The Plant Cell, Vol. 17, 1033–1045, April 2005, www.plantcell.org © 2005 American Society of Plant Biologists

RESEARCH ARTICLES

Molecular Basis of Evolutionary Events That Shaped the *Hardness* Locus in Diploid and Polyploid Wheat Species (Triticum and Aegilops) [W]

Nathalie Chantret,[a,1] Jérôme Salse,[b,1] François Sabot,[c] Sadequr Rahman,[d] Arnaud Bellec,[b] Bastien Laubin,[c] Ivan Dubois,[e] Carole Dossat,[e] Pierre Sourdille,[c] Philippe Joudrier,[f] Marie-Françoise Gautier,[f] Laurence Cattolico,[e] Michel Beckert,[c] Sébastien Aubourg,[b] Jean Weissenbach,[g] Michel Caboche,[b] Michel Bernard,[c] Philippe Leroy,[c] and Boulos Chalhoub[b,2]

[a] Institut National de la Recherche Agronomique-Centre de Cooperation Internationale en Recherche Agronomique pour le Développement, Biotrop, F-34398 Montpellier Cedex 5, France
[b] Laboratory of Genome Organization, Unité de Recherches en Génomique Végétale-Institut National de la Recherche Agronomique, F-91057 Évry Cedex, France
[c] Unité Mixte de Recherche 1095-Institut National de la Recherche Agronomique Amélioration et Santé des Plantes, Domaine de Crouël, F-63039 Clermont-Ferrand Cedex 2, France
[d] Commonwealth Scientific and Industrial Research Organization Plant Industry, ACT 2601, Camberra, Australia
[e] GENOSCOPE, Consortium National de Recherche en Génomique, F-91057 Evry Cedex, France
[f] Institut National de la Recherche Agronomique-Unité Mixte de Recherche, Polymorphismes d'Intérêt Agronomique, 34060 Montpellier Cedex 01, France
[g] GENOSCOPE, Consortium National de Recherche en Génomique, Centre Nationale de Recherche Scientifique, Unité Mixte de Recherche 8030, F-91057 Evry Cedex, France

The *Hardness* (*Ha*) locus controls grain hardness in hexaploid wheat (*Triticum aestivum*) and its relatives (Triticum and Aegilops species) and represents a classical example of a trait whose variation arose from gene loss after polyploidization. In this study, we investigated the molecular basis of the evolutionary events observed at this locus by comparing corresponding sequences of diploid, tertraploid, and hexaploid wheat species (Triticum and Aegilops). Genomic rearrangements, such as transposable element insertions, genomic deletions, duplications, and inversions, were shown to constitute the major differences when the same genomes (i.e., the A, B, or D genomes) were compared between species of different ploidy levels. The comparative analysis allowed us to determine the extent and sequences of the rearranged regions as well as rearrangement breakpoints and sequence motifs at their boundaries, which suggest rearrangement by illegitimate recombination. Among these genomic rearrangements, the previously reported *Pina* and *Pinb* genes loss from the *Ha* locus of polyploid wheat species was caused by a large genomic deletion that probably occurred independently in the A and B genomes. Moreover, the *Ha* locus in the D genome of hexaploid wheat (*T. aestivum*) is 29 kb smaller than in the D genome of its diploid progenitor *Ae. tauschii*, principally because of transposable element insertions and two large deletions caused by illegitimate recombination. Our data suggest that illegitimate DNA recombination, leading to various genomic rearrangements, constitutes one of the major evolutionary mechanisms in wheat species.

INTRODUCTION

Genomic duplications have been proposed as a major force of genome evolution because they supply genetic raw material for more tolerable DNA diversification and allow species to evolve new functions and adapt to a wide range of habitats and environmental conditions (Ohno, 1970). Contrasting with duplications of individual genes or genomic segments, polyploidy consists of whole genome duplication either by doubling the genome of one species (autopolyploidy) or through hybridization of two related species and chromosome doubling (allopolyploidy). Although still poorly understood, polyploidy is widespread in plant species and has already been shown to induce genetic and epigenetic events, including sequence elimination and gene silencing (Mayer and Aguilera, 1990; Wendel and Wessler, 2000; Eckardt, 2001; Ozkan and Feldman, 2001; Ozkan et al., 2001; Pikaard, 2001; Shaked et al., 2001; Kashkush et al., 2002, 2003; Kellis et al., 2004).

[1] These authors contributed equally to this work.
[2] To whom correspondence should be addressed. E-mail chalhoub@evry.inra.fr; fax 33-160874549.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Boulos Chalhoub (chalhoub@evry.inra.fr).
[W] Online version contains Web-only data.
Article, publication date, and citation information can be found at www.plantcell.org/cgi/doi/10.1105/tpc.104.029181.

Domestication of plant species is another important factor that affects genome evolution. First, domestication can favor spreading (and even formation) of interesting polyploid species. The hexaploid bread wheat (*Triticum aestivum*), for example, appeared 7000 to 9500 years ago, probably favored by human cultivation (Nesbitt and Samuel, 1995) when early farmers started domestication of the Triticum species (Feldman et al., 1995). Second, domestication involves few individuals and consequently leads to reduced genetic diversity of cultivated plant species as compared with their wild progenitors (Tanksley and McCouch, 1997). Third, domestication leads also to the selection and spreading of specific genes and alleles that control traits of agronomic importance and distinguish crops from their wild relatives (Clark et al., 2004).

Within the Poaceae, the grass genera Aegilops and Triticum constitute an interesting model to study effects of polyploidy and domestication on plant genome evolution. They include several diploid species (2n = 14) that, via allopolyploidization, produced several tetraploid and hexaploid wheat species, most of which have been domesticated (Figure 1) (Feldman et al., 1995; Eckardt, 2001; Huang et al., 2002). *T. turgidum* (2n = 28, AABB) was derived from a hybridization event that happened ~0.5 to 3 million years ago between a diploid donor of the A genome (*T. monococcum* ssp *urartu*, 2n = 14, AA) and another unknown species of the section Sitopsis (close to *Ae. speltoides*), donor of the B genome (2n = 14, BB) (Figure 1) (Feldman et al., 1995; Blake et al., 1999; Huang et al., 2002). The hexaploid wheat (*T. aestivum*, 2n = 21, AABBDD) originated from an additional

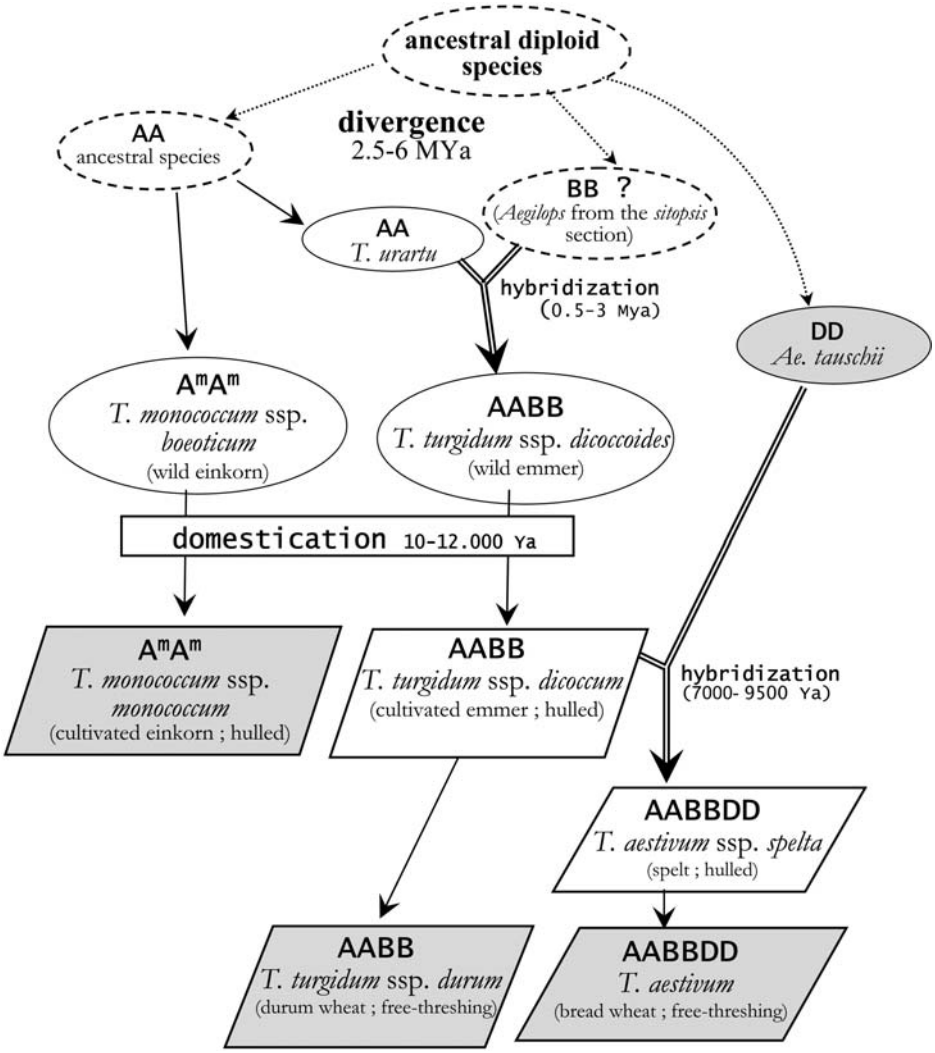


Figure 1. Schematic Representation of the Evolutionary History of Wheat Species (Triticum and Aegilops).

