAC shotgun sequencing was performed at the ‘Centre National de Sequencage’ (Evry, France) essentially as described by CHANTRET et al. (2005). Genes, TEs and other repeats were identified by computing and integrating results based on BLAST algorithms (ALTSCHUL et al. 1990, 1997), predictor programs, and different software and procedures, detailed below. Cross-analysis of the information obtained for genes, TEs as well as repeats and unassigned DNA was integrated into ARTEMIS (RUTHERFORD et al. 2000). Sequence annotation and analysis were performed as described in Supplemental Method 1. The 10 BAC clone sequences were submitted to EMBL and references are: TA3B54F7: AM932680; TA3B63B13: AM932681; TA3B63B7: AM932682; TA3B81B7: AM932683; TA3B95C9: AM932684; TA3B95F5: AM932685; TA3B95G2: AM932686; TA3B63C11: AM932687; TA3B63E4: AM932688; TA3B63N2: AM932689. References for the three publicly available genomic sequences from the wheat B genome (SABOT et al. 2005; GU et al. 2006; DVOvak et al. 2006) are: CT009588, AY368673, DQ267103.

Publicly available genomic sequences from the wheat A genome

The retained publicly available A genomic sequences consist of 19 sequenced and well annotated BAC clones or contigs (YAN et al. 2002, 2003; SANMIGUEL et al. 2002; WICKER et al. 2003b; CHANTRET et al. 2005; ISIDORE et al. 2005; GU et al. 2006; DVOvak et al. 2006; MILLER et al. 2006), representing more than 3.5 Mb. References (accession numbers) of the analyzed BAC sequences are below:
**Diploid A genome:** AF326781, AF488415, AY146588, AY188331, AY188332, AY188333, AY491681, AY951944, AY951945, DQ267106, AF459639.

**Tetraploid A genome:** AY146587, AY485644, AY663391, CT009587, DQ267105.

**Hexaploid A genome:** AY663392, CT009586, DQ537335.

**Chromosome 3B BAC clones and fluorescent in situ hybridisation (FISH)**

The 10 BAC clones and/or their subclones were originally mapped by fluorescence in situ hybridization (FISH) on flow-sorted 3B chromosomes using the Cot-1 fraction as blocking DNA to suppress hybridization of repeated sequences (SAFAR et al. 2004; DOLEZEL et al. 2004; M. KUBALAKOVA and J. DOLEZEL, personal communication). Further FISH hybridization experiments were conducted, without Cot-1 DNA, on mitotic metaphase chromosomes of hexaploid wheat (*T. aestivum*) cv. Chinese Spring (CS). The FISH hybridization protocol is presented in Supplemental Method 2.

**Estimation of LTR-retrotransposon insertion dates**

For all genomic sequences of the B and A genomes of wheat, retrotransposon copies with both 5’ and 3’ LTR and target site duplications (TSD) were considered as corresponding to original insertions and analyzed by comparing their 5’ and 3’ LTR sequences. The two LTRs were aligned and the number of transition and transversion mutations was calculated using MEGA3 software (KUMAR et al. 2004). A mutation rate of 1.3x10^{-8} substitutions/site/year (SANMIGUEL et al. 1998; MA et al. 2004; WICKER et al. 2005; MA and BENNETZEN 2004; Gu et al. 2006) was used. The insertion dates and their standard errors (SE) were estimated using the formula \( T = \frac{K2P}{2r} \) (KIMURA 1980).
Statistical analysis

All statistical analyses and the different tests (KOLMOGOROV-SMIRNOV, Bootstrap, probability density functions) were done with the R-package (http://www.r-project.org). KOLMOGOROV-SMIRNOV tests (FÉRIGNAC, 1962) were applied to check whether the distribution of insertion dates of retrotransposons deviates from uniformity, and whether they are different when comparing different TE families or superfamilies within and between the B and A genomes. Probability density of TE insertion dates was estimated using Gaussian kernel density estimation (SILVERMAN 1986), taking into account measured standard deviation for each individual insertion date (KIMURA 1980).
RESULTS

Constitution of a genomic sequence data set representative of the wheat B genome

Analysis of 10 BAC sequences from the wheat chromosome 3B: Only three large well-annotated genomic sequences (BAC clones), representing 0.55 Mb of sequence, were available for the wheat B genome (SABOT et al. 2005; GU et al. 2006; DVORAK et al. 2006). To obtain more representative genomic sequences, we sequenced and annotated 10 BAC clones of wheat chromosome 3B, representing 0.15% of the chromosome length (1.43 Mb) (Figure 1). Detailed annotation files are deposited at EMBL / GenBank Data Libraries. These sequenced genomic regions show high proportion of TEs, which represent 79.1% of the cumulative sequence length (Figure 1, Supplemental Table 2). Other repeated DNA sequences represents 2.4% and unassigned DNA sequences account for 17.5% of the cumulative sequence length.

We conducted gene prediction analysis for the remaining 18.5% non-TEs and non-repeated DNA, using different search programs (see Supplemental Method 1 and Supplemental Text 1 for detailed description). Genes of known and unknown functions, or putative genes were defined based on predictions and the existence of rice or other Triticeae homologs. Hypothetical genes were identified based on prediction programs only. Pseudogenes were not well predicted and frameshifts need to be introduced within the CDS structure to better fit a putative function based on BLASTX (mainly with rice). Truncated pseudogenes (genes disrupted by large insertion or deletion) and highly degenerated CDS sequences were considered as gene-relics. Combined together, all these types of gene sequence information
(GSI) account for only 1.0% of the sequence and are present in seven BAC clones (one or two genes per clone) while the remaining three BAC clones (TA3B95C9, TA3B95G2, TA3B63N2) contain no genes (indicated in Figure 1A and detailed in Supplemental Text 1, Supplemental Table 3 and Supplemental Table 4).

Six genes (of known or unknown function), and two putative genes were identified using the FGENESH prediction software (http://www.softberry.com) and by identification of homologs in rice (Figure 1A, Supplemental Table 3). Six additional ‘gene-relics’ or ‘pseudogenes’ were also identified based on colinearity with rice (Figure 1A, Supplemental Table 3). Finally 10 CDS, designated as ‘hypothetical genes’, were identified according to the FGENESH prediction program only (Figure 1A, Supplemental Table 4).

TE prediction, annotation, classification and nomenclature were performed essentially as suggested by the unified classification system for eukaryotic TEs (WICKER et al. 2007) with two modifications. The Athila retrotransposons were analyzed separately from the other Gypsy retrotransposons (see also Supplemental Methods 1). The Sukkula retrotransposons were considered as belonging to the Gypsy superfamily because of similarities with the Erika (Gypsy) elements. The 79.1% of TEs were shown to be composed of a wide variety of TEs, distributed as follows: 61.9% class I (171 TEs from 48 families), 16.2% class II (113 TEs from 28 families) and 1.0% unclassified TEs (18 TEs from nine families) (Figure 1). The CACTA TEs represent the majority (96%) of class II TEs. More details about the TE composition in the 10 different BAC clones of wheat chromosome 3B are provided in Supplemental Text 2.

