Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels

Edwige Isidore, Beatrice Scherrer, Boulos Chalhoub, et al.

Genome Res. 2005 15: 526-536
Access the most recent version at doi:10.1101/gr.3131005

Letter

Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels

Edwige Isidore,[1,4] Beatrice Scherrer,[1,4] Boulos Chalhoub,[2] Catherine Feuillet,[1,3] and Beat Keller[1,5]

[1]*Institute of Plant Biology, University of Zürich, 8008 Zürich, Switzerland;* [2]*Institut National de la Recherche Agronomique (INRA), Unité de Génomique Végétale, 91057 Evry, France*

Plant genomes, in particular grass genomes, evolve very rapidly. The closely related A genomes of diploid, tetraploid, and hexaploid wheat are derived from a common ancestor that lived <3 million years ago and represent a good model to study molecular mechanisms involved in such rapid evolution. We have sequenced and compared physical contigs at the *Lr10* locus on chromosome 1AS from diploid (211 kb), tetraploid (187 kb), and hexaploid wheat (154 kb). A maximum of 33% of the sequences were conserved between two species. The sequences from diploid and tetraploid wheat shared all of the genes, including *Lr10* and *RGA2* and define a first haplotype (H1). The 130-kb intergenic region between *Lr10* and *RGA2* was conserved in size despite its activity as a hot spot for transposon insertion, which resulted in >70% of sequence divergence. The hexaploid wheat sequence lacks both *Lr10* and *RGA2* genes and defines a second haplotype, H2, which originated from ancient and extensive rearrangements. These rearrangements included insertions of retroelements and transposons deletions, as well as unequal recombination within elements. Gene disruption in haplotype H2 was caused by a deletion and subsequent large inversion. Gene conservation between H1 haplotypes, as well as conservation of rearrangements at the origin of the H2 haplotype at three different ploidy levels indicate that the two haplotypes are ancient and had a stable gene content during evolution, whereas the intergenic regions evolved rapidly. Polyploidization during wheat evolution had no detectable consequences on the structure and evolution of the two haplotypes.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. AY663391 and AY663392 for the BACs 1156G16 (*T. turgidum* subsp. *durum*, cv. Langdon) and 930H14 (*T. aestivum* cv. Renan), respectively.]

Comparative analysis of related genomes can reveal the molecular mechanisms of genome evolution. Species from the grass family provide an excellent model for such studies, as extensive genetic colinearity among several grass species has been described despite very heterogenous genome sizes and evolutionary divergence times of over 60 million years (for review, see Devos and Gale 2000; Keller and Feuillet 2000). Recent large-scale comparative studies in maize, sorghum, rice, barley, and wheat have revealed a mosaic organization of conserved and nonconserved genes at orthologous loci and many small rearrangements (Dubcovsky et al. 2001; Ramakrishna et al. 2002; Song et al. 2002; Brunner et al. 2003; Gu et al. 2003, 2004; Ilic et al. 2003; Guyot et al. 2004). Interestingly, intraspecific comparative studies in maize have also shown disruption of colinearity, not only in the intergenic regions, but also in the gene space (Fu and Dooner 2002; Song and Messing 2003). The partial absence of microcolinearity observed between grass species, as well as between inbred maize cultivars, is a rich resource for the identification of molecular mechanisms involved in genome evolution.

The wheat A genomes that diverged from a common ancestor living about 0.5–3 million years ago (Mya) (Huang et al. 2002; Wicker et al. 2003a) are particularly suitable for comparative analysis. The A genomes are found in modern wheat species of different ploidy such as Einkorn wheat (*Triticum monococcum*, diploid), Emmer wheat (*T. turgidum*, tetraploid), and bread wheat species (*T. aestivum*, hexaploid) (Feldman 2001). A number of recent studies have demonstrated rapid and massive local changes in wheat genomes after polyploidization (Liu et al. 1998; Ozkan et al. 2001). Therefore, the comparison at the molecular level of wheat A genomes at different ploidy levels could also give insight into the molecular consequences of the "genomic shock" following hybridization of complete genomes. The development of large insert bacterial artificial chromosome (BAC) libraries from diploid wheat *T. monococcum* DV92 (AA, Lijavetzky et al. 1999), tetraploid wheat *T. turgidum* subsp. *durum* cv. Langdon (AABB, Cenci et al. 2003), and the donor of the D genome *Aegilops tauschii* (Moullet et al. 1999) has allowed the isolation and sequencing of large wheat genomic fragments. Recent studies have compared orthologous loci in the three A, B, and D homoeologous wheat genomes (Gu et al. 2004; Kong et al. 2004),

[3]Present address: UMR INRA-UBP ASP Clermont-Ferrand, 63039 Clermont-Ferrand Cedex 2, France.
[4]These two authors contributed equally to this work.
[5]Corresponding author.
E-mail bkeller@botinst.unizh.ch; fax 41-1-634-8204.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3131005.