Wild and domesticated species are represented in circles and squares, respectively. Ancestral species or unknown species are surrounded by a dotted circle, whereas species surrounded by a plain line circle are actual species. Species from which the *Ha* locus region has been sequenced in this study are highlighted in gray. Estimated dates for divergence, polyploidization, and domestication are taken from Feldman et al. (1995), Nesbitt and Samuel (1995), Allaby et al. (1999), Eckardt (2001), and Huang et al. (2002).

polyploidization event between the early domesticated tetraploid *T. turgidum* ssp *dicoccum* and the diploid donor of the D genome, *Ae. tauschii* (2n = 14, DD), 7000 to 9500 years ago (Figure 1).

Grain endosperm texture or grain hardness (i.e., whether the endosperm is physically hard or soft) is an important trait of these species and determines their use (Giroux and Morris, 1998; Morris, 2002). The cultivated tetraploid durum wheat (*T. turgidum* ssp *durum*) is mainly used for pasta because its endosperm is hard. In hexaploid bread wheat (*T. aestivum* ssp *aestivum*), hardness is of crucial importance, and both soft and hard varieties are selected and marketed (reviewed in Morris, 2002). It is controlled by the *Ha* (*Hardness*) locus located on the short arm of chromosome 5D, at a subtelomeric position (Law et al., 1978; Sourdille et al., 1996). The *Ha* locus encodes friabilins that are composed of three related 15-kD lipid binding proteins: puroindoline a, encoded by the *Pina* gene, puroindoline b, encoded by the *Pinb* gene, and a less abundant protein called Grain Softness Protein, encoded by the *Gsp-1* gene (Gautier et al., 1994; Rahman et al., 1994).

The *Ha* locus represents a classical example of a locus controlling a trait whose variation arose as a result of gene elimination after polyploidization. The *Pina* and *Pinb* genes are absent from the A and B genomes of *T. turgidum* subspecies, domesticated and wild, but present in all their diploid progenitors (Gautier et al., 2000). These observations suggest that this particular evolution of the *Ha* locus occurred only in tetraploid wheat species, probably early after the tetrapolyploidization events (Figure 1). *Pina* and *Pinb* genes were restored in hexaploid wheat (*T. aestivum*, 2n = 21, AABBDD), upon allopolyploidization between the diploid D genome donor *Ae. tauschii* and the early domesticated tetraploid *T. turgidum* ssp *dicoccum*, ~7000 to 9500 years ago (Figure 1) (Feldman et al., 1995). The *Ha* locus in hexaploid wheat has been subjected to human selection, leading to hard wheat cultivars because of mutations or genomic rearrangements (such as *Pina* gene loss) (Giroux and Morris, 1998; Giroux et al., 2000; Morris, 2002).

We know very little about the molecular basis of the evolutionary events that shaped the *Ha* locus in wheat. In this study, we investigated the molecular basis of genomic rearrangements that occurred at the *Ha* locus by comparing corresponding sequences of diploid, tetraploid, and hexaploid wheat species (Triticum and Aegilops), which diverged relatively recently (Figure 1).

RESULTS

BAC clones anchored by the *Gsp-1* gene from four species (*T. monococcum*, *Ae. tauschii*, *T. turgidum* ssp *Durum*, and *T. aestivum*), representing seven genomes in three different ploidy contexts, were isolated and sequenced (Figure 1). Comparative sequence analysis and annotation of genes, transposable elements (TEs; transposons and retrotransposons), and unassigned DNA are presented in Figure 2. Distribution of these different sequence classes and description of coding DNA sequences (CDSs) are detailed in Supplemental Tables 1 and 2 online. The entire annotation (EMBL files) for each of the seven Triticum genomic sequences can be found in Supplemental Data 2 online.

Pina, *Pinb*, and *Gsp-1* are three related genes previously shown to be implicated in hardness variability (Giroux and Morris, 1998; Giroux et al., 2000; Morris, 2002; Hogg et al., 2004). The *Ha* locus region in the genomes presenting the puroindoline genes (i.e., *T. monococcum*, *Ae. tauschii*, and the D genome of *T. aestivum*) was defined as the genomic segment that carries *Gsp-1*, *Pina*, and *Pinb* genes (Figures 2A and 2B). To delimit the *Ha* locus, especially in the genomes carrying only the *Gsp-1* gene and lacking the puroindoline genes (i.e., the A and B genomes of *T. turgidum* and *T. aestivum*), we determined orthologous CDSs at the 5′ and 3′ boundaries of the *Ha* locus region using Multipipmaker analysis (Schwartz et al., 2000) (Figure 2A) and comparison with the rice (*Oryza sativa*) orthologous region (Figure 3).

The sequence comparison shows that the TEs constitute an important proportion of the loci and that they are not present at orthologous positions (Figure 2), suggesting that they were active after the divergence of the A, B, and D Triticum and Aegilops genomes as well as after the polyploidization events.

5′ Boundary of the *Ha* Locus Region

The first CDS (*Gene1*) found upstream the *Gsp-1* gene in the seven Triticum genomes (Figure 2B) is also present in the orthologous rice region (Chantret et al., 2004) (Figure 3). This gene encodes a putative protein of 425 amino acids, which is 39% similar to the β-1,3-galactosyl-*O*-glycosyl-glycoprotein β-1,6-*N*-acetylglucosaminyltransferase (Core 2 branching enzyme) from *Homo sapiens*. Therefore, *Gene1* was designated as the *BGGP* gene. A *BGGP* gene homolog (named *GlcNAc* gene) was also found at the recently sequenced *Ha* locus from barley (*Hordeum vulgare*) (Caldwell et al., 2004). However, its position (gene order) is different from that of its wheat and rice homologs (Caldwell et al., 2004; discussed hereafter).

3′ Boundary of the *Ha* Locus Region

Comparison of the CDS of the D genome of *T. aestivum* with the orthologous rice region (Figure 3) extends the colinearity beyond the *Pinb* gene to a block of tandemly repeated genes (*Gene7*) and one gene (*Gene8*) of unknown function (Figures 2A and 2B). The block of *Gene8*, surrounded by one or more copies of *Gene7*, is also observed in the A and B genomes of *T. turgidum* and *T. aestivum*, thus delimiting the *Ha* locus region (Figures 2A and 2B).

The predicted proteins from the *Gene7* cluster are 39% similar to a cell division protein ftsH homolog (integral membrane protein belonging to the AAA ATPase family) from *Helicobacter felis.* Thus, *Gene7* repeats were named *ATPase* genes. Several wheat ESTs were found to match with *Gene8*. This gene encodes a 534–amino acid protein that is 43% similar to a hypothetical 60.0-kD protein from *Saccharomyces cerevisiae* (Q03795).

The conservations of intergenic sequence indicate that the block composed of *Gene8* and the three *ATPase* genes (*7-1*, *7-2*, and *7-3*), identified in all the studied A, B, and D wheat genomes (Figure 2B), is an ancestral organization present before their divergence. Three copies of *ATPase* genes and several *ATPase* duplication events were also identified at the barley *Ha* locus (Caldwell et al., 2004). As for the *BGGP* gene, described above,