Twenty-one transposable element families, some of which are present in several copies, were identified for the first time in this study (Figure 1A, indicated by arrows). They account for
9.8% by number and 7.9% by length of the overall sequences. Class I retrotransposons are the category for which we found the majority of novel TE families (17). Description of these novel TEs, their features as well as the suggested nomenclature are presented in Supplemental Text 2 and Supplemental Table 5.

The 10 sequenced BAC clones or their subclones were originally mapped by fluorescence in situ hybridization (FISH) on flow-sorted 3B chromosomes, using the C,ot-1 fraction as blocking DNA to suppress hybridization of repeated sequences (SAFAR et al. 2004; DOLEZEL et al. 2004; M. KUBALAKOVA and J. DOLEZEL, personal communication). As described by DEVOS et al. (2005) and PAUX et al. (2006), specific PCR markers, based on TE-TE or TE-unassigned DNA junctions were used to confirm the different BAC clone map positions on the deletion BINs (Qi et al. 2003) of chromosome 3B (except TA3B63E4) (Figure 1B). Details of PCR markers and genotyping results are given in Supplemental Table 6.

**Representation of transposable elements and the wheat B genome:** Five BAC clone sequences were publicly available from the B genome of wheat (SABOT et al. 2005; GU et al. 2006; DVORAK et al. 2006). Four of these were sequenced for two orthologous regions in tetraploid and hexaploid wheat species (one BAC clone per region and per species) (SABOT et al. 2005; GU et al. 2006). As they share nearly identical sequences (99%) with common TEs insertions, they were considered as redundant in our study and only the longest BAC clone sequences (three in total) were counted in calculation and appreciation of TEs proliferation. These, added to the above-described 10 genomic region sequences of wheat chromosome 3B, constitute 1.98 Mb of sequence from the wheat B genome. Four main TE superfamilies occupy 66.5% of the analyzed B genome loci: the Athila superfamily (54
elements), the *Copia* superfamily (57 elements), the *Gypsy* superfamily (79 elements), and the *CACTA* superfamily (70 elements) (Table 1). Interestingly, proportions of the *Athila*, *Copia* and *Gypsy* retrotransposons (respectively 10.8%, 14.2% and 28.1%) (Table 1) are very similar to estimates based on 11 Mb of the chromosome 3B sequences BAC-end (Paux *et al.* 2006). The major deviation concerns the proportion of *CACTA* class II TEs, which is higher in the 13 genomic regions (13.4%) than in the overall BAC-end sequences (4.9%), probably due to their clustering in some BAC clones that we have sequenced, such as TA3B54F7 (40.5% of *CACTA* TEs) (Figure 1).

The 13 sequences represent only ~0.03% of the B genome. However, statistical tests, using standard errors (SE) as well as a bootstrap analysis with 10,000 re-samplings, confirm robustness of estimations of sequence proportions of the *Gypsy*, *Copia*, *Athila* and *CACTA* TE superfamilies (Table 1). We also evaluated variation of mean sequence proportions estimated for the four TE superfamilies by comparing all possible clone number representations and combinations (from one to 12 BAC clones) (Figure 2). Results show that representing the wheat B genome with a low number of BAC clones results in very variable proportions of the TE sequences (Figure 2). These variations decrease significantly by increasing the number of considered BAC clones (Figure 2). This confirms the usefulness of our effort in sequencing more BAC clones for better representation of the wheat B genome.

It is also interesting to note that direct FISH hybridization, using the whole BAC clone as a probe, resulted in dispersed and mostly homogenous signals across all wheat chromosomes for eight out of all 10 BAC clones of wheat chromosome 3B (except TA3B63C11 and TA3B54F7) (Safar *et al.* 2004 and Supplemental Figure 1), thus confirming sequencing results that show high TE composition.
Constitution of a genomic sequence data set representative of the wheat A genome

The publicly available A genomic sequences that we were able to use are more abundant and consist of 20 sequenced and well annotated BAC clones or contigs. 10 of these were comparatively sequenced for five orthologous regions of the wheat A genome at the diploid, tetraploid and/or hexaploid levels and were partially overlapping (Wicker et al. 2003b; Chantret et al. 2005; Isidore et al. 2005; Dvorak et al. 2006; Gu et al. 2006); while others were determined at only one ploidy level (mostly diploid) (Sanmiguel et al. 2002; Yan et al. 2002, 2003; Miller et al. 2006). Comparisons show that no shared TE insertions were observed between orthologous regions (from two ploidy levels), except in the region of the high molecular weight (HMW) glutenin gene, the sequences of which were nearly identical at the tetraploid and hexaploid levels (Gu et al. 2006). Thus, we only used the sequence from hexaploid wheat to represent the HMW glutenin gene region and considered all the other different orthologous regions (from different ploidy levels) separately. This led to 19 BAC clones, representing 3.63 Mb of sequence, that were analyzed for the wheat A genome.

The Gypsy TEs were found to occupy 19.7%, the Athila TEs 10.4%, the Copia TEs 21.8% and the CACTA TEs 9.4% of the cumulative sequence length (Table 1). Similarly to the B genomic sequences, we also analyzed and validated the robustness of the estimation of sequence proportions of the main TEs superfamilies and their representation of the A genome (Figure 2). It is also interesting to note that similar proportions of the Gypsy, Copia, Athila and CACTA TEs were found whether the 11 genomic sequences from the diploid A genome
or those determined from A genomes of tetraploid (six regions) and hexaploid (three regions) wheat species were considered separately or combined (data not shown).

**Comparison of TE sequence proportions and ratios of complete to truncated copies**

Our analysis showed a significantly higher number of *Gypsy* retrotransposons in the wheat B genome sequences than in the A genome (Table 1). Conversely, a higher proportion of *Copia* retrotransposons is observed in genomic sequences of the wheat A genome than in the B genome (Table 1). Proportions of the *Athila* and *CACTA* TEs were not statistically different between the two genomes (Table 1).