15:526–536 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

which diverged 2.5–4 Mya (Huang et al. 2002). Rapid genome evolution was observed that was mostly due to insertion of retroelements after the divergence of the three subgenomes. A comparison of the A genome of tetraploid (cv. Langdon with A^u genome from *T. urartu*) with the A^m genome of the diploid *T. monococcum* DV92 that diverged 0.5–3 Mya (Huang et al. 2002; Wicker et al. 2003a), has recently revealed rapid genome divergence in the intergenic regions even between these closely related genomes (Wicker et al. 2003a). Conservation was restricted to small regions and a large proportion of the sequence, which contains mainly repetitive elements, was completely different, providing evidence for a dynamic and rapid evolution. So far, none of the previous comparative studies have included orthologous sequences from hexaploid wheat, and the evolution of A genomes derived from *T. urartu* in different polyploid backgrounds has not been investigated. Two BAC libraries from hexaploid wheat have been recently constructed from the cultivars Chinese Spring (Allouis et al. 2003) and Renan (B. Chalhoub, unpubl.), allowing such analysis.

The leaf rust resistance gene *Lr10* was recently isolated (Feuillet et al. 2003) using a combination of subgenome map-based cloning (Stein et al. 2000; Wicker et al. 2001) and haplotype studies (Scherrer et al. 2002). *Lr10* is closely associated with a second resistance gene analog (*RGA2*), but the two genes are very different from each other at the molecular level (Feuillet et al. 2003). Analysis of the wheat gene pool by Southern hybridization revealed the presence of two groups of lines with distinct haplotypes (H1 and H2) at the *Lr10* locus, defined by the presence (H1) or absence (H2) of the two full-length *Lr10* and *RGA2* genes on chromosome 1A (Scherrer et al. 2002). Thus, the *Lr10* resistance locus differs from most resistance loci in plants, where clusters of closely related genes are present, but is similar to the *Arabidopsis RPM1* locus (Grant et al. 1995). The *RPM1* locus also exists in two stable haplotypes in the *Arabidopsis* gene pool that are defined by the presence or absence of the gene.

Studies of molecular haplotype diversity in plants are limited and so far restricted to a few species, such as *Arabidopsis*, soybean, barley, and maize (Collins et al. 2001; Nordborg et al. 2002; Charlesworth et al. 2003; Zhu et al. 2003; for review, see Rafalski and Morgante 2004). To assess the molecular basis and mechanisms of genomic rearrangements during evolution of the *Lr10* locus in the two different haplotypes, we have compared BAC sequences at the orthologous *Lr10* loci of the A genomes from diploid, tetraploid, and hexaploid wheat species. These comparisons revealed that the two haplotypes were very stable during evolution from diploid to tetraploid and hexaploid wheat, although conservation was restricted to the gene space, whereas intergenic sequences diverged rapidly. A large deletion and inversion was found at the origin of the H2 haplotype that, despite gene loss, was highly conserved during wheat evolution.

Results

Contig establishment at the *Lr10* locus in tetraploid and hexaploid wheat

We have isolated BAC clones from the orthologous *Lr10* loci on chromosome 1AS from *T. turgidum* subsp. *durum* cv. Langdon and *T. aestivum* cv. Renan, which belong to the H1 and H2 haplotypes, respectively (Fig. 1). Seven BAC clones were identified from the tetraploid Langdon BAC library (data not shown) using the probes *rga1NBS* and *rga2NBS2,* which are derived from the *Lr10* and *RGA2* genes, respectively (Scherrer et al. 2002) and are specific for the A genome. Previous hybridization studies have shown that H2 lines (including Renan) have no *Lr10* gene, only part of *RGA2* on chromosome 1A and a full copy of *RGA2* on chromosome 1D (Scherrer et al. 2002). No specific primers could be designed for the partial fragment of *RGA2* on chromosome 1A. Therefore, the hexaploid Renan BAC library was screened by PCR of pooled BAC DNA with two primer pairs, one amplifying specifically the *RGA2* copy on the D genome (DF/DR) and the other one amplifying both *RGA2* copies on the A and D genomes (ADF/ADR). In this way, we could identify three BACs from the A genome by screening for amplification with the primers ADF/ADR but not with DF/DR (data not shown). Based on DNA hybridizations with probes derived from the *T. monococcum* DV92 sequence (Wicker et al. 2001), physical contigs were established for the tetraploid and hexaploid wheat species. Hybridization experiments of NotI fingerprints of both tetraploid and hexaploid BAC clones revealed that the order of the probes was conserved in both wheat species (Fig. 1). However, the *Lr10* and *RGA2* probes corresponding to the 5′ end of both genes were not detected in the hexaploid Renan (H2 haplotype). The probe *F467* derived from low-pass sequencing of the *T. monococcum* DV92 BAC clone 111I4 (Stein et al. 2000) was found on two NotI frag-