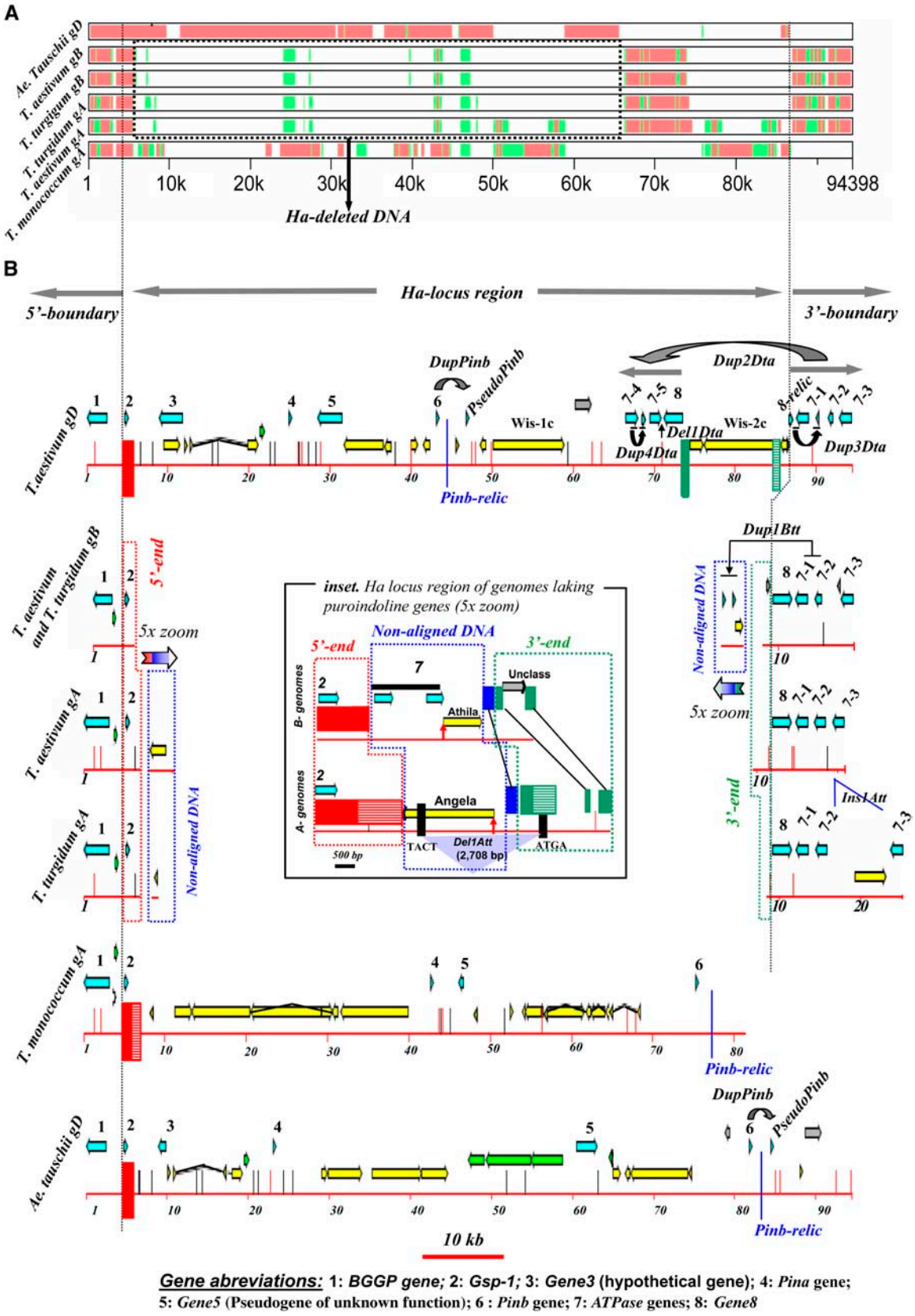


Figure 2. Identification of the *Ha* Locus Region and Comparative Annotation Overview of the Seven Wheat Species (Triticum and Aegilops) Genomes.

their position (gene order) is different from that of their wheat and rice homologs (Caldwell et al., 2004; detailed hereafter).

The D and B genomes of wheat contain additional duplications (complete or partial) of the *Gene8* and *ATPase* genes (Figure 2B). DOTTER program (Sonnhammer and Durbin, 1995) analysis and multisequence alignment between both intergenic and genic regions spanning the *Gene8/ATPase* gene (*7-1*, *7-2*, and *7-3*) interval in the D genome of *T. aestivum* against the A and B genome sequences showed that the block of *Gene8* and the three *ATPase* genes was duplicated in inverse orientation (Figure 2B, Dup2Dta). These data do not permit us to determine whether this event occurred before the A, B, and D genome divergence or exclusively in the D genome because this region was deleted from the A and B genomes of *T. turgidum* and *T. aestivum*. *Gene8* has been partially deleted in the D genome (after duplication), and only 221 bp showing a strong identity (93% nucleic acid identity and 89% AA identity) with the 3′ end of *Gene8* remain (called *Gene8*-relic) (Figure 2B). Further *ATPase* gene (*7-1*, *7-2*, and *7-3*) duplications or deletions are observed within the D genome of *T. aestivum* (Figure 2B). These additional *ATPase* gene rearrangements seem to be specific to the D genome of *T. aestivum* because they are not observed in the A or B genome (Figure 2B).

In the B genomes of *T. turgidum* and *T. aestivum*, in addition to the block of *Gene8* and the *ATPase* gene cluster (composed of one functional copy and two pseudogenes), two truncated *ATPase* genes were identified upstream of *Gene8* (Figure 2B, Dup1Btt). Coding and intergenic nucleic acid sequence comparison showed that these two truncated *ATPase* genes were generated by a duplication/inversion event, which involved the end of the *ATPase7-1* gene, the beginning of the *ATPase7-2* gene, and their intergenic interval. This duplicated region is 98% identical, suggesting a recent inversion/duplication event (Figure 2B) that differs from that observed in the D genome of *T. aestivum*.

The *Ha* Locus Region

Between the 5′ and 3′ boundaries defined above, the *Ha* locus varies in gene content, depending on the Triticum genome considered. In the A genome of *T. monococcum* and the D genomes of *Ae. tauschii* and *T. aestivum*, the *Ha* locus region carries the *Pina*, *Pinb*, and *Gsp-1* genes, two additional degenerated copies of the *Pinb* gene (*PseudoPinb* and *Pinb*-relic), and two nonrelated CDSs (*Gene3*, which is present in the D genomes only, and *Gene5*) (Figure 2B). *Pina* and *Pinb*, as well as *Gene3* and *Gene5*, are absent from the A and B genomes of *T. turgidum* and *T. aestivum* so that the *Ha* locus region contains only the *Gsp-1* gene in these two taxa.

Gsp-1 (*Gene2*) is 495 bp long, structured as a single exon present in all seven Triticum genomes. *Gene3* is hypothetical (see Methods for definition) and consists of one CDS that is present only in *Ae. tauschii* and the D genome of *T. aestivum*. The *Pina* gene (*Gene4*) is 447 bp long, structured as a single exon, and encodes the puroindoline a protein. *Gene5* lies between the *Pina* and *Pinb* genes of *T. monococcum*, *Ae. tauschii*, and the genome D of *T. aestivum* and shares 83% identity with the *T. aestivum* meiotic anther cDNA clone WHE3241_D05_G09 (CA498167, 678 bp). This gene was designated as a pseudogene because its CDS structure lakes the consensus GT/AG intron/exon transitory motifs. *Gene5* is shorter in *T. monoccocum* as compared with that present in the D genomes of *T. aestivum* and *Ae. tauschii*. The *Pinb* gene (*Gene6*) is 447 bp long and structured as a single exon that encodes the puroindoline b protein. A pseudogene (*PseudoPinb*) is present in the D genomes of *T. aestivum* and *Ae. tauschii* ~2500 bp downstream from the *Pinb* gene. Comparison of *PseudoPinb* and *Pinb* sequences revealed a conserved segment that comprises ~290 bp of the 5′ noncoding region (5′ untranslated region) and a highly degenerated CDS. In addition to the *PseudoPinb* (Figure 2B), a shorter duplication of *Pinb* gene or *PseudoPinb* (common sequences include <190 bp of 5′ untranslated region and 60 nucleotides of CDS) was also identified in the three genomes that contained the puroindoline genes (Figure 2B, called *Pinb*-relic).

Large Genomic Deletions from the *Ha* Locus of the A and B Genomes of Polyploid Wheat Species

The *Ha* locus region is relatively large in genomes containing the puroindolines genes compared with that of the A and B genomes

Figure 2. (continued).

(A) Multipipmaker (Schwartz et al., 2000) alignment using the sequence of the *Ha* locus of the D genome of *T. aestivum* as a matrix compared with the other six available wheat sequences. Colored blocks show the percentage of sequence identity (>90 in red; between 50 and 90% in green). The identified *Ha*-deleted interval in the A and B genomes of *T. turgidum* and *T. aestivum* is indicated (box surrounded by dotted line).

(B) Scaled diagram of annotation results of the *Ha* locus region (and 5′ and 3′ boundaries) from the seven Triticum genomes (detailed in Supplemental Data 1 online). Only one B genome is presented because the two B genomes share 99% sequence identity. Genes (CDS) (light blue), class I TEs (yellow blocks), class II TEs (green blocks), unclassified elements (gray), *MITE*s (vertical red bars), and repeats (vertical black bars) are shown. The remaining white spaces correspond to unassigned DNA (no features of annotation). Orthologous CDS between the different genomes have the same number or gene name. CDS duplication (Dup) and deletion (Del) events are indicated by arrows. Sequences of the genomes lacking the puroindoline genes (i.e., the A and B genomes of *T. turgidum* and *T. aestivum*) are continuous and are presented discontinuously to illustrate conserved 5′ and 3′ boundaries of the *Ha* locus region, conserved 5′ and 3′ ends of this same locus, and the DNA at the *Ha* locus region that cannot be aligned with the genomes presenting *Pina* and *Pinb* genes. The inset shows a 5× zoomed presentation of the *Ha* locus region of the A and B genomes of *T. aestivum* and *T. durum* (starting from *Gene 2* [*GSP-1* gene] to the 5′ end of *Gene8*) and detailing sequence conservation with the other genomes presenting the puroindoline genes at the 5′ and 3′ ends as well as the nonaligned sequences. Boxes represent sequence conservation between the different genomes at the 5′ and 3′ ends. The 5′ end of *Ha* locus region: conservations between the different Triticum and Aegilops genomes (red); conservation between the A genomes only (boxes with red lines). The 3′ end of *Ha* locus region: conservation with the D genome of *T. aestivum* at region flanking duplicated *Gene8* (green boxes); conservation with the D genome of *T. aestivum* at region flanking *Gene8*-relic (boxes with green lines); conservation between the A and B genomes only (blue).