Major differences were found between the three main retrotransposon superfamilies, in the ratio of complete (intact) copies, defined as having both long terminal repeats (LTRs) and target site duplication (TSD), as compared to degenerated and truncated copies that resulted from LTR-mediated unequal homologous recombinations (UHR) or illegitimate DNA recombination ([Devos et al. 2002; Ma et al. 2004; Ma and Bennetzen 2004; Vitte and Bennetzen 2006]) (Table 1). In the B genome sequences, the *Athila* and *Copia* retrotransposons show low ratios of complete to incomplete retrotransposons (respectively 0.32 and 0.46), whereas the *Gypsy* retrotransposons show the highest ratio (0.98) (Table 1). In comparison, the 3.63 Mb of genomic sequence of the wheat A genome shows a lower ratio (0.45) of complete to incomplete *Gypsy* retrotransposons whereas proportions of intact *Copia* retrotransposons are relatively higher than those observed in the B genome (0.67) (Table 1). The *Athila* retrotransposon ratio in the A genome is comparable to the ratio in the B genome (0.36 and 0.32 respectively).
CACTA TE original insertions are characterized by the “CACTA” sequence and three bp TSD sequence motifs, surrounding terminal inverted repeats (TIR) at both ends. We used these signatures to define complete CACTA copies, where the “CACTA”, TIR and TSD sequence motifs are observed at both ends, and truncated copies, where the “CACTA” and TSD motifs are absent from one or both ends. Ratio of complete to incomplete copies of the CACTA class II TE was about five times lower in the wheat B genome (ratio of 0.37) than in the A genome (ratio of 1.52) (Table 1).

Insertion dates and proliferation of LTR-retrotransposons

To understand differences in sequence proportions and the ratios of complete to truncated copies between retrotransposon superfamilies, as well as between the B and A genomes, we compared TE proliferation periods and rates.

The two LTRs are identical at the time of retrotransposon insertion and their sequence divergence reflects time lapsed since the insertion (SanMiguel et al. 1998). Several studies have shown that LTRs evolve at approximately twice the rate of genes and UTR regions, and we used a rate of 1.3x10^-8 substitutions/site/year (Ma et al. 2004; Ma and Bennetzen, 2004; Wicker et al. 2005; Gu et al. 2006).

We calculated the LTR-divergence and dates of insertion of the Athila, Copia and Gypsy retrotransposon (complete copies with both LTRs and TSD) found in the wheat B and A genomes (Figure 3). Such TE insertion dates offer a very important insight into relative timing of various events, regardless of the approaches used to estimate nucleotide substitution rates or molecular clock calibration points used in these calculations.
The vast majority of complete retrotransposons in the B and A genomes of wheat (86% and 92% respectively) were estimated to be less than three million years old (Figure 2), in agreement with several previous studies of grasses and other plants species (SanMiguel et al. 1998, 2002; Wicker et al. 2003b, 2005; Gao et al. 2004; Ma et al. 2004; Du et al. 2006; Piegu et al. 2006; Wicker and Keller 2007). This is explained by the fact that LTR retrotransposons are continuously removed by unequal homologous recombination and illegitimate DNA recombination as new ones are inserted (Vicient et al. 1999; Devos et al. 2002; Ma et al. 2004; Pereira 2004). Insertion of Egug element (RLGa_Egug_TA3B95C9-1 approximately 5 MYA (divergence: 0.131) is the oldest such event found in our study and the most recent one is the Sukkula insertion (RLG_Sukkula_TA3B63B7-2) for which only one-base indel differentiates the two LTRs of 4192/4193 bp.

Comparison of LTR divergence dates revealed that different LTR-retrotransposon superfamilies and families proliferated at different periods and rates during evolution of the wheat B and A genomes (Figure 3). We applied Kolmogorov-Smirnov tests to check whether within the last three million years (0.078 divergence), the distribution of insertion dates of retrotransposons deviate from uniformity (thus confirming a burst of higher proliferation), and whether they are different when comparing different retrotransposon families or superfamilies within and between the wheat B and A genomes (thus illustrating differential proliferation). This was done for all complete copies of the three main retrotransposon superfamilies as well as for the most abundant retrotransposon families (nine) that have five or more complete copies in the B and/or A genomes (Figure 3).
Superfamily level comparison: The combination of all complete retrotransposon copies at the superfamily level (Figure 3A), indicated that the distribution of the Gypsy retrotransposon insertion dates in both B and A genomes and that of Copia retrotransposons in the A genome were significantly different from uniform (P-value <0.01) because of their higher proliferation during the last two million years (Figure 3A). Proliferation of the Copia retrotransposons in the B genome was uniform and low all across the three million year period, whereas proliferation of the Athila retrotransposons was different from a uniform distribution in both genomes at P-value < 0.1.

One possible reason for the non-uniform distributions of retrotransposon insertion dates within the three million year period is because older insertions are more likely to be removed (completely or partially) from the genome (see above). Therefore, we checked whether distributions of insertions are significantly different from a uniform distribution for the most recent period of evolution during which the impact of DNA removal should be lower. To carry out this analysis, we divided the LTR-retrotransposon insertions according to the median (of their distribution) that varies depending on the retrotransposon superfamily and family (Figure 3A, gray dot). Kolmogorov-Smirnov (Féronac 1962) tests were then conducted on the half of complete copies, which show the most recent insertion dates. Distribution of insertion dates of the Gypsy retrotransposons in the wheat B genome and that of the Copia retrotransposons in the B and A genomes can be considered as uniform (P-value > 0.05, Figure 3A), indicating that they have constantly proliferated during this most recent period. In contrast, the distribution of Athila retrotransposons in the wheat B and A genomes and that of Gypsy retrotransposons in the A genome are not uniform (P-value <0.05, Figure 3A), consistent with a decreasing proliferation during the most recent period.
Comparison of the proliferation of the three retrotransposon superfamilies, shows that distribution of the Athila retrotransposons is statistically different from that of the Gypsy retrotransposons (Figure 3A, P-value <0.05) in the B genome. The Athila distribution is significantly different from that of the Gypsy and Copia retrotransposons (Figure 3A, P-value <0.05) in the A genome.

Comparison of the distributions of the three retrotransposon superfamilies between the B and A genomes shows that Copia distributions are significantly different (Figure 3A, P-value =0.628), due to their higher proliferation and more recent insertions in the A genome. Both genomes show similar old distribution of the Athila retrotransposons (Figure 3A). Distributions of the Gypsy retrotransposons were not statistically different between the two genomes for the whole three million years period (Figure 3A, P-value >0.05). However, separate KOLMOGOROV-SMIRNOV tests for the most recent period show that these have less proliferated in the wheat A genome (P-value =0.052, Figure 3A), unlike in the wheat B genome (P-value =0.38, Figure 3A).

**Distribution of the most abundant retrotransposon families:** Some specific retrotransposon families were abundant in the B and/or A genomes. This is the case of the Angela and Wis families, together representing 72% and 85% of the Copia superfamily in the B and A genomes respectively (Figure 3B). This is also the case of the Sabrina family representing 62% and 63% of the Athila superfamily in the B and A genomes respectively (Figure 3B). There are more families that compose the Gypsy retrotransposon superfamily, the most abundant being Fatima, representing 25% in both genomes (Figure 3B).
KOLMOGOROV-SMIRNOV tests show non-significant deviations (P-value > 0.05) from uniform distributions for all nine retrotransposon families (with five or more observed complete copies in at least one genome), with the exception of the *Jeli* (*Gypsy*) elements in the B genome and the *Angela* (*Copia*) elements in the A genome, which have more recently proliferated (Figure 3B). Separate analysis for the most recent period, corresponding to half of the complete copies, shows that, as expected from the superfamily level analysis, the *Wham* family in the A genome and the *Sabrina* family in the B genome have not recently proliferated (P-value < 0.05, Figure 3B).