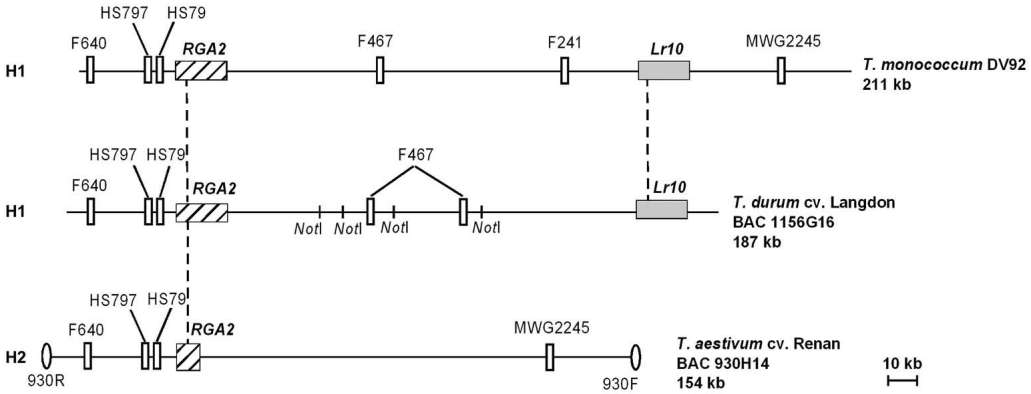

Figure 1. Physical maps of BAC clones from the *Lr10* orthologous loci in diploid, tetraploid, and hexaploid wheat. All probes indicated are derived from the *T. monococcum* DV92 sequence (Wicker et al. 2001) except for 930F and 930R, which are PCR-amplified BAC end probes of the hexaploid BAC clone and were not found in any of the other sequences. The NotI restriction sites are also indicated.

ments on the *T. turgidum* BAC, suggesting a duplication in cv. Langdon. In addition, the distance between *RGA2* and MWG2245 (Stein et al. 2000) was shorter in the hexaploid H2 haplotype, indicating a large deletion, including the *Lr10* gene in Renan. Based on these results, one BAC clone from cv. Langdon (BAC 1156G16) and one BAC clone from cv. Renan (BAC 930H14) were selected for complete sequencing (Fig. 1). A total of 187,054 bp of sequence was assembled with a 9.18-fold coverage for the Langdon 1156G16 BAC clone. For the cv. Renan BAC 930H14, a total of 154,778 bp of sequence was assembled with a similar coverage.

The diploid and tetraploid wheat lines studied here have the same haplotype structure (H1) at the *Lr10* locus, as they each comprise full-length *RGA2* and *Lr10* genes (Scherrer et al. 2002; Fig. 2A) as well as the three genes *ACT*, *CCF*, and *CCF*$^{(p)}$ (Fig. 2). All of the genes are conserved in order and orientation. The size of the large intergenic region between *Lr10* and *RGA2* is nearly identical (around 130 kb) in both sequences (Fig. 2A). In the haplotype H2 of the hexaploid cv. Renan, the genes *ACT*, *CCF*, *CCF*$^{(p)}$, *A5HY*, and *NLL2* (Fig. 2B) are conserved in the same order and orientation as in the H1 haplotype. However, major differences are observed in the *RGA2/NLL1* interval. First, in the hexa-

Sequence organization at the *Lr10* locus in tetraploid and hexaploid wheat

The tetraploid and hexaploid BAC sequences were annotated based on comparison with the diploid *T. monococcum* DV92 sequence analyzed by Wicker et al. (2001). The 187,054-bp sequence of tetraploid wheat is comprised of 8% genic regions and 55.3% identifiable, repetitive elements. Class I LTR retrotransposons are the most abundant elements, and *Angela* LTR retrotransposons represent >48% of all repetitive sequences. The 154,778-bp sequence of hexaploid wheat is comprised of 11.5% genic regions and 33.3% repetitive elements. *Angela* LTR retrotransposons represent >50% of all repeats and are mainly located in a single region of ~25 kb between the 3′ end of *RGA2* (*RGA2-b*) and *NLL1* (Fig. 2). The only class II elements identified in the three orthologous sequences are CACTA transposons (Wicker et al. 2003b). Five genes (*ACT*, *CCF*, *CCF*$^{(p)}$, *RGA2*, and *Lr10*), of which one is a putative gene (*CCF*$^{(p)}$), were found in the Langdon BAC 1156G16 sequence (Fig. 2). Orthologs of all five genes have been previously identified in *T. monococcum* DV92 (Wicker et al. 2001; Guyot et al. 2004; Fig. 2). The *ACT*, *CCF*, and *CCF*$^{(p)}$ genes were also present in the orthologous Renan BAC 930H14 sequence. In this sequence, three additional genes were detected (*NLL1*, *A5HY*, and *NLL2*), for which orthologs also exist in the orthologous region of *T. monococcum* DV92. The *NLL1* gene was described by Wicker et al. (2001), whereas *NLL2* was recently identified by Guyot et al. (2004). The *A5HY* gene, which was identified in *T. monococcum* by low-pass sequencing (data not shown), encodes a conserved domain of cytochrome P450s, proteins usually involved in oxidative degradation of various compounds. The best BLASTP hit for this protein is an aldehyde-5-hydroxylase (e value of e-159).