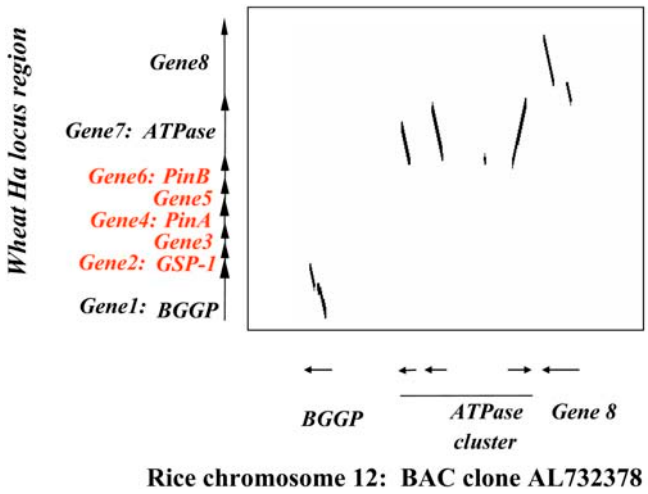


Figure 3. Microsynteny between Rice and Wheat at the *Ha* Locus.

The dot plot was performed using the DOTTER program (Sonnhammer and Durbin, 1995) with default parameters between the pseudomolecule (without TEs and repeats) of the D genome of *T. aestivum* (vertical) and rice BAC clone AL732378 (horizontal) sequences. Annotation features identified for these sequences are reported on the corresponding axes. Diagonals on the dot plot output represent nucleotide conservation between the two sequences analyzed. Gene numbers and names are as in Figure 2. Wheat genes that are deleted in the A and B genomes of *T. turgidum* and *T. aestivum* or that do not have homologs on the orthologous rice region are in red.

of polyploid wheat species, where the puroindoline genes are absent. In the D genome of *T. aestivum*, the entire *Ha* locus region (from the *Gsp-1* gene included in the 5′ end up to the 3′ end of *Gene8*) covers 82,353 bp (Figure 2B). In *T. monococcum* and *Ae. tauschii*, the available sequences of the *Ha* locus region (from the *Gsp-1* gene included in the 5′ end up to the 3′ end of the BAC because *Gene8* is not available) cover 75,156 and 89,806 bp, respectively (Figure 2B). This corresponds to an orthologous segment of only 61,135 bp compared with that available for the D genome of *T. aestivum*. In the wheat genomes lacking the puroindolines genes (i.e., the A and B genomes of *T. turgidum* and *T. aestivum*), the *Ha* locus region covers 3923 and 5323 bp, respectively, for *T. turgidum* and 6630 and 5321 bp for *T. aestivum*. The *Gsp-1* gene is the only CDS present in the *Ha* locus region of these genomes, whereas *Pina*, *Pinb*, *PseudoPinb*, *Pinb-relic*, *Gene3*, and *Gene5* are absent (Figures 2A and 2B). These results show that the previously reported loss of *Pina* and *Pinb* genes from both the A and B genomes of polyploid wheat species (*T. turgidum* and *T. aestivum*) (Gautier et al., 2000) is due to one or more genomic deletions of the whole genomic segment carrying *Gene3*, *Pina*, *Gene5*, *Pinb*, and *PseudoPinb* genes, rather than sequence divergence (Figures 2A and 2B).

Deletions of *Pina* and *Pinb* Genes Occurred Independently in the A and B Genomes of Polyploid Wheat Species

The extent of sequence homology between the different wheat genomes was analyzed to determine sequence conservation breakpoints (abrupt end of clear sequence similarity) and the genomic sequence at which the deletion(s) of the genomic segments carrying the *Pina* and *Pinb* genes occurred.

At the 5′ end of the *Ha* locus region, different extents of sequence conservation between the A, B, and D genomes were observed (Figure 2B, inset). For the A genomes of *T. turgidum* and *T. aestivum*, sequence conservation covers the *Gsp-1* gene and extends to 1473 bp at the 3′ end of the *Gsp-1* gene of the A genome of *T. monococcum*, the closest related diploid *Triticum* genome analyzed in this study (Figure 2B, inset). One conservation breakpoint corresponds to the insertion site of an imperfect *Angela* solo long terminal repeat (LTR) in the A genomes of both *T. turgidum* and *T. aestivum*; this LTR is absent in *T. monococcum* (Figure 2B, inset). For the B genomes, maximum sequence conservation extends to 754 bp at the 3′ end of the *Gsp-1* gene of the D genome of *T. aestivum*. Thus, the breakpoint of sequence conservation is distinct from the one identified for the A genomes (Figure 2B, inset) and is located 130 bp before the beginning of the relatively recent inversion/duplication of two truncated copies of the *ATPase* genes discussed above (Figure 2B).

At the 3′ end of the *Ha* locus (flanking the 5′ end of *Gene8*), the A and B genomes of *T. aestivum* and *T. turgidum* contain several stretches of conserved sequences found also in the D genome of *T. aestivum* and disrupted by a miniature inverted repeat transposable element (MITE) and an unclassified retroelement (Figure 2B, inset). The A genome of *T. aestivum* and B genomes of both *T. aestivum* and *T. turgidum* have nearly identical breakpoints of sequence conservation, which differs from that of the A genome of *T. turgidum* because of a 2708 bp deletion (Figure 2B, inset, *Del1Att*).

The sequence remaining between these two breakpoints could not be aligned between the A and B genomes of *T. turgidum* and *T. aestivum* or to any other Triticum genome (Figure 2B, inset, termed nonaligned DNA). The A genomes contain the truncated imperfect *Angela* retroelement (Figure 2B, inset). The two B genomes harbor completely different sequences composed of one retroelement relic belonging to the class I *Athila* LTR retrotransposon, in addition to two truncated recently duplicated *ATPase*-relics (Figure 2B, inset).

The presence of different truncated retroelements in the *Ha* locus suggests their involvement in the deletion of the *Pina* and *Pinb* genes from the A and B genomes of tetraploid and hexaploid wheats, apparently by illegitimate recombinations. These events also appear to have occurred independently. The breakpoint of the truncated *Athila* retroelement in the B genomes and the breakpoint of the truncated *Angela* retroelement in the A genomes could correspond to those of deletion events at the *Ha* locus (Figure 2B, inset, red arrows).

Other Genomic Rearrangements Observed at Different Ploidy Levels

In addition to the large deletions described above, other genomic rearrangements are also evident between the A and D genomes of different ploidy levels. Only the B genomes of *T. aestivum* and *T. turgidum* did not show any rearrangements.

High Sequence Conservation of the B Genomes of Allopolyploid Wheat Species

The *Ha* locus region and 5′ and 3′ boundaries of the B genome of the hexaploid *T. aestivum* (19,274 bp) show 99% sequence identity with that of the B genome of the tetraploid *T. turgidum* (19,229 bp) and only 27 small indels (1 to 16 bp in a TA-repeated microsatellite sequence motif), and 33 single nucleotide polymorphisms account for the overall difference (only one B genome is presented in Figure 2B).

Genomic Rearrangements in the A Genomes of Allopolyploid Wheat Species

The *Ha* locus region and 5′ and 3′ boundaries of the A genome of *T. turgidum* is 25,216 bp, whereas that of the A genome of *T. aestivum* is 20,745 bp. The first genomic rearrangement event responsible for this size difference is a 2708 bp deletion from the *T. turgidum* A genome relative to the *T. aestivum* A genome (Figure 2B, *Del1Att*, inset). Sequence comparison between the two very closely related A genomes allows the exact identification of the deletion breakpoint and an antisense/complement TACT/ATGA sequence motif that flanks the deletion (Figure 2B, *Del1Att*, inset). Therefore, the observed deletion might have been generated by an illegitimate recombination event in the A genome of *T. turgidum* ssp *durum* compared with that of *T. aestivum*, which occurred after the deletion of puroindoline genes (both genomes have a common A genome ancestor derived from the tetraploid *T. turgidum*). The second genomic rearrangement event, which is responsible for the size difference between the two closely related A genomes, is at the 3′ boundary of the *Ha* locus. The A genome of *T. turgidum* contains a fragment of 7325 bp, located between *ATPase7-2* and *ATPase7-3*, that is absent in the *A* genome of *T. aestivum* (Figure 2B, *Ins1Att*). This corresponds to two *LINE* insertions in the A genome of *T. turgidum.*

Beside these two major rearrangement events, the remaining sequences show a similar level of sequence conservation (97% identity).

Genomic Rearrangements at the Ha Locus of the D Genomes

The region between *Pina* and *PseudoPinb* is two times shorter in the D genome of *T. aestivum* compared with that of *Ae. tauschii* (42,575 bp versus 80,171 bp). Figure 4 illustrates sequence conservation between *Ae. tauschii* and the D genome of *T. aestivum.* Seven colinearity breakpoints corresponding to TE insertions or genomic inversions or deletions are observed (Figure 4, breakpoints 1 to 7).