Distribution of insertion dates of the *Wham* and *Sabrina* families are different from almost all the other seven families within and between the B and A genomes (P-value < 0.05). Distribution of insertion dates of the *Angela* family in the wheat A genome is statistically different (P-value < 0.05) from that of the *Fatima* family in both genomes. Distributions of insertion dates of the remaining families do not show statistical differences (P-value > 0.05) within and between the wheat B and A genomes (Figure 3B).

Moreover, some retrotransposons families were abundant and present in several complete copies in only one genome (*Romani, Daniela, Erika* and *Wham* for the A genome; *Egug* and *Jeli* in the B genome) but absent or presenting few copies in the other (Figure 3B). It is likely that this corresponds to differential proliferation of the considered retrotransposons, as different copies were detected in different genomic regions of wheat B or A genomes.

**LTR retrotransposon proliferation was neither enhanced nor repressed by the allotetraploidization event**
The allotetraploidization event that brought the B and A genomes of wheat together in one nucleus was estimated to occur no more than 0.5-0.06 MYA (HUANG et al. 2002; DVORAK et al. 2006, CHALUPSKA et al. 2008). This corresponds to a divergence-interval of 0.013-0.016, using the corrected rate of $1.3 \times 10^{-8}$ substitutions/site/year for more rapid divergence of LTRs (MA et al. 2004; MA and BENNETZEN 2004, DVORAK et al. 2006). Comparisons show that retrotransposon insertions continued in wheat B and A genomes during the last 0.5-0.6 million years, apparently without being enhanced nor repressed by the allotetraploidization event (Figure 3). For example, analysis of genomic sequences available from the three ploidy levels of the A genome do not show differences in proliferation periods and rates of retrotransposons (Figure 3).

In order to check the accuracy of these observations and to calibrate the divergence rate used for coding sequences on one hand and that used for LTRs of retrotransposons on the other hand, we traced several retrotransposons for their insertion prior or posterior to the allotetraploidization event. A PCR-based tracing strategy, derived from the retrotransposon-based insertion polymorphism (RBIP) method (FLAVELL et al. 1998; DEVOS et al. 2005; PAUX et al. 2006), was developed for 21 retrotransposon insertions from the B genome, sampled as having different estimated insertion dates (Figure 3, indicated by gray triangles). It simply relies on primers designed in both the retrotransposon and its flanking sequences (either unassigned DNA or an older pre-inserted TE sequence) so that PCR amplification will be specific to the retrotransposon insertion. As the diploid wheat species donor of the B genome is unknown (FELDMAN et al. 1995; BLAKE et al. 1999; HUANG et al. 2002), we analyzed the occurrence (ie. presence or absence) of the 21 retrotransposon insertions in hexaploid (*T. aestivum*) and tetraploid (*T. turgidum*) wheat genotypes, which carry the wheat
Examples of PCR-based tracing of the 21 original retrotransposons insertions in the wheat genotypes compared with their estimated insertion dates (± standard errors: SE) are presented in Figure 4. Full tracing results are supplied in Supplemental Table 7 and sequences of the PCR primers in Supplemental Table 1. With the exception of except the Jeli_TA3B95C9-1, all the other seven most recently inserted retrotransposons, which have calculated insertion dates intervals (means ± SE) equal to or less than the 0.5-0.6 MYA interval (divergence 0.013-0.016), were detected in some but not all genotypes carrying the B genome, suggesting their occurrence after the tetraploidization event (Figure 4 and Supplemental Table 7). In contrast, all 13 retrotransposon insertions, which have calculated insertion intervals (means ± SE) older than 0.7 MYA, were detected in all tested genotypes carrying the B genome, suggesting their occurrence prior to the allotetraploidization event (Figure 4 and Supplemental Table 7). Given the uncertainty in calculating intervals of insertion dates, the PCR-based tracing method confirms the calibration of LTR divergence on that of gene divergence. More importantly, it also confirms that retrotranspositions (insertions) were not enhanced nor repressed by the allotetraploidization event.

Relative proliferation periods of the CACTA class II transposable elements

The CACTA class II DNA TEs represent an important proportion of the B and A genomic sequences (13.4% and 9.4% respectively). As for the main LTR-retrotransposon superfamilies, ratios of complete to truncated copies are very different for B (0.37) and A (1.52) genomes (Table 1). In contrast to LTR-retrotransposons, the CACTA TEs do not have long repeats or other features, which would allow determination of their insertion dates based on sequence divergence. Therefore, their proliferation periods and rates were evaluated
indirectly, relative to their level of insertions -into or -by other CACTA TEs and, more importantly, by elements of the three main LTR-retrotransposon superfamilies for which proliferation periods and rates were evaluated based on dates of insertions (described above). This was calculated for all CACTA TE copies as well as for complete and truncated copies separately (Table 2).

In the wheat B genome, the majority of CACTA TE insertions (mainly those detected as truncated copies) occurred in DNA annotated as unassigned (Table 2). For the rest, significantly higher insertions of CACTA TEs into Athila and other CACTA TEs than into Copia and Gypsy retrotransposons were observed. The two latter retrotransposon superfamilies were significantly more inserted into, than by, CACTA TEs (Table 2). These observations indicate that proliferation of the CACTA TEs in the B genome of wheat started before, and continued during and after Athila retrotransposon proliferation, whereas very few insertions occurred during the last waves of high proliferation of Copia and Gypsy.

Similarly, a high level of insertions into unassigned DNA was observed for the CACTA TEs in the A genome. However, for the remaining insertions, no clear period of proliferation could be determined as these show similar levels of insertions into or by all other TE superfamilies (Table 2). These observations, combined with the observed higher level of complete copies (Table 1), suggest that the CACTA TE proliferation continued in the wheat A genome during the last waves of proliferation of Copia and Gypsy, unlike those in the B genome.
DISCUSSION

To constitute representative genomic sequences of the wheat B genome, we have sequenced in the present study 10 BAC clones of the chromosome 3B, representing the most important number of genomic regions sequenced for a single wheat chromosome and a cumulative sequence length of 1.429 Mb (0.15% of the chromosome length). As expected, transposable elements (TEs) proliferation was pronounced (representing 79.1%). Five of these were revealed as gene-containing BAC clones, at a density of one or two genes per clone, two other BAC clones contain gene-relics or pseudogenes, whereas the three remaining BAC clones were missing genes. This confirms previous conclusion about the more random distribution of genes on the wheat genome (DEVOS et al. 2005). Interestingly and in comparison with rice, a high level of ‘truncated-genes” (six gene-relic or pseudogenes), several of which by TEs insertions (three confirmed cases), was revealed. If the confirmed gene number (excluding hypothetical genes), identified in the 1.43 Mb sequences (eight), is extrapolated to the whole wheat chromosome 3B of 1Gb estimated size, then 5594 genes might be present. A slightly higher number (6000) was calculated from BAC-end sequence analysis (PAUX et al. 2006).