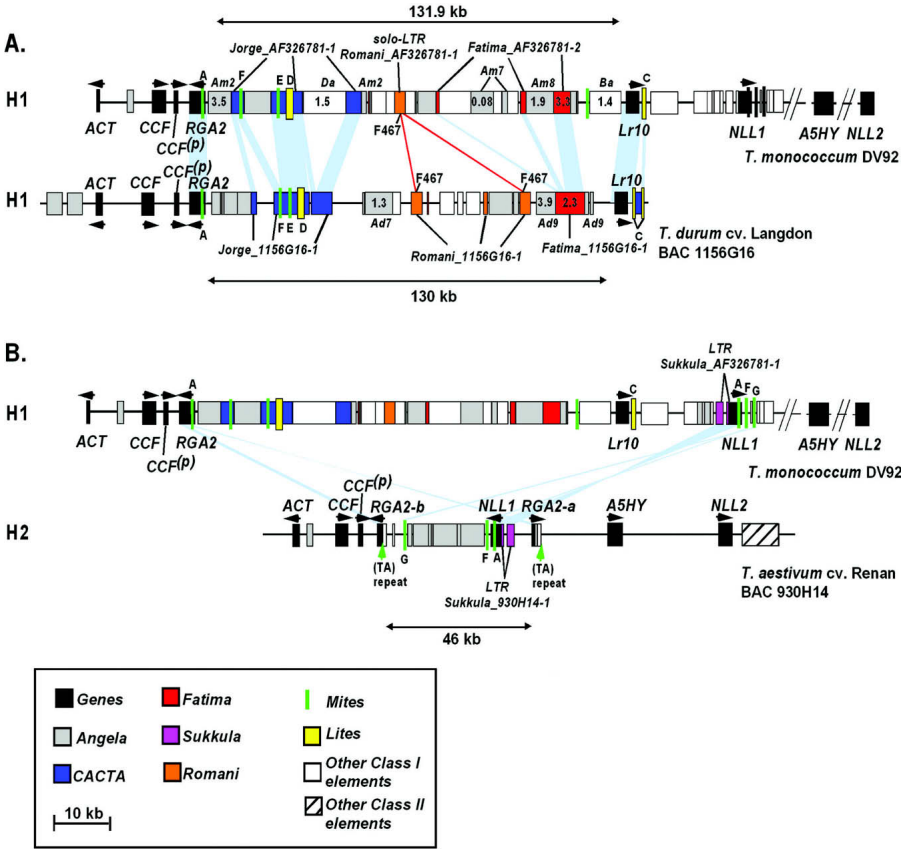

Figure 2. Schematic representation of the sequence organization and detailed comparison at the *Lr10* orthologous loci in diploid, tetraploid, and hexaploid wheat. (*A.*) Organization of the orthologous loci in *T. monococcum* and *T. turgidum* and comparison of the genomic region between *RGA2* and *Lr10* (haplotype H1). (*B.*) Organization of the orthologous loci in *T. monococcum* and *T. aestivum* and comparison between *T. monococcum* (H1) and *T. aestivum* (H2) in a region delimited by the two *RGA2* fragments of the *T. aestivum* sequence. Light-blue areas indicate conservation between the orthologous regions in the *RGA2–Lr10* interval. The genes are indicated by black boxes. The short arrows located *above* or *below* the genes indicate the transcriptional orientation. Different repetitive elements are displayed with different colors. Full names of conserved retroelements are given according to the complete annotation of the BAC sequences AY663391 and AY663392. *Am2*, *Am7*, *Am8*, *Ad7*, and *Ad9* correspond to *Angela* retroelements *Angela*_AF326781-2, *Angela*_AF326781-7, *Angela*_AF326781-8, *Angela*_1156G16-7, *Angela*_1156G16-9, respectively. *Da* and *Ba* correspond to *Daniela*_AF326781-7, *Barbara*_AF326781-7, respectively. The foldback elements (MITES and LITES) are only given initials in bold type font. (A) Athos; (F) Fortuna; (E) Eos; (D) Deimos; (C) Charon. F467 is a RFLP probe derived from the diploid sequence (Wicker et al. 2001) corresponding to an LTR fragment of a *Romani* retroelement, and the red lines *between* the diploid and tetraploid sequences indicate that these LTRs are derived from a common ancestor element. Green arrows indicate the (TA) repeats. Numbers in the boxes indicate the estimated insertion dates of the respective retroelements in million years ago. Numbers *below* the sequences indicate intergenic distances in kilobases delimited by double-headed arrows. The sequence in the region delimited by dashed lines at the *right* end of the *T. monococcum* sequence was only low-pass sequenced, and the intergenic distances in this region were as estimated by Stein et al. (2000). H1 and H2 indicate the two haplotypes, H1 and H2, according to Scherrer et al. (2002).

ploid sequence (H2), 2.5 kb of *RGA2*, corresponding to the 5′ half of the gene, are missing. The remaining part (bases 2537–4769) is split into two parts, *RGA2-a* and *RGA2-b*, which are separated from one another by >46 kb (Fig. 2B). In addition, *RGA2-a* is inverted compared with the original gene sequence. The gene fragment *RGA2-a* comprises the last 73 bp of the second exon, the second intron, and the first 352 bp of the third and last exon of *RGA2* (bases 2538–3272 of the complete *RGA2* gene). The gene fragment *RGA2-b* comprises the remaining part of the last exon (bases 3273–4769 bp). The second major difference to the sequence of the H1 haplotype is the complete absence of *Lr10* in the hexaploid H2 sequence, confirming previous hybridization experiments (Scherrer et al. 2002). Finally, the two haplotypes differ in the position of the *NLL1* gene between the *RGA2* fragments, and the inverted orientation of *NLL1* and *RGA2-a* compared with the order observed in the H1 sequences. This suggests that the *T. aestivum* H2 haplotype originated from the H1 haplotype after extensive rearrangements including deletions, inversions, and insertions.