Two TE insertions in the D genome of *T. aestivum* correspond to synteny breakpoints 1 and 7 (Figure 4, Teins1Dta). Breakpoint 1 (Figure 4, TEins1Dta) is located within *Gene3* where the second intron is larger in the D genome of *T. aestivum* than *Ae. tauschii* (1986 bp versus 110 bp) because of insertion of a retroelement (1872 bp) that we called *Morgane* (Figure 4). The LTR_STRUC prediction program (McCarthy and McDonald, 2003) and alignment of the *Morgane* retroelement against itself allowed us to identify complete 5′ and 3′ end LTRs of 287 bp, surrounded by a GTGGC direct repeat. Breakpoint 7 (Figure 4, TEins4Dta) corresponds to the insertion of the copia LTR retrotransposon *WIS* (*WIS*_1611A10_1.1) in the D genome of *T. aestivum D*, absent in the orthologous *Ae. tauschii* region. Breakpoints 3 (Figure 4, TEins2Dat) and 5 (Figure 4, TEins3Dat) correspond to TE insertions (*Fatima*_BAC_10_1.1, *Wilma*_ BAC_10_1.1, and *Sabrina*_BAC_10_1.1, respectively) within the D genome of *Ae. tauschii*.

A large inversion is evident between the two D genomes (Figure 4). In *Ae. tauschii*, the 5′ sequence flanking this inversion contains a CCGATGTTATTA motif (present also in the D genome of *T. aestivum*), which is also found in inverse (complementary) orientation (TAATAACATCGG) at the 3′ sequence flanking the inversion (in *Ae. tauschii* only). It is possible that this motif facilitated the large inversion by illegitimate recombination via cDNA strand exchange during replication. The site of DNA strand exchange would correspond to these motifs.

Two deletions in the *Ha* locus region of the D genome of *T. aestivum*, compared with the D genome of *Ae. Tauschii*, were observed (Figure 4, breakpoints 4 and 6). At breakpoint 4, a large genomic fragment (15,486 bp) is deleted in the D genome of *T. aestivum* (Figure 4, Del3Dta). In *Ae. tauschii*, a GGTTCCCTG motif is present at each end of the deleted fragment in the same orientation. In the D genome of *T. aestivum*, one copy of the same motif is found at the deletion breakpoint position (Figure 4, Del3Dta). These results show that this deletion (Del3Dta) is probably due to an illegitimate recombination involving both GGTTCCCTG motifs flanking the deleted sequence, which occurred by replication slippage. The synteny breakpoint 6 (Figure 5, Del2Dta) corresponds to a deletion of another large DNA fragment (6538 bp) in the D genome of *T. aestivum*, which may also have been caused by illegitimate recombination involving a CAT-ATTAA/TTAATCATG complement/reverse motif flanking the deleted sequence (Figure 4).

DISCUSSION

Both expansion and reduction mechanisms of wheat genomes have been revealed in this study. TEs appear to have been very active since the divergence of the A, B, and D Triticum genomes as well as after the allopolyploidization events. Such activity results in genome size increases, as previously described in plants (Bennetzen, 2002; Devos et al., 2002; Ramakrishna et al., 2002; Wicker et al., 2003; Ma et al., 2004). The observed genomic duplications also result in genome size increases. By contrast, several genomic deletions were observed that apparently resulted from illegitimate recombination events. The high rate of genomic rearrangements observed at the *Ha* locus provides evidence for a very plastic and dynamic nature of plant genomes (Petrov, 2001; Bennetzen, 2002).

Evolution of the *Ha* Locus in the Pooideae Subfamily by Duplication of the Puroindoline Genes

Gsp-1, *Pina*, and *Pinb* are present only in species of the Pooideae subfamily, such as the genera Triticum, Aegilops, Hordeum (barley: hordoindoline), Secale (rye: secaloindoline), and Avena (avenoindoline), where soft endosperm texture is a dominant trait

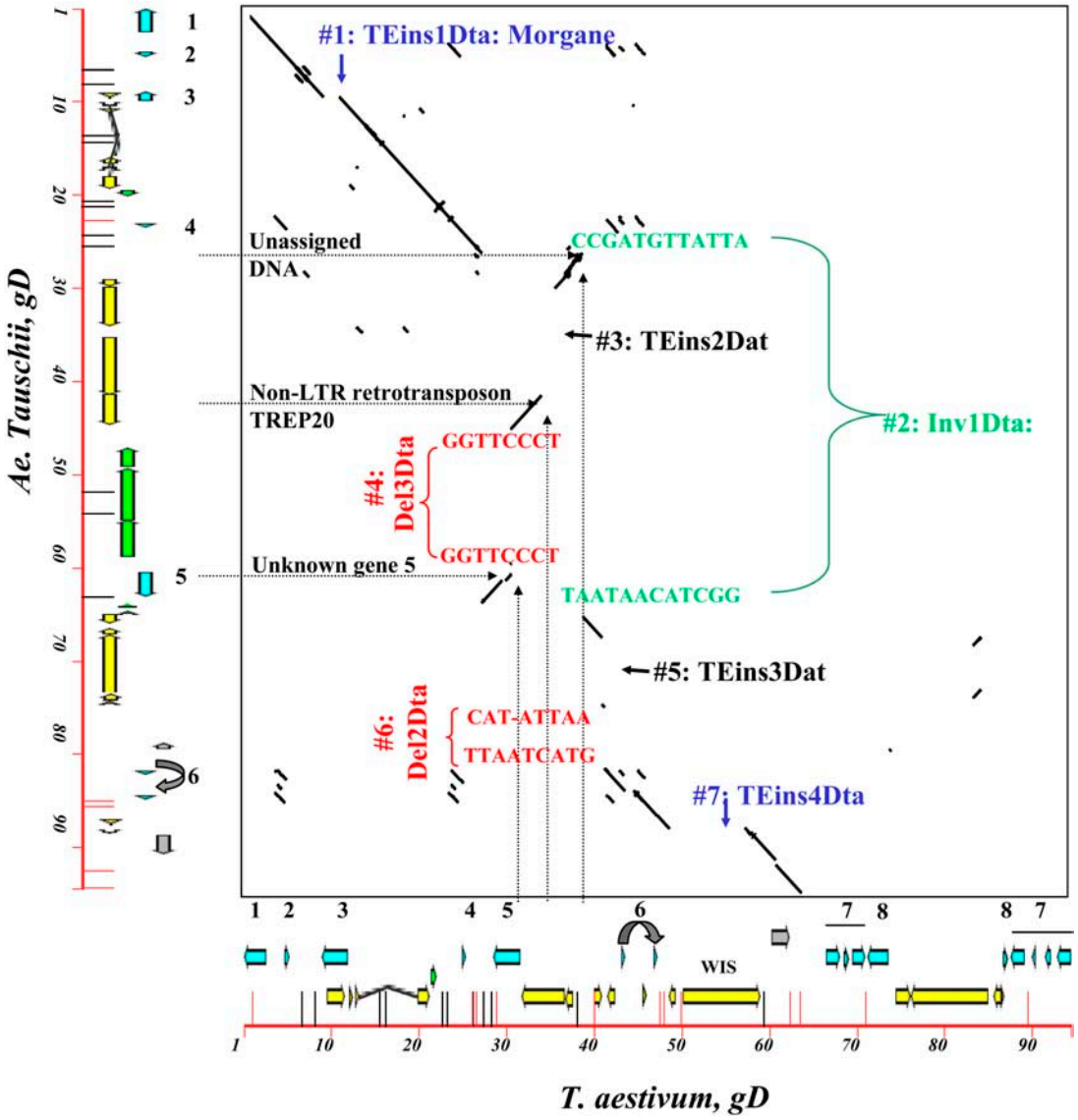


Figure 4. Comparison of the D Genome of *Ae. tauschii* with the D Genome of *T. aestivum*.

The dot plot was performed using the DOTTER program (Sonnhammer and Durbin, 1995) with default parameters between *Ae. tauschii* (vertical) and the *T. aestivum* D genome (horizontal) sequences. Annotation features identified for these sequences are reported on the corresponding axes. Gene numbers and names as well as color codes for TEs and other DNA sequence classes are as in Figure 2. Diagonals on the dot plot output represent nucleotide conservation between the two analyzed sequences. Synteny breakpoints representing rearrangements are shown with numbers (from 1 to 7): TEins, TE insertion; Inv, inversion; Del, deletion. TE insertion events in the *Ae. tauschii* (black). TE insertion events in the D genome of *T. aestivum* are in blue. Inversion events and identified sequence motifs at breakpoint (in *Ae. tauschii*) are in green. Deletion events in the D genome of *T. aestivum* and identified sequence motifs (in *Ae. tauschii*) are in red.