**Representation of transposable elements**

In the present study, TE dynamics, proliferation and evolutionary pathways were analyzed and compared between 1.98 Mb of sequence from 13 BAC clones of the wheat B genome and 3.63 Mb of sequence from 19 BAC clones of the wheat A genome. These genomic sequences represent very small fractions (less than 0.03%) of their respective genomes. Nevertheless, it has been argued that for studying abundant repeats, sequencing and
annotation of a small proportion of the genome can be representative (BRENNER et al. 1993; VITTE and BENNETZEN 2006; LIU et al. 2007). We have been able to confirm the adequate representation where less variation in the proportion of the main TE superfamilies was observed when analyzing a large number of BAC clones (Figure 2). Interestingly, TE proportions observed in the 13 genomic regions of the B genome of wheat are similar to those obtained from 11 Mb of BAC-end sequences of wheat chromosome 3B (PAUX et al. 2006). Similarly, TE proportions were not significantly different for the wheat A genome when they were compared between the different ploidy levels (see results).

Although they are representative of abundant wheat TEs available in the TREP database (WICKER et al. 2002; http://wheat.pw.usda.gov/ITMI/Repeats), the class I and class II TEs observed in the genomic sequences of the wheat B and A genomes may not cover all wheat TEs. It is expected that more wheat TEs will be identified, as further wheat genomic sequences will become available. This is particularly supported by the identification in this study of more than 21 different novel TE families, most of which (17) are retrotransposons. We also believe that low-copy TEs and those that tend to ‘compartmentalize’ in specific regions, like pericentromeric heterochromatin regions (which is not the case of our regions) would be missed, over- or underrepresented in this study (MA and BENNETZEN, 2006; Liu et al. 2007). This could be the case of the CACTA TEs, which show the highest variation in sequence proportion between regions because they tend to be clustered in the Triticeae genomes (our results and WICKER et al. 2003a, 2005).

**Transposable elements proliferated differentially in the B and A genomes of wheat**
Abundance of TEs varies widely across different organisms. Human (*Homo sapiens*) DNA is composed of 45% (Lander *et al.* 2001) repetitive sequences, *Drosophila melanogaster* of 3.9% (Kaminker *et al.* 2002) and maize of 67% (Haberer *et al.* 2005; Liu *et al.* 2007) whereas TE content in the wheat genomic sequences analyzed in this study or in other studies (Li *et al.* 2004; Gu *et al.* 2006; Paux *et al.* 2006) is around 80%. Proportions of different classes of TEs also vary between organisms. Class II TEs are almost more than 10 times less abundant than class I TEs and constitute a small fraction (less than 2%) of the human, rice (Piegu *et al.* 2006), maize (Kronmiller and Wise 2008), *Arabidopsis* and cotton (Hawkins *et al.* 2006) genomes. In comparison, class II TE abundance is relatively high in the wheat B and A genomes (14.1% and 9.9% respectively) the majority of which (95%) are CACTA TEs, which are particularly abundant in the *Triticeae* genomes (Wicker *et al.* 2003a, 2005). Class I retrotransposon abundance is relatively high in several plant genomes, 58.7% and 56.6% estimated in this study for the wheat B and A genomes respectively, 40-50% in cotton species (Hawkins *et al.* 2006), 35-60% in rice species (Piegu *et al.* 2006) and 64% in maize (Kronmiller and Wise 2008; Liu *et al.* 2007).

In the present study, combination of TE sequence analysis and classification, comparison of proportions of complete to incomplete copies, TE insertion date estimations as well as PCR-based tracing of insertions, allow us to compare their proliferation periods and rates in the wheat B and A genomes (Figure 5). It is evident that TEs appear to proliferate differentially in waves of high activity followed by periods of low activity (Figure 5). Both genomes show similar rates and relatively old proliferation periods for the *Athila* retrotransposons (Figure 5). However, the *Copia* retrotransposons have relatively proliferated more recently in the A genome whereas a more recent *Gypsy* proliferation is observed in the B genome. Due to their
biology and replication mechanism, it was not possible to directly estimate the CACTA class II TE insertion dates. We have estimated their proliferation periods and rates relative to that of the three main LTR retrotransposon superfamilies. In the wheat B genome, the CACTA TE high proliferation period started before and overlaps with that of the Athila retrotransposons. In the wheat A genome, in addition to the relatively old proliferation similar to that in the B genome, CACTA TEs continued to proliferate during the same period as Gypsy and Copia retrotransposons. Determining the ancient proliferation periods of CACTA TEs partially explain why CACTA TEs often tend to be clustered together (our Results and Wicker et al. 2003a, 2005), although they were detected in almost all analyzed BAC clones. Differential proliferation of TEs provides a valid explanation for the size variation of closely related wheat genomes (Bennett and Smith 1976, 1991; http://data.kew.org/cvalues/homepage.html).

This is the first time that dynamics as well as proliferation periods and rates of TEs are compared between two closely related wheat genomes. This was only possible because we sequenced, in the present study, 10 different genomic regions and constituted a genomic sequence data set representative of the wheat B genome. For the wheat A genome, more representative genomic sequence data were rendered publicly available. There have been initial attempts to evaluate TE proliferation in the wheat genomes. Li et al. (2004) analyzed the D genome of the diploid Ae. tauschii and showed that the copy number of most TEs have increased gradually following polyploidization. However, they used dot blots, which are not very accurate. Sabot et al. (2005) have updated TEs annotation in wheat genomic sequences and reported their composition and distribution in relation to genes. They suggested that Copia TEs have been most active in the wheat A, B and D genomes, combined together.
Accurate comparison of dynamics as well as proliferation periods and rates between individual genomes of wheat could not be conducted in the study of SABOT et al. (2005) as, in the genomic sequences available at that time, the A genome was overrepresented whereas the B genome was underrepresented. By using more representative genomic sequences in the present study, we showed the more recent activation of the Copia and CACTA TEs in the wheat A genome but not in the B genome in which a more recent Gypsy proliferation is observed. Overrepresentation of the A genomic sequences in the study of SABOT et al. (2005) may explain the reason for which they found that Copia TEs have been most active in the wheat A, B and D genomes combined together. Thus our analysis, using representative sequence data sets, shows for the first time differential proliferation of TEs between the wheat A and B genomes and illustrates the inadequacy of combining together sequence data sets from different genomes as it has been previously done.