A gene-rich region is highly conserved in the three homoeologous A genomes

At the three orthologous loci, the first 31 kb of common sequence correspond to a gene-rich region with three genes (*ACT*, *CCF*, and *CCF*$^{(p)}$) (Fig. 3) and a partial gene (*RGA2*) that are conserved in order and orientation. The sizes of the intergenic regions are short, the largest being 14.6 kb between *ACT* and *CCF* in the *T. monococcu*m DV92 sequence (Fig. 3). Nucleic acid sequences of all the genes are highly conserved, as there is >89.8% identity between the orthologous genes, independent of haplotype and ploidy level (data not shown). In addition, sequences ranging from 5 bp to >6.8 kb are also conserved in the intergenic regions with >80% of identity at the nucleic acid level (Fig. 3). The conserved regions downstream of the genes are, in general, shorter (from 5 to 1.3 kb) than the upstream regions (from 28 to 6.8 kb). For both the *ACT* and *CCF* genes, the conserved 5′ upstream regions are longer in *T. turgidum* and *T. aestivum* (1.3 and 6.8 kb, Fig. 3) than in *T. turgidum*/*T. aestivum* and *T. monococcum* (<750 bp in both). In *T. aestivum* (haplotype H2), only the bases 3273–4769 of *RGA2* (*RGA2-b*), corresponding to the last 1496 bp of the last exon, are found at the same position as in *T. monococcum* and *T. turgidum* (haplotype H1). Nevertheless, the 1496 bp of sequence is highly conserved (>96% of sequence identity). In addition, 1.3 kb downstream of the stop codon of *RGA2* are conserved, in all three genomes, with >90% identity (Fig. 3), whereas 4.6 kb are conserved upstream of the start codon of *RGA2* only in the diploid and tetraploid species. A similar conservation of both upstream and downstream regions was found for *Lr10* (data not shown) in the H1 haplotype, and the gene sequence itself is even more highly conserved (>99% identity). Such conservation in the gene and regulatory sequences of both *Lr10* and *RGA2* indicates a strong selective pressure that maintains them intact and functional in the species of the H1 haplotype. To study whether selective pressure acts on the functional *RGA2* gene, and to assess whether this is also true for the partial *RGA2* sequence found in the hexaploid wheat cultivar, we have compared the last exon (exon 3) of *RGA2* (nucleotides 2920–4769) in *T. monococcum*, *T. turgidum*, and *T. aestivum*. Estimation of the synonymous (*Ks*) vs. nonsynonymous (*Ka*) substitution rates yielded a *Ka/Ks* ratio of 0.125 and revealed strong purifying selection on the H1 sequences with a functional allele (*T. monococcum* and *T. turgidum*). In contrast, the *Ka/Ks* ratio of 0.35 calculated for exon 3 in the *T. aestivum* (H2) sequence indicates relaxation of the selection on the partial gene.

The large intergenic regions between *RGA2* and *Lr10* evolved differentially in *T. monococcum* and *T. turgidum* of the HI haplotype

The intergenic regions between *RGA2* and *Lr10* in *T. monococcum* and *T. turgidum* are mainly comprised of nested repetitive elements (Fig. 2A). Repetitive elements represent 64% of the *T. turgidum* interval and 83% of the *T. monococcum* interval, which is higher than the content of repetitive elements in the complete sequences of both species (54% in *T. turgidum* and 70% in *T. monococcum*). This suggests preferential insertion of repetitive elements in this intergenic region. The largest complete elements common to both species are a *Jorge* CACTA transposon of 15.6 kb (*Jorge*_AF326781-1 and *Jorge*_1156G16-1) and a *Fatima* LTR retrotransposon of 9.1 kb (*Fatima*_AF326781-2 and *Fatima*_1156G16-1). These elements are true orthologs, as they are found in both species at corresponding positions, indicating that they were present already in the common ancestor. In addition, five foldback elements (three MITES and two LITES) (Fig. 2A) located in *RGA2*, the *Jorge* element and downstream of *Lr10*, respectively, are also conserved in both sequences. These data indicate that these elements inserted between *RGA2* and *Lr10* in the common ancestor of the H1 lines before the divergence of the A genome in the *T. monococcum* and *T. urartu* lineages. Both *Jorge* and *Fatima* elements are interrupted by other elements in diploid wheat, but only *Jorge* is interrupted in the tetraploid sequence (Fig. 2A). In total, 35.7 kb are conserved between the *RGA2* and *Lr10* genes at the two orthologous loci of the H1 haplotype structure. This represents only 27% of the two intergenic regions, despite their almost identical length (130 kb).

The presence of the conserved *Jorge* and *Fatima* elements both in *T. monococcum* DV92 and *T. turgidum* cv. Langdon allowed the comparison of base substitution rates in repetitive elements vs. genes in this region. The average substitution rate of the complete sequences of the two repetitive elements calculated as described in San Miguel et al. (1998) is 0.032. The synonymous substitution rates of the coding sequences of three

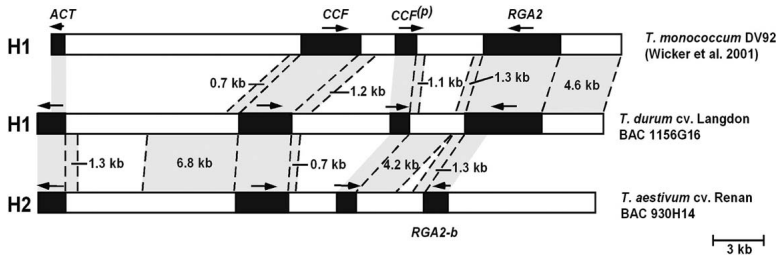

Figure 3. Schematic comparison of the 31-kb gene-rich region conserved on the proximal side of the *Lr10* locus in *T. monococcum*, *T. turgidum*, and *T. aestivum*. The four genes *ACT*, *CCF*, *CCF*$^{(p)}$, and *RGA2* located in this gene-rich region are represented by black boxes. The arrows located *above* the genes indicate the transcriptional orientation. Areas of similarity between the orthologous sequences are indicated in gray. The sizes of conserved intergenic regions are given in kilobases and are delimited by dashed lines. H1 and H2 indicate the haplotypes according to Scherrer et al. (2002).