(Gautier et al., 2000; Darlingto et al., 2001; Morris, 2002). These genes were not detected in species of other Poaceae subfamilies, such as rice, maize (*Zea mays*), or sorghum (*Sorghum bicolor*) (Gautier et al., 2000). However, genes flanking the puroindoline genes cluster at 5′ and 3′ boundaries of *Ha* locus allow the identification of the rice orthologous region and reveal an ancestral organization in which wheat and rice are collinear for these boundary genes (i.e., *Gene1* [*BGGP gene*], the cluster of *ATPase* genes, and *Gene8*) (Figure 3). In the orthologous barley *Ha* locus, these boundary genes as well as *Gsp-1*, *Pina*, and *Pinb* genes were not found in the same order as that of wheat and rice (Caldwell et al., 2004). The authors (Caldwell et al., 2004) suggested genomic translocations and intrachromosomal rearrangements in the barley genome relative to the ancestral grass (rice and wheat) organization. In the orthologous rice region, no significant sequence similarity with the puroindoline genes was

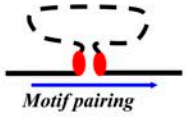
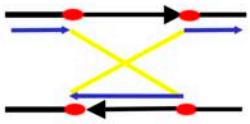
Events and extent of rearrangements	Breakpoints positions	Motifs types and sequence*	Proposed mechanism and schematic presentation
#1 Deletion of Pina and Pinb genes from the A genome of polyploid wheats (Figure2B)	TE/ND	ND	ND
#2 Deletion of Pina and Pinb genes from the B genome of polyploid wheats (Figure 2B)	TE/ND	ND	ND
#3 Del1Att (Figure 2B, inset):	TE/Unass	Antisense of complementary:	ND
#4 Del2Dta (Figure 4)	Unass/Unass	Complementary	 Motif pairing
#5 Del2Dta (Figure 4)	Unass/Unass	Forward	Recombination Slippage
#6 Inv1Dta (Figure 4)	Unass/Unass	Complementary	 Slipped-strand mispairing

Figure 5. Different DNA Illegitimate Recombination Events Observed at the *Ha* Locus Region, Breakpoint Positions (Sequence Classes in Which Rearrangements Occurred), Identified Sequence Motifs and Schematic Presentation of the Proposed Rearrangement Mechanism.

The precise sequence motif positions are supplied in Supplemental Data 2 online. ND, not determined; TE, TE sequences; Unass, unassigned sequences. References: 1 and 2 (Devos et al., 2002; Ma et al., 2004), 3 (none), 4 and 6 (Bzymek et al., 1999; Demura et al., 2002; Molinier et al., 2004; Tufarelli et al., 2004), and 5 (Demura et al., 2002). Red circles represent motifs observed at rearrangement breakpoints. Black lines indicate nonrearranged DNA segments; blue arrows indicate nascent rearranged DNA genomic segments. Yellow lines represent the direction of synthesis of nascent rearranged DNA.

detected in our study or in that of Caldwell et al. (2004). However, the presence of a putative *Gsp-1-relic* was previously suggested in rice based on the comparison with the *Ha* locus of *T. monococcum* (Chantret et al., 2004). Nevertheless, these different observations suggest that the puroindoline genes are specific to the Pooideae. On the other hand, the fact that *Pina* and *Pinb* genes are more similar to each other (70% AA similarity) than to the *Gsp-1* gene (56 to 58% AA similarity) suggests that they originated from two independent duplications of an ancestral Pooideae *Gsp-1* gene (Tranquilli et al., 1999; Chantret et al., 2004).

Other duplication events involving puroindoline genes were revealed in our study. In addition to the *PseudoPinb* found in the D genomes (*Ae. tauschii* and *T. aestivum*), the *Pinb-relic* is present in *Ae. tauschii*, *T. monococcum*, and the D genome of *T. aestivum* (Figure 2B). The presence of the same *Pinb-relic* in both D genomes and *T. monococcum* indicates that the duplication event occurred before the divergence of the A and D genomes. The equal sequence divergence in the common parts of the three duplicated regions did not allow us to determine whether the *Pinb*-relic is derived from a *Pinb* or *PseudoPinb* duplication event.

In barley, two homologs of the *Pinb* gene, the hordoindoline b genes (*Hindb1* and *Hindb2*), (GenBank accession number AJ276143; Darlingto et al., 2001; Caldwell et al., 2004) are closer to each other (95% of sequence identity) than to any wheat *Pinb* or *PseudoPinb* genes. These observations do not allow us to determine whether the duplication of *Pinb* and *Hindb* genes (and pseudogenes) took place in a common ancestor, followed by rapid divergence of nonfunctional *Pinb* genes in the Triticum genus, or from independent duplication events in the Triticum and Hordeum genera. The later hypothesis is favored by the frequency of duplication events revealed in this study.

Evolution of the *Ha* Locus in Polyploid Wheat Species by Genomic Deletions

If the *Ha* locus has evolved in the Pooideae by duplication of puroindoline genes, it also represents a classical example of a trait whose variations arose from gene elimination after polyploidization. We showed that the previously reported *Pina* and *Pinb* gene loss from the A and B genomes of polyploid wheat species (Gautier et al., 2000) is due to one or more genomic deletions of a large genomic segment containing additional genes (*Gene3*, *Gene5*, *PseudoPinb*, and *Pinb-relic*) rather than sequence divergence. These gene deletion event(s) occurred only in the A and B genomes of allopolyploid wheat species (wild and domesticated), as both the *Pina* and *Pinb* genes are present in all the diploid wheat species without exception, including *T. urartu* (the A genome progenitor) and Aegilops species from the Sitopsis section (described as the species most closely related to the B genome progenitor) (Gautier et al., 2000; Morris, 2002) (Figure 1). Polyploidy has been shown to induce DNA sequence elimination in the first generations (first to third) of synthetic wheat polyploids (Ozkan et al., 2001; Shaked et al., 2001), but nothing is yet known about the mechanisms of the observed genomic rearrangements. By comparison, the link between allopolyploidization and the evolutionary events that occurred and shaped the *Ha* locus in the A and B genomes of allopolyploid wheat species cannot be clearly established. However, our study allows identification of the extent and sequences of deleted genomic regions at this locus and suggests a mechanism by which they occurred. Detailed analysis of the remaining sequences at the *Ha* locus region in the A and B genomes of polyploid wheat revealed different truncated retroelements. As a consequence, *Pina* and *Pinb* gene deletions may have been driven by one or more interelement as well as intraelement illegitimate recombination events (Devos et al., 2002; Ma et al., 2004). We propose that these illegitimate recombination events occurred independently in the A and B genomes because they involved different retroelements and unassigned DNA sequences. Such events appear to be frequent at the *Ha* locus of wheat species because one additional deletion event in the A genome of *T. turgidum* ssp *durum* (Figure 2A, Del1Att) and other deletion events in the D genome of *T. aestivum* (Figure 4) were observed.

Since the emergence of allohexaploid bread wheat and its spreading to a more northern cultivation area, the *Ha* locus (of the D genome) has been subjected to selection pressure leading to hard wheat cultivars as a result of mutations or genomic rearrangements (such as *Pina* gene loss) (Giroux and Morris, 1998; Giroux et al., 2000; Morris, 2002). Furthermore, our study revealed that additional genomic rearrangements by illegitimate recombination, including two deletion events (different than the *Pina* and *Pinb* genes deletion), had occurred in the *Ha* locus of the D genome of hexaploid wheat as compared with that of the diploid *Ae. tauschii*. The genomic rearrangements in the D genome of hexaploid wheat could not be unequivocally attributed to polyploidization, domestication, or divergence between the *Ae. tauschii* accession sequenced here and the real donor(s) of the D genome. All three evolutionary forces most likely shaped the *Ha* locus.

Molecular Mechanism of the Evolutionary Events at the *Ha* Locus

Based on analysis of truncated TEs in Arabidopsis and rice genomes, DNA elimination by illegitimate recombination has been suggested as a major evolutionary force of plant genomes that counteract genome expansion by TE insertion or genomic duplication (Devos et al., 2002; Ma et al., 2004). However, it has not been possible to determine the extent and sequence composition of the rearranged genomic regions or whether illegitimate recombination involved sequences other than TEs. To elucidate these mechanisms, our strategy was based on (1) comparison between rearranged and nonrearranged genomes with different polyploidy levels, (2) identification of rearrangement breakpoints and specific sequence motifs, and (3) comparison with similar mechanisms previously proposed for eukaryotic as well as prokaryotic organisms (Bzymek et al., 1999; Demura et al., 2002; Molinier et al., 2004; Tufarelli et al., 2004). In contrast with homologous recombination, which requires pairing of two copies of genes or long repeats, illegitimate recombination events require smaller sequence motifs and occur in any genomic region (Kirik et al., 2000; Gregory, 2004). The ubiquitous identification of short conserved sequence motifs at rearrangement breakpoints observed in this study suggests illegitimate recombination, although different mechanisms leading to various rearrangement events (deletions, duplications, and inversions) are implicated (Figure 5). Thus, it was found that such events could take place by errors in DNA replication, such as replication slippage (Figure 5, event 5) (Demura et al., 2002), intermolecular and intramolecular unequal crossover (Figure 5, events 4 and 6) (Bzymek et al., 1999; Demura et al., 2002; Molinier et al., 2004; Tufarelli et al., 2004), or other unknown mechanisms (Figure 5, event 3). Illegitimate recombination has also been shown to be implicated in DNA double-stand repair (Gorbunova and Levy, 1999; Kirik, 2000). Contrary to previous observations (Devos et al., 2002; Ma et al., 2004), the illegitimate recombination events proposed in this article did not always occur in TE sequences (Figure 5). The two genomic deletion events in the *Ha* locus of the D genome of *T. aestivum* occurred in DNA sequences classified as unassigned (Figure 5). Thus, our results show that the rejoining of DNA fragments separated by several kilobases, and carrying different types of sequence classes, may occur by illegitimate recombination that is driven by short sequence motifs and could explain the diverse distribution, size, and sequence composition of the eliminated genomic DNA fragments (Figure 5).