Four families (Angela, Wis, Sabrina, Fatima) were abundant, representing the majority of LTR-retrotransposons in the B and A genomes of wheat, some of which proliferated differentially (see results). Proliferation of specific types of TEs in specific genomes (or species), leading to rapid genome size variation and sequence divergence, has also been observed in other plant species. Analysis of maize (Zea mays) genomic sequences suggests that the high percentage of LTR retrotransposons is due to proliferation of just a few families of TEs (MEYERS et al. 2001; LIU et al. 2007; KRONMILLER and WISE 2008). Similarly, comparison of TE proportions between various cotton species (Gossypium species) revealed differential lineage-specific expansion of various LTR-retrotransposons superfamilies and families, leading to threefold genome size differences (HAWKINS et al. 2006). Species-
specific differential retrotransposon expansions are also the cause of size doubling of the *Oryza australiensis* genome as compared to cultivated rice (*O. sativa*) (PIEGU et al. 2006).

**Neither enhancement nor repression of transposable elements proliferation following allotetraploidization**

As estimated from their insertion dates and confirmed by PCR-based tracing analysis, the majority of differential proliferation of TEs in B and A genomes of wheat (87% and 83% respectively) occurred prior to the allotetraploidization event that brought them together in *T. turgidum* and *T. aestivum*, less than 0.5 MYA (HUANG et al. 2002; DVORAK et al. 2006, CHALUPSKA et al. 2008). More importantly, the allotetraploidization event appears to have neither enhanced nor repressed retrotranspositions. We suggest that in addition to the *Ph1* gene, preventing homeologous recombination (GRIFFITHS et al. 2006), differential proliferation of TEs has also contributed to the rapid divergence of the B and A genomes of the wheat diploid progenitors and the relative stability of the natural wheat allopolyploids that occurs thereafter.

Different levels of stability, estimated as elimination of DNA sequences, were observed in newly synthesized wheat allopolyploids, depending on wheat genome combinations (FELDMAN and LEVY 2005 and our unpublished results). The natural wheat allopolyploids combining the B and A genomes are relatively stable and cannot be exactly re-synthesized, because the diploid progenitor of the B genome is unidentified (FELDMAN et al. 1995; BLAKE et al. 1999; HUANG et al. 2002; DVORAK et al. 2006). Nevertheless, by studying a synthetic wheat allotetraploid, combining the A and S genomes (the closest identified diploid relatives to the progenitors of the A and of the B genome of natural wheat polyploids), the same group
(Kashkush et al. 2003) reported on transcriptional activation of the *Wis* LTR retrotransposon but not its transposition following allotetraploidization. This is in agreement with the lack of enhancement of transpositions observed in this study in wheat natural allopolyploids combining the A and B genomes. Comparatively, less TE proliferation, estimated as increased rate of deletions and decreased rate of insertions, was recently observed in the cotton polyploid species *Gossypium hirsutum* as compared to its diploid progenitors *G. arboretum* and *G. raimondii* (GROVER et al. 2008).

**Apparent transposable element proliferation as a balance between their transposition and removal evolutionary forces**

As in the present study, the vast majority of complete retrotransposons studied so far were also estimated to be less than three million years old (SANMIGUEL *et al.* 1998, 2002; WICKER *et al.* 2003b, 2005; GAO *et al.* 2004; MA *et al.* 2004; DU *et al.* 2006; PIEGU *et al.* 2006; WICKER and KELLER 2007). These findings imply that there are mechanisms of active deletion of LTR retrotransposons from the genome, such as unequal homologous recombination and illegitimate recombination (VICIENT *et al.* 1999; DEVOS *et al.* 2002; MA *et al.* 2004; PEREIRA 2004). Proliferation periods and rates estimated for TEs at a given evolutionary period are the result of both antagonist evolutionary forces: TE insertion activity (transpositions) (BENNETZEN and KELLOGG 1997) but also their removal (PETROV *et al.* 2000; PETROV 2002a). Thus, it is not clear whether the insertions and/or truncation (removal) rates of TEs are constant or vary during genome evolution. ‘Burst of insertions’, so far described for TEs, could correspond to periods of (i) high insertion activity, (ii) low rates of TE removal and/or (iii) combinations of both evolutionary forces.
The fact that *Copia* retrotransposons have been active until recently in the *Arabidopsis thaliana* genome allowed Pereira (2004) to calculate a rate of their elimination (or half-life) as 472,000 years, outside of centromeric regions. Using this method and assuming that repetitive sequences are removed from the genome at a constant rate, a higher half-life (79,000 years) was calculated for *Copia* removal in rice (Wicker and Keller 2007). As the insertion-date distribution of *Copia* retrotransposons in Triticeae (wheat and barley) is not exponential, Wicker and Keller (2007) suggested that their half-life is much longer than in rice, thus representing a major difference between small and large genomes of plants. Similar distributions are observed in our study for all three retrotranspon superfamilies in both B and A genomes of wheat. Our analysis suggests that lower proliferation of the LTR-retrotransposons during the most recent period could account for these apparent non-exponential distributions of insertion dates (including *Copia* retrotransposons) (Figure 5).

Our study clearly shows that, during their evolution, specific types of TEs have undergone differential proliferation in specific wheat genomes (or species) but not in others, leading to rapid sequence divergence. Little is known about the mechanistic causes that lead to differential proliferation of a single or related group of TEs across the genome of a specific species. These rapid TE expansions could correspond to periods of relaxed selection pressure such as genome duplication, interspecific hybridizations (although this was not revealed in our study), or stress conditions. It is also possible that TE proliferation could be caused by advantageous mutations in the TE sequence. A third alternative is differential deregulation of epigenetic silencing that allows specific TE families to proliferate in specific genomes.
ACKNOWLEDGEMENTS

This project was supported by the National Center for Sequencing (CNS-Génoscope)/APCNS2003-Project: *Triticum species* comparative genome sequencing in wheat (http://www.genoscope.cns.fr/externe/English/). PCR-based tracing of retrotransposons insertions was funded by the ANR (Agence Nationale pour la Recherche)-Biodiversité project (ANR-05-BDIV-015) and the ANR-05-Blanc project-ITEGE. We sincerely thank Dr. J. Dolezel, and Dr. M. Kubalakova (Institute of Experimental Botany, Olomouc, Czech Republic) for providing FISH mapping information for BAC clones B95G2, B95C9, B63B7, B54F7; Dr. Joseph Jahier (INRA-Rennes, France) and Professor Feldman (Weizmann Institute of Science, Israel) for valuable discussions and for providing wheat genotypes; Dr. Catherine Feuillet (INRA-Clermont-Ferrand, France) for providing the wheat deletion lines; Dr. Thomas Wicker (Zurich University, Switzerland) for valuable advice on novel transposable element classifications and *CACTA* TE evolution; Dr. Piotr Gornicki (University of Chicago, Chicago) and anonymous reviewers for valuable discussion and constructive criticisms; Dr. Heather McKhann (CNG, EPGV-INRA, Evry, France) for valuable discussion and revision of the manuscript.
LITERATURE CITED


ISIDORE, E., B. SCHERRER, B. CHALHOUB, C. FEUILLET and B. KELLER, 2005 Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. Genome Res. 15: 526-536.