complete genes (*CCF*, *Lr10*, and *RGA2*) are 0.054, 0.050, and 0.019, respectively. *RGA2* appears to have evolved more slowly than the repetitive elements. However, the data show that the other two genes appear to have evolved faster than the repetitive elements, and highlight the high variability of synonymous substitution rates in different genes. Recently, such a large variability in substitution rates was also found in a set of 24 genes in rice (Ma and Bennetzen 2004).

Insertion times were estimated for nine complete LTR retroelements in the *RGA2*/*Lr10* intergenic region in the *T. monococcum* and *T. turgidum* sequences using an average base substitution of 6.5×10^{-9} mutations per site per year (Gaut et al. 1996) and following the method used by SanMiguel et al. (1998). This revealed a "recent" group (0.08–1.5 Mya) and a more "ancient" group (2.3–3.9 Mya) of retrotransposon insertions. The old group is comprised of one element conserved in both sequences (*Fatima*_AF326781-2 and *Fatima*_1156G16-1) and two *Angela* elements (*Am2* and *Ad9*). Although *Am2* is only present in *T. monococcum* and *Ad9* only in *T. turgidum*, both are interrupted by conserved elements (*Jorge* and *Fatima*) (Fig. 2A). Therefore, these *Angela* elements were most likely already present in the common ancestor of the two species and subsequently deleted from one of them. Although *Fatima*_AF326781-2 and *Fatima*_1156G16-1 are true orthologs (i.e., they inserted in the common ancestor), their calculated insertion times are slightly different (3.3 and 2.3 Mya, respectively) (Fig. 2A). This can be explained by the fact that the LTRs of this element are short (490 bp), compared with an average LTR size of 1438 bp for *Angela* elements, and this resulted in a calculated substitution rate with a higher standard deviation (SD) (in both sequences, SD was at least 0.85 million years). The more recent group of retrotransposon insertions comprises five elements with insertion times ranging from 0.08 to 1.9 Mya, four of them in *T. monococcum* (*Da*, *Am7*, *Am8*, and *Ba*) and only one in *T. turgidum* (*Ad7*), with *Am7* being the most recent insertion (0.08 Mya). None of these elements are common to both sequences, indicating either that they inserted after the divergence of the wheat A genomes, or that they were present in a common ancestor and were subsequently deleted from one of the two species. The divergence of the wheat A genomes is estimated to have occurred 0.5 to 3 Mya ago (Huang et al. 2002; Wicker et al. 2003a). Genome divergence is thus in the range of the calculated insertion times of the four remaining "recent" elements (on average 1.5 Mya), but there may be differences in evolution rates between LTR sequences and other sequences (SanMiguel et al. 1998). We cannot, therefore, exclude the possibility that these elements inserted after the species divergence.

Only 27% of sequence in the *RGA2*/*Lr10* intergenic region is conserved between *T. monococcum* and *T. turgidum*. This is the result of independent rearrangements that occurred after the divergence of the wheat A genomes. Comparison of the conserved features identified several molecular mechanisms underlying these rearrangements. In addition to insertions and deletions of LTR retrotransposons, which have also been described in previous studies in wheat (Wicker et al. 2003a; Gu et al. 2004), the pattern of conservation of the *Romani* retrotransposons provides interesting insight into the evolution of this interval (Fig. 2A). Both *T. monococcum* and *T. turgidum* contain a *Romani* element at corresponding positions, indicating that the element was already present in the common ancestor. In *T. monococcum*, only a solo-LTR with identical TSDs (*Romani*_AF326781-1), the result of an intraelement recombination, is present, whereas the complete element is present in the tetraploid wheat sequence (*Romani*_1156G16-1) (Fig. 2A). A model for the evolution of this *Romani* element in the two sequences is proposed in Figure 4. Interestingly, *Romani* in *T. turgidum* is interrupted by other retroelements in its internal domain. The insertions of these four elements date back 1.8–2.8 Mya (dated according to SanMiguel et al. 1998). It is therefore very likely that at least some of them were already present in the common ancestor of *T. monococcum* and *T. turgidum*. Thus, the intraelement recombination in *T. monococcum* led to the loss of at least 7 kb of genomic DNA (if all elements inserted after the divergence of the two species), and at the most, 33 kb (if all elements were present in the common ancestor). The presence of the two LTRs in *T. turgidum* has actually been misinterpreted as a duplication of the *T. monococcum* sequence in *T. turgidum*, based on hybridization experiments of the tetraploid BAC with the probe *F467* (Fig. 1), which hybridized to two NotI fragments. This probe was derived from the internal sequence of the *Romani* solo-LTR of *T. monococcum* and detected both LTRs in the *T. turgidum* sequence.