METHODS

BAC Clone Isolation

The BAC clone from *Aegilops tauschii* (genome D) carrying the *Gsp-1*, *Pina*, and *Pinb* genes was isolated previously (Turnbull et al., 2003) and kindly provided by S. Rahman. The *Triticum monococcum* (genome A) BAC clone carrying the *Ha* locus was sequenced previously (GenBank accession number AY491681; Chantret et al., 2004) and has been incorporated in our comparative analysis. BAC clones containing the orthologous *Ha* locus region in the A and B genomes of durum wheat and the A, B, and D genomes of hexaploid wheat were obtained from *T. turgidum* ssp *durum* cv Langdon65 (Cenci et al., 2003) and *T. aestivum* cv Renan BAC libraries by screening with PCR primers or probes specific to the *Gsp-1*, *Pina*, and/or *Pinb* genes (Gautier et al., 1994). Assignment to the A, B, or D genomes of the BAC clones from these two polyploid species was based on their further characterization by restriction fragment length polymorphisms and fingerprinting on agarose gels. To ensure maximum coverage of the *Ha* locus, the longest BAC clones were sequenced.

BAC Sequencing and Sequence Assembly

BAC shotgun sequencing was performed at the Centre National de Séquençage (Evry, France). Briefly, BAC DNA was extracted by partial alkaline bacterial lysis and purified by two consecutive centrifugations in a caesium chloride gradient. Approximately 40 μg of sheared BAC DNA (Genemachines hydroshear; San Carlos, CA) resulted in fragments that were separated on preparative LPM agarose gels (FMC, Rockland, ME). DNA fragments of 5 kb were eluted with β-agarase (Biolabs, Hitchin, UK) and ligated to the *Bst*XI-digested pcDNA2.1 vector (Invitrogen, Carlsbad, CA). All sequence reads of subclone ends were performed using dye primer sequencing on ABI3730 machines (PE-Applied Biosystems, Foster City, CA). A target of 10-fold coverage was chosen. Sequences were assembled into contigs using PHRED (Ewing and Green, 1998) and PHRAP (P. Green, unpublished data) software. Gaps between the contigs were filled and sequenced by primer walking and transposition reactions (Finnzymes, Espoo, Finland). Poor-quality sequences were determined using specific primers and dye terminator sequencing on an automated ABI3730 sequencer (PE-Applied Biosystems). Consensus sequences were considered as valid when at least 98.0% of the nucleotides were base-called with a PHRAP score above 40. The consensus sequence was obtained after analysis of at least three sequence reads (on both strands) or using sequencing methods based on two different labeling procedures applied on one strand. The final assembling as well as individual single nucleotide polymorphisms between BAC clones were systematically PCR checked.

Sequence Annotation System

Genes and repeated elements (TEs and short repeats) were identified by computing and integrating results based on BLAST algorithms (Altschul et al., 1990, 1997), predictor programs, and different software detailed as follows.

Gene Structure Analysis

Gene structures and putative functions on the seven BAC clones analyzed were identified through BLASTN and BLASTX alignments against dbEST (http://www.ncbi.nlm.nih.gov/) and SwissProt databases, as well as using the gene predictor program GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi; Lukashin and Borodovsky, 1998) with three training versions (i.e., rice [*Oryza sativa*], wheat, and *Arabidopsis thaliana*). To incorporate heterologous information, we only recovered ESTs showing a minimum of 70% identity on at least 70% of the EST length. No cutoff was imposed on the BLASTX alignment against Swissprot to identify all potential gene encoding regions.

The CDS structure consists of a consensus that we derived from the three information sources previously described. The gene content parameter represents the sum of known genes, hypothetical genes, unknown genes, and pseudogenes. Known genes were named based on BLASTX results against proteins with known functions (SwissProt). CDSs were considered as hypothetical genes if their identification was only based on the GeneMark.hmm predictor (as a consensus of the structures suggested by the rice, wheat, and Arabidopsis trained versions), without any evidence of putative function based on BLASTX results; as unknown genes if the identification was only based on matching ESTs, without any evidence of putative function based on BLASTX results; and finally as pseudogenes if frame shifts need to be introduced within the CDS structure to better fit a putative function based on BLASTX results. Truncated pseudogenes, corresponding to genes disrupted by large insertion or deletion events as highly degenerated CDS sequences, were considered gene relics.

TEs

TEs were detected by comparison with two databases of repetitive elements. Core domains (nucleic coordinates of known elements) were identified through BLASTN alignments against TREP (nucleic nonredundant database, TREPnr), a public annotated version of repetitive elements from Triticeae species, mainly wheat and barley (*Hordeum vulgare*) (http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml) (Wicker et al., 2002). LTRs and limits were identified through BLASTN alignments against Repbase (http://www.girinst.org/Repbase_Update.html) (Jurka, 2000). Finally, putative and hypothetical polyproteins were identified by BLASTX alignments against TREP (hypothetical protein database, TREPprot). We used $1e^{-4}$ as a cutoff for BLASTN alignment results (either on TREPnuc or Repbase). No cutoff was imposed for BLASTX results on TREPprot. Nested insertions of TEs were considered only when complete reconstruction of the split element was possible with no ambiguity. Other TE structures (either novel or highly degenerate TEs) were identified within the remaining unassigned DNA either by LTR_STRUC or BLASTX against the nonredundant database (National Center for Biotechnology Information). When it was possible (complete elements), target-site duplications were indicated in the commentary of the element. TEs were subgrouped as class I, including LTR retrotransposons (such as copia and gypsy elements or *LARDs* and *TRIMs*) and non-LTR retrotransposons (such as *SINEs* and *LINEs*), class II, including DNA transposons (*CACTA*-like or not) and *MITEs*, and unclassified elements within TREP or Repbase databases. Unassigned DNA corresponds to sequences in which neither CDS nor TE was identified. Such unassigned DNA may contain short repetitive units (tandem repeats or inverted repeats) (Figure 2B).

Short repeated motifs were identified either as inverted repeats (EINVERTED; http://bioweb.pasteur.fr/seqanal/interfaces/einverted.html) or tandem repeats (Tandem Repeat Finder; http://tandem.bu.edu/trf/trf.advanced.submit.html), with default parameters. Only repeated intervals longer than 100 bp were kept in our annotation results.

Pairwise comparisons of the seven BAC clones, including the analysis of each BAC sequence against itself, were performed using the program *Dotter* (Sonnhammer and Durbin, 1995) to identify or confirm direct repeats, LTRs, local duplications, and deletion events as well as *MITEs*. Multiple sequence comparisons were performed with PIPMAKER software (Schwartz et al., 2000).

As a final screening, unassigned DNA (free of annotated genes or TEs) was aligned using BLASTX against the National Center for Biotechnology Information nonredundant database (including translation of rice and Arabidopsis annotated CDS). This BLASTX analysis allows the extension of several TE features already identified.

Cross-analysis of the information obtained for genes and TEs as short repeats was integrated into ARTEMIS (Rutherford et al., 2000) using PERL scripts and is available in Supplemental Data 2 online. Genes are named with the initial letters when functions are known or just numbered. TEs are named with the classical nomenclature (i.e., element name, BAC name, appearance rank) and also designated as complete, truncated, and degenerated sequences as suggested in the TREP database.

Sequence data from this article have been deposited with the EMBL/GenBank data libraries under the following accessions numbers: *T. aestivum* genome A, CR626929; *T. aestivum* genome B, CR626930; *T. aestivum* genome D, CR626934; *T. turgidum* genome A, CR626933; *T. turgidum* genome B, CR626932; *Ae. tauschii*, CR626926.

ACKNOWLEDGMENTS

We sincerely thank Heather McKhann and Ian Small (Institut National de la Recherche Agronomique-Unité de Recherches en Génomique Végétale) for valuable comments and discussions, Ivans Lagudah for supplying the *Ae. taushii* BAC library, Jorge Dubcovsky for the *T. durum* BAC library, the Genoplante consortium (http://www.genoplante.com) for the *T. aestivum* library, Lorenzo Cerutti for his help in the development of the annotation platform at Institut National de la Recherche Agronomique (Clermont-Ferrand, France), and Frank Samson for BAC annotation visualization-tool development. Sequencing of the BAC clone from *Ae. tauschii* was supported by the Genoplante consortium. Sequencing of the other BAC clones was supported by research grants from the French Ministry of Research and Agriculture.

Received November 9, 2004; accepted January 10, 2005.

REFERENCES

Allaby, R.G., Banerjee, M., and Brown, T.A. (1999). Evolution of the high molecular weight glutenin loci of the A, B, D, and G genomes of wheat. Genome **42,** 296–307.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215,** 403–410.