PATERSON, A. H., J. E. BOWERS and B. A. CHAPMAN, 2004 Ancient polyploidization predating


PEREIRA, V., 2004 Insertion bias and purifying selection of retrotransposons in the Arabidopsis thaliana genome. Genome Biol. 5: R79.


FIGURE LEGENDS

Figure 1. Detailed annotation, BIN map positions and sequence composition of 10 sequenced BAC clones of wheat chromosome 3B.

(A) Detailed annotations of the 10 sequenced BAC clones. Main TEs, other repeats and genes sequence information (GSI) are represented with distinct features and motifs (detailed in the feature details box). (g) genes, (pg) putative genes, (gr) gene-relics, and (psg) pseudogenes. For nested insertions of TEs: the newly inserted TE is presented above the split one. Complete reconstruction of split TEs was done and the different parts are linked with a line to visualize the entire element. Some BAC clones are represented by several unordered contigs (TA3B63E4, TA3B63C11, TA3B63N2). EMBL BAC clone references and annotation files are given in the Methods section and deposited at the EMBL / GenBank Data Libraries. Detailed coding sequence and TE descriptions are supplied in Supplemental Text 1 and Supplemental Text 2. Arrows indicate novel TEs identified in this study, described in Supplemental Text 2 and Supplemental Table 5.

(B) BIN map position of nine of the BAC clones. The wheat chromosome 3B BINs are according to Qi et al. (2003). Details of genotyping results are given in Supplemental Table 6.

(C) Proportions of the main sequence classes and types. Features and motifs legends are the same as for Figure 1A. Details are given in Supplemental Table 2

Figure 2. Changes of coefficient of variation of proportions (%) of the main transposable element superfamilies calculated over all possible BAC clone combinations and simulated
over a size varying from one to 12 BAC clones for the wheat B genome and one to 18 for the wheat A genome (combination size).

For each number of considered BAC clones (X axis), sequence proportions (%) were calculated for all possible BAC clone combinations and the coefficient of variation between these proportions was calculated (Y axis).

**Figure 3.** Distribution of insertion dates estimated for LTR-retrotransposons in the B and A genomic sequences of wheat (divergence and MYA).

(A) All dated LTR-retrotransposons combined at the three main superfamily levels (*Athila*, *Gypsy* and *Copia*).

(B) The most abundant retrotransposon families, showing five or more dated copies in at least one of the A or B wheat genomes.

Mean insertion dates calculated for retrotransposons are represented with vertical bars. For the A genomic sequences: ‘blue’ for retrotransposons detected from the diploid and ‘red’ from the polyploid genomic sequences. The genomic sequences of the B genome (red) were obtained from the polyploid wheat. Copies of a given retrotransposon superfamily or family showing identical mean insertion dates are presented by adjacent vertical bars that are joined with a lower horizontal gray bar. The number between brackets corresponds to the total number of considered retrotransposon copies. Gray triangles indicate retrotransposon insertions that have been traced using PCR in a collection of genotypes of *T. aestivum* and *T. turgidum*. The interval-period of the allotetraploidization event (0.5-0.6 MYA, divergence 0.013-0.016) is highlighted in gray.
Uniformity Test: Kolmogorov-Smirnov (Féignac 1962) tests determining probabilities (P-value) that the distribution of insertion dates of retrotransposons deviate from uniformity (thus confirming a burst of higher proliferation): All: for the last three million years (0.078 divergence); ‘Recent’: for the most recent periods, estimated when dividing the LTR-retrotransposon insertions by the median (indicated by gray dots). Tests were done on those families that show five copies or more.

Comparison of distributions: The same Kolmogorov-Smirnov tests determining probabilities that distributions of insertion dates for the last three million years (0.078 divergence) are different between the different retrotransposons superfamilies and families as well as the B and A genomes of wheat.

Figure 4. PCR-based tracing of series of retrotransposons, inserted at different dates in wheat chromosome 3B, across a collection of wheat tetraploid (T. turgidum) and hexaploid (T. aestivum) genotypes. Averages and intervals (means ± Standard errors) of retrotransposon insertion dates are presented. The interval of the allotetraploidization event (0.5-0.6 MYA), calculated according to gene sequence divergence (Huang et al. 2002; Dvorak et al. 2006, Chalupská et al. 2008), is in gray. Retrotransposons for which insertion dates were estimated based on the divergence of their LTR, prior to the tetraploidization event were generally detected in almost all genotypes, whereas those posterior to the tetraploidization event were detected in only some genotypes. Gels show PCR-based detection of: insertion of \textit{RLC_{Alixa TA3B95C9-1}} (into \textit{DTC_{Caspar TA3B95C9-1}}) in all tested genotypes, except one; insertion of \textit{RLG_{Nathalia TA3B63E4-1}} (into \textit{DTC_{Vincent TA3B63E4-2}}) in some genotypes. Primer sequences, details of insertions dates (averages and intervals), and PCR-
based detection in different wheat genotypes are given in Supplemental Table 1 and Supplemental Table 7.


**Figure 5.** Proliferation periods and rates of the main retrotransposon superfamilies in the wheat B and A genomes.

Expressed as probability density functions, where the area under each curve was calculated based on estimated insertion dates of retrotransposons (presented in Figure 3) and their corresponding standard errors, using Gaussian kernel density estimation (Silverman 1986). The curves have been scaled with respect to the number of observations, so that the sum of their areas (given for each retrotransposon superfamily in the legend box) equals the probability one and comparisons between genomes and retrotransposon superfamilies can be performed. When calculated standard errors were very low, a minimum value of 80,000 years
(corresponding to 0.002 divergence) was used. The grayed-out field is due to uncertainty in very recent insertion date estimations.
Table 1: Details of transposable elements (TEs) from the four most represented superfamilies, in 13 genomic regions of the wheat B genome (1.98 Mb), compared to publicly available sequences from 19 genomic regions of the wheat A genome (3.63 Mb).

<table>
<thead>
<tr>
<th></th>
<th>13 genomic regions of the wheat B genome*</th>
<th>19 publicly available genomic regions of the wheat A genome**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Athila</td>
<td>Copia</td>
</tr>
<tr>
<td>Observed number of TEs</td>
<td>54</td>
<td>57</td>
</tr>
<tr>
<td>Sequence proportion (means±SE)%***</td>
<td>10.8±1.6</td>
<td>14.2±2.5</td>
</tr>
<tr>
<td>Bootstrap means deviation****</td>
<td>-0.07</td>
<td>+0.02</td>
</tr>
<tr>
<td>Complete TEs with TSD (%)</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>Incomplete (truncated) TEs</td>
<td>41</td>
<td>39</td>
</tr>
<tr>
<td>-LTR-mediated homologous recombination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Entire TE without TSD</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>-Solo-LTR</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>-Illegitimate recombination</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td>Complete TEs / incomplete (truncated) TEs</td>
<td>0.32</td>
<td>0.46</td>
</tr>
</tbody>
</table>

* This corresponds to 1.43 Mb from the 10 genomic regions sequenced in this study and 0.55 Mb from three other publicly available genomic regions from Sabot et al. (2005), Gu et al. (2006), Dvorak et al. (2006). See Methods section for BAC clone sequence references.