The two haplotypes, H1 and H2, originate from ancient and extensive rearrangements

The presence of the conserved *Jorge* and *Fatima* repetitive elements both in *T. monococcum* and *T. turgidum* (a total of 24,797 bp) allowed us to estimate the time of divergence between the two species, because these conserved elements must have been identical at the time of divergence. Based on an average base substitution of 6.5×10^{-9} mutations per site per year (Gaut et al. 1996), the two loci diverged ~2.4 Mya. Similar values for the divergence of the two species were previously obtained by Wicker et al. (2003a) (2.9–3.3 Mya). Based on this divergence time and the synonymous substitution rate of *RGA2* between *T. monococcum* and *T. turgidum*, a molecular clock of 3.93×10^{-9} mutations per site per year was established for *RGA2*. Assuming a constant substitution rate in both species, an estimated date of 4 Mya was calculated for the partial deletion and disruption of *RGA2* in Renan. Thus, the disruption of *RGA2* in Renan is older than the estimated divergence time of the wheat A genomes (2.4 Mya) used in this study. This result indicates an ancient origin of the

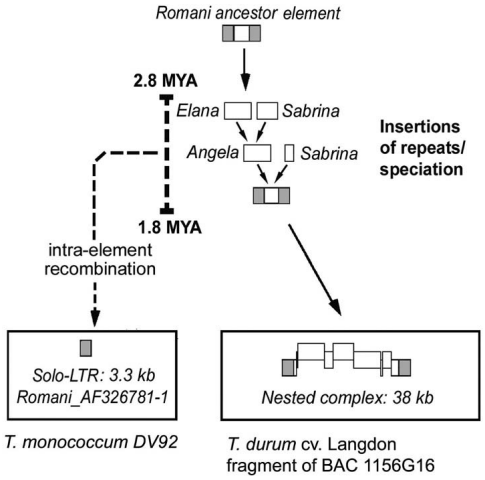

Figure 4. Model for the evolution of the *Romani* element in the *T. monococcum* and *T. turgidum* H1 haplotypes. The complete *Romani* retrotransposon is indicated by rectangles comprising two LTR (dark-gray boxes) and internal sequences (white box). The same *Romani* element is at the origin of the solo-LTR in the *T. monococcum* sequence resulting from an unequal intra-element recombination event and at the origin of the nested complex in *T. turgidum*.

H2 haplotype and is in agreement with the presence of this haplotype at different ploidy levels.

The longest available sequence of the H1 haplotype (*T. monococcum* DV92) (Wicker et al. 2001; Fig. 2A,B) was compared with the sequence of the H2 haplotype (*T. aestivum*) (Fig. 2B). The region where the two haplotypes differ dramatically is delimited by the two *RGA2* fragments *RGA2-b* and *RGA2-a* in the *T. aestivum* sequence (50 kb), and by *RGA2* and the foldback element *Gorgon* (G, Fig. 2B) upstream of the *NLL1* gene in *T. monococcum* (170 kb). On the proximal (left) side of both sequences, *RGA2-b* is conserved in the same orientation (Fig. 2B). On the distal (right) side, conserved sequences including a fragmented LTR of a *Sukkula*-type retroelement (*Sukkula*_AF326781-1 and *Sukkula*_930H14-1), the *NLL1* gene, three foldback elements (MITES *Athos* or A, *Fortuna* or F, and *Gorgon* or G), as well as *RGA2-a* are in reverse orientation (Fig. 2B). Such a complex pattern of conservation and divergence indicates extensive rearrangements at the origin of the two haplotypes. A model for these rearrangements is presented in Figure 5, starting with a common ancestor sequence containing *RGA2*, *Lr10*, *NLL1*, *Sukkula*, and the conserved foldback elements. In this ancestral "A" locus, insertions and deletions of repetitive sequences must have occurred between *RGA2* and *Lr10*, resulting in an H1 haplotype progenitor, which was then further modified to give the modern A^m and A^u haplotype of *T. monococcum* DV92 and *T. turgidum* Langdon (Fig. 5).

In the evolution of the H2 haplotype, a first deletion event removed a sequence containing the beginning of *RGA2* (corresponding to the first 2537 bp of *RGA2* of *T. monococcum*) and the complete *Lr10* gene (Fig. 5). Then, the second half of *RGA2* (bases 2538–4769) was split into *RGA2-a* and *RGA2-b* at position 3272 and a large sequence including *RGA2-a*, the foldback elements, the LTR of the *Sukkula*-type element, and the *NLL1* gene were inverted (Fig. 5). The two *RGA2* fragments were disrupted without any sequence loss; at the disruption point, the fragment *RGA2-a* ends at the base 3272, and the fragment *RGA2-b* starts at the base 3273. The lack of TSDs or inverted repeats at the breakpoints did not allow the identification of a precise mechanism responsible for this sequence disruption and inversion. Although the same types of fragmented LTR retrotransposons (data not shown) are found in the vicinity of these breakpoints, they do not start exactly at the breakpoints (at least 20 bp after) and are highly degenerated. It is therefore not clear whether they were involved in the rearrangement. Sequence inversion usually requires inverted repeats that align next to each other and act as recombination sites. Upon cross-over, the DNA sequence between these sites is inverted.