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

Blake, N.K., Lehfeldt, B.R., Lavin, M., and Talbert, L.E. (1999). Phylogenetic reconstruction based on low copy DNA sequence data in an allopoyploid: The B genome of wheat. Genome **42,** 351–360.

Bennetzen, J.L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica **115,** 29–36.

Bzymek, M., Saveson, C.J., Feschenko, V.V., and Lovett, S.T. (1999). Slipped misalignment mechanisms of deletion formation: In vivo susceptibility to nucleases. J. Bacteriol. **181,** 477–482.

Caldwell, K.S., Langridge, P., and Powell, W. (2004). Comparative sequence analysis of the region harboring the hardness locus in barley and its colinear region in rice. Plant Physiol. **136,** 3177–3190.

Cenci, A., Chantret, N., Kong, X., Gu, Y., Anderson, O.D., Fahima, T., Distelfeld, A., and Dubcovsky, J. (2003). Construction and characterization of a half million clone BAC library of durum wheat (*Triticum turgidum* ssp. *durum*). Theor. Appl. Genet. **107,** 931–939.

Chantret, N., Cenci, A., Sabot, F., Anderson, O., and Dubcovsky, J. (2004). Sequencing of the *Triticum monococcum Hardness* locus reveals good microcolinearity with rice. Mol. Genet. Genomics **271,** 377–386.

Clark, R.M., Linton, E., Messing, J., and Doebley, J.F. (2004). Pattern of diversity in the genomic region near the maize domestication gene tb1. Proc. Natl. Acad. Sci. USA **101,** 700–707.

Darlingto, H.F., Rouster, J., Hoffmann, L., Halford, N.G., Shewry, P.R., and Simpson, D.J. (2001). Identification and molecular characterisation of hordoindolines from barley grain. Plant Mol. Biol. **47,** 785–794.

Demura, M., Takeda, Y., Yoneda, T., Furukawa, K., Usukura, M., Itoh, Y., and Mabuchi, H. (2002). Two novel types of contiguous gene deletion of the AVPR2 and ARHGAP4 genes in unrelated Japanese kindreds with nephrogenic diabetes insipidus. Hum. Mutat. **19,** 23–29.

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

Eckardt, N.A. (2001). A sense of self: The role of DNA sequence elimination in allopolyploidization. Plant Cell **13,** 1699–1704.

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

Feldman, M., Lupton, F.G.H., and Miller, T.E. (1995). Wheats. In Evolution of Crops, 2nd ed., J. Smartt and N.W. Simmonds, eds (London: Longman Scientific), pp. 184–192.

Gautier, M.F., Aleman, M.E., Guirao, A., Marion, D., and Joudrier, P. (1994). *Triticum aestivum* puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant Mol. Biol. **25,** 43–57.

Gautier, M.F., Cosson, P., Guirao, A., Alary, R., and Joudrier, P. (2000). Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid *Triticum* species. Plant Sci. **153,** 81–91.

Giroux, M.J., and Morris, C.F. (1998). Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. Proc. Natl. Acad. Sci. USA **95,** 6262–6266.

Giroux, M.J., Talbert, L., Habernicht, D.K., Lanning, S., Hemphill, A., and Martin, J.M. (2000). Association of *puroindoline* sequence type and grain hardness in hard red spring wheat. Crop Sci. **40,** 370–374.

Gorbunova, V.V., and Levy, A.A. (1999). How plants make ends meet: DNA double-strand break repair. Trends Plant Sci. **4,** 263–269.

Gregory, T.R. (2004). Insertion-deletion biases and the evolution of genome size. Gene **324,** 15–34.

Hogg, A.C., Sripo, T., Beecher, B., Martin, J.M., and Giroux, M.J. (2004). Wheat puroindolines interact to form friabilin and control wheat grain hardness. Theor. Appl. Genet. **108,** 1089–1097.

Huang, S., Sirikhachornkit, A., Su, X.J., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. Proc. Natl. Acad. Sci. USA **99,** 8133–8138.

Jurka, J. (2000). Repbase update: A database and an electronic journal of repetitive elements. Trends Genet. **9,** 418–420.

Kashkush, K., Feldman, M., and Levy, A.A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics **160,** 1651–1659.

Kashkush, K., Feldman, M., and Levy, A.A. (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. **33,** 102–106.

Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature **428,** 617–624.

Kirik, A., Salomon, S., and Puchta, H. (2000). Species-specific double-strand break repair and genome evolution in plants. EMBO J. **19,** 5562–5566.

Law, C.N., Young, C.F., Brown, J.W.S., Snape, J.W., and Worland, A.J. (1978). The study of grain protein control in wheat using whole chromosomes substitution lines. In Seed Protein Improvement by Nuclear Techniques, I.A.E. Agency, ed (Vienna, Austria: I.A.E. Agency), pp. 483–502.

Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res. **26,** 1107–1115.

Ma, J., Devos, K.M., and Bennetzen, J.L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res. **14,** 860–869.

Mayer, V.W., and Aguilera, A. (1990). High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. Mutat. Res. **231,** 177–186.

McCarthy, E., and McDonald, J. (2003). LTR_STRUC: A novel search and identification program for LTR retrotransposons. Bioinformatics **19,** 362–367.

Molinier, J., Ries, G., Bonhoeffer, S., and Hohn, B. (2004). Onterchromatid and interhomolog recombination in *Arabidopsis thaliana*. Plant Cell **16,** 342–352.

Morris, C.F. (2002). Puroindolines: The molecular genetic basis of wheat grain hardness. Plant Mol. Biol. **48,** 633–647.

Nesbitt, M., and Samuel, D. (1995). From staple crop to extinction? The archeology and history of the hulled wheats. In Hulled Wheats. Proceedings of the First International Workshop on Hulled Wheats, 1st ed., S. Padulosi, K. Hammer, and J. Heller, eds (Rome: International Plant Genetic Resources Institute), pp. 40–99.

Ohno, S. (1970). Evolution by Gene Duplication. (London: George Allen and Unwin).

Ozkan, H., and Feldman, M. (2001). Genotypic variation in tetraploid wheat affecting homoeologous pairing in hybrids with *Aegilops peregrina*. Genome **44,** 1000–1006.

Ozkan, H., Levy, A.A., and Feldman, M. (2001). Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. Plant Cell **13,** 1735–1747.

Petrov, D.A. (2001). Evolution of genome size: New approaches to an old problem. Trends Genet. **17,** 23–28.

Pikaard, C.S. (2001). Genomic change and gene silencing in polyploids. Trends Genet. **17,** 675–677.

Rahman, S., Jolly, J.C., Skerritt, J.H., and Wallosheck, A. (1994). Cloning of a wheat 15-kDa grain softness protein (GSP). GSP is a mixture of puroindoline-like polypeptides. Eur. J. Biochem. **223,** 917–925.

Ramakrishna, W., Emberton, J., Ogden, M., SanMiguel, P., and Bennetzen, J.L. (2002). Structural analysis of the maize rp1 complex reveals numerous sites and unexpected mechanisms of local rearrangement. Plant Cell **14,** 3213–3223.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: Sequence visualization and annotation. Bioinformatics **16,** 944–945.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker—A web server for aligning two genomic DNA sequences. Genome Res. **10,** 577–586.

Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, A.A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. Plant Cell **13,** 1749–1759.

Sonnhammer, E.L., and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene **167,** GC1–10.

Sourdille, P., Perretant, M.R., Charmet, G., Leroy, P., Gautier, M.F., Joudrier, P., Nelson, J.C., Sorrells, M.E., and Bernard, M. (1996). Linkage between RFLP markers and genes affecting kernel hardness in wheat. Theor. Appl. Genet. **93,** 580–586.

Tanksley, S.D., and McCouch, S.R. (1997). Seed banks and molecular maps: Unlocking genetic potential from the wild. Science **277,** 1063–1066.

Tranquilli, G., Lijavetzky, D., Muzzi, G., and Dubcovsky, J. (1999). Genetic and physical characterization of grain texture-related loci in diploid wheat. Mol. Gen. Genet. **262,** 846–850.

Tufarelli, C., Hardison, R., Miller, W., Hughes, J., Clark, K., Ventress, N., Frischauf, A.M., and Higgs, D.R. (2004). Comparative analysis of the alpha-like globin clusters in mouse, rat, and human chromosomes indicates a mechanism underlying breaks in conserved synteny. Genome Res. **14,** 623–630.

Turnbull, K.M., Turner, M., Mukai, Y., Yamamoto, M., Morell, M.K., Appels, R., and Rahman, S. (2003). The organization of genes tightly linked to the Ha locus in *Aegilops tauschii*, the D-genome donor to wheat. Genome **46,** 330–338.

Wendel, J.F., and Wessler, S.R. (2000). Retrotransposon-mediated genome evolution on a local ecological scale. Proc. Natl. Acad. Sci. USA **97,** 6250–6252.

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

Wicker, T., Yahiaoui, N., Guyot, R., Schlagenhauf, E., Liu, Z.D., Dubcovsky, J., and Keller, B. (2003). Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A_m genomes of wheat. Plant Cell **15,** 1186–1197.