** 19 genomic regions available for the A genome (Yan et al., 2002, 2003; SanMiguel et al., 2002; Wicker et al., 2003b; Chantret et al., 2005; Isidore et al., 2005; Dvorak et al., 2006; Gu et al., 2006; Miller et al., 2006). See Methods section for BAC clone sequence references.

*** Relative to cumulative sequence length. SE: Standard errors for estimated means.

**** Differences between arithmetic means (line above) and that from bootstrap analysis (Efron, 1979), with 10000 re-samplings.

Charles et al.
**Table 2:** Associations of *CACTA* transposable elements with the four most represented TE superfamilies and other DNA sequence classes, in 13 genomic regions of the wheat B genome (1.98 Mb) and 19 publicly available genomic sequences of the wheat A genome (3.63 Mb).

<table>
<thead>
<tr>
<th>DNA sequence classes</th>
<th>13 genomic regions of the wheat B genome*</th>
<th>19 publicly available genomic regions of the wheat A genome**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>CACTA</em> TEs inserted into other DNA sequences</td>
<td>Other DNA sequences inserted into <em>CACTA</em> TEs</td>
</tr>
<tr>
<td><em>Athila</em> TEs</td>
<td>12: 4 / 8 ***</td>
<td>6: 0 / 6</td>
</tr>
<tr>
<td><em>Copia</em> TEs</td>
<td>1: 0 / 1</td>
<td>5: 0 / 5</td>
</tr>
<tr>
<td><em>Gypsy</em> TEs</td>
<td>1: 1 / 0</td>
<td>6: 0 / 6</td>
</tr>
<tr>
<td><em>CACTA</em> TEs</td>
<td>7: 4 / 3</td>
<td>6: 1 / 5</td>
</tr>
<tr>
<td>Other TEs</td>
<td>3: 3 / 0</td>
<td>0: 0 / 0</td>
</tr>
<tr>
<td>Unclear TE associations</td>
<td>2: 0 / 2</td>
<td>2: 0 / 2</td>
</tr>
<tr>
<td>Unassigned DNA</td>
<td>44: 7 / 37</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>70: 19 / 51</td>
<td>25: 1 / 24</td>
</tr>
</tbody>
</table>

* *, ** Same sequences and references as in Table 1.

** Results are detailed as follows: Total *CACTA* TE copies: Complete *CACTA* TE copies / Truncated *CACTA* TE copies

**** From cases where we can’t be certain that a *CACTA* TE is inserted -into or -by another TE element.
SUPPLEMENTAL DATA FILES.

Supplemental Figure 1. Fluorescence in situ hybridization (FISH) pattern observed for 10 sequenced BAC clones from wheat chromosome 3B on mitotic metaphase chromosomes of *Triticum aestivum* cv. Chinese Spring (CS).

Supplemental Table 1. Sequences and positions on the corresponding BAC clones of PCR primers used for tracing of the TE insertions (Figure 4 and Supplemental Table 7) as well as for the mapping of the 10 BAC clones on the deletion BINs of wheat chromosome 3B (Figure 1 and Supplemental Table 6).

Supplemental Table 2. Overall (A) and detailed (B) sequence class proportion of 10 sequenced BAC clones of wheat chromosome 3B (cumulative length: 1.43 Mb).

Supplemental Table 3. Details of complete and truncated genes predicted from 10 sequenced BAC clones of wheat chromosome 3B and their classification (according to criteria developed in Supplemental Method 2).

Supplemental Table 4. Hypothetical genes (predicted based on FGENESH prediction only) found in 10 sequenced BAC clones of wheat chromosome 3B.

Supplemental Table 5. Characteristics of novel transposable elements (TEs) identified in the 10 sequenced genomic regions of wheat chromosome 3B. The similarities (BLASTN / BLASTX) and structural properties found are also presented.

Supplemental Table 6. Mapping of 10 sequenced BAC clones on wheat chromosome 3B deletion lines, with PCR-specific markers derived from transposable elements (TE)-TE junctions.
Supplemental Table 7. PCR-based tracing of series of retrotransposons, inserted at different dates in wheat chromosome 3B, across a collection of genotypes of tetraploid (T. turgidum) and hexaploid (T. aestivum) wheat species.

Supplemental Text 1. Sequence annotation of 10 BAC clones from wheat chromosome 3B: Gene prediction, description and synteny with rice.

Supplemental Text 2. Sequence annotation of 10 BAC clones from wheat chromosome 3B: Transposable element prediction, annotation, classification and composition.

Supplemental Method 1. Annotation method of the 10 BAC clone sequences from wheat chromosome 3B.

Supplemental Method 2. Fluorescent in situ hybridisation (FISH) protocol.
Figure 1

Features and motifs

A

B

C
Figure 2
### Table: Allotetraploidization

<table>
<thead>
<tr>
<th>Classification</th>
<th>Genome Count</th>
<th>Uniformity test (×100)</th>
<th>Comparison of distribution (×100)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gypsy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (41)</td>
<td></td>
<td>0.52 0.09</td>
<td></td>
</tr>
<tr>
<td>B (39)</td>
<td></td>
<td>38.02 1.08</td>
<td></td>
</tr>
<tr>
<td><strong>Copia</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (60)</td>
<td></td>
<td>12.86 0.00</td>
<td></td>
</tr>
<tr>
<td>B (18)</td>
<td></td>
<td>21.65 73.46</td>
<td></td>
</tr>
<tr>
<td><strong>Athila</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A (19)</td>
<td></td>
<td>1.35 9.73</td>
<td></td>
</tr>
<tr>
<td>B (13)</td>
<td></td>
<td>0.53 6.62</td>
<td></td>
</tr>
</tbody>
</table>

### Diagram: Figure 3

- **Classification:** Count
- **Genome:** Unifor ming test (×100)
- **Comparison of distribution:** (×100)

**Gypsy**
- Daniela A (5)
- Erika A (6)
- Fatima A (10)
- Jeli A (2)
- Laura A (5)

**Copia**
- Angela A (39)
- Wis A (12)

**Athila**
- Sabrina A (12)
- Wham A (7)
Figure 4

Features and motifs
- Athila
- Gypsy
- Copia
- Unknown
- PCR-based detection in all genotypes
- PCR-based detection in some genotypes

PCR-based detection in all genotypes
PCR-based detection in some genotypes
Figure 5