The only features that could have acted as inverted repeats are (TA) microsatellites of 56 and 59 bp found on the hexaploid sequence at the positions 34,333 and 85,957 bp. However, their position relative to *RGA2-b* and *RGA2-a* (90 bp and 5 kb downstream, respectively, Fig. 2B) does not allow to identify them as the origin of this rearrangement. Afterward, insertions and partial deletions involving *Angela* LTR retrotransposons occurred in the inverted sequence, leading to nested *Angela* retrotransposons (Fig. 2B).

Identical H2 haplotypes are found at three ploidy levels, and H2 subhaplotypes result from different types of rearrangements

In an earlier study, we did not find the H2 haplotype in tetraploid wheat lines (Scherrer et al. 2002), a finding which contrasted with the large number of H2 lines in hexaploid wheat that evolved from tetraploid wheat (Feldman 2001). Therefore, we have studied haplotypes at the *Lr10* locus in an additional large set of 300 tetraploid lines, consisting of 67 accessions of *T. tur-*

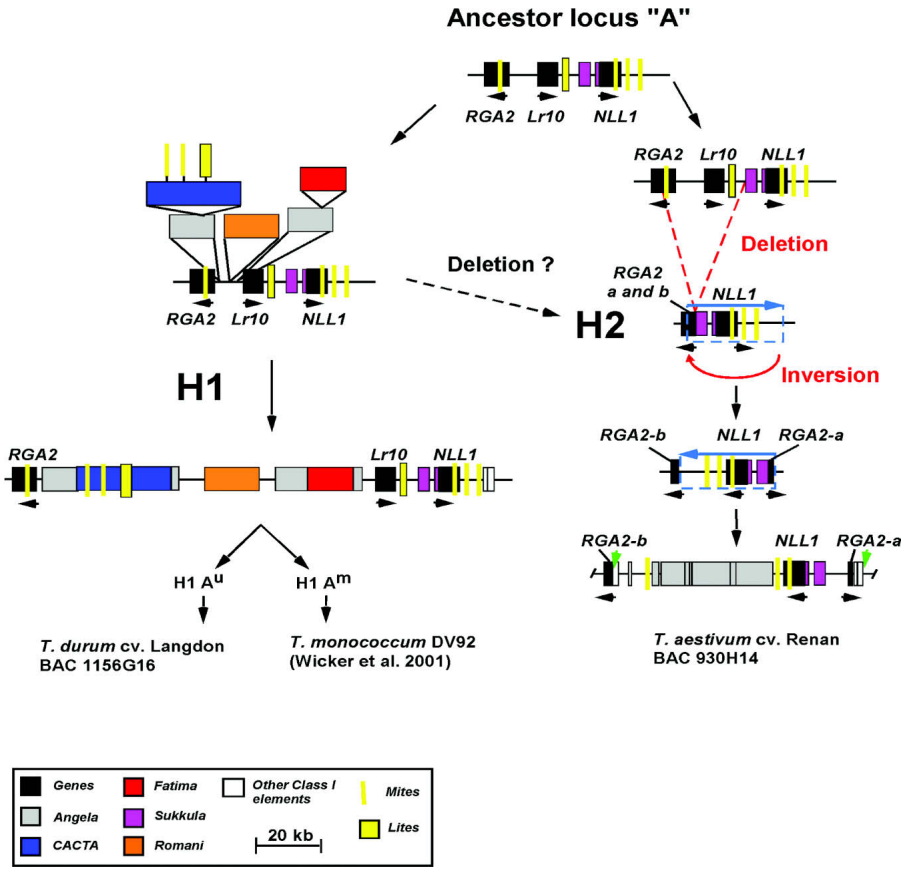

Figure 5. Model for the evolution of the H1 and H2 haplotypes at the *Lr10* locus. Genes are indicated by black boxes. The short arrows located *below* the genes indicate the transcriptional orientation. Identified nested repetitive elements are shown by colored boxes. The evolution of the H1 haplotype is represented on the *left* side and the evolution of the H2 haplotype on the *right* side of the figure. The ancestor of the H1 haplotype comprises elements common to both *T. monococcum* and *T. turgidum*, and subsequent independent events of insertions/deletions occurred resulting in the modern loci. The H2 ancestor has undergone a deletion comprising *Lr10* and the first exon of *RGA2*. The deletion is indicated by red dashed lines. An inversion indicated by a red arrow and delimited by blue arrows and a dashed rectangle also occurred in the H2 ancestor leading to a disruption of *RGA2* that is at the origin of the modern *T. aestivum* sequence. Green arrows indicate the (TA) repeats possibly involved in the sequence inversion.

gidum subsp. *dicoccoides*, 27 *T. turgidum* subsp. *carthlicum*, 34 *T. turgidum* subsp. *dicoccum*, 20 *T. turgidum* subsp. *turgidum*, and 152 accessions of *T. timopheevi* subsp. *armeniacum*. A total of 38 of these 300 tetraploid accessions showed the H2 haplotype by Southern hybridization with *Lr10* and *RGA2* probes (data not shown). The 38 lines consisted of eight *T. turgidum* subsp. *dicoccoides* accessions, 16 *T. turgidum* subsp. *dicoccum*, two *T. turgidum* subsp. *durum*, and 12 *T. timopheevi* subsp. *armeniacum* lines, demonstrating the presence of the H2 haplotype in a variety of tetraploid species.

To assess conservation of the H2 haplotype in the wheat gene pool and to determine whether the same rearrangements are at the origin of diploid, tetraploid, and hexaploid lines with the H2 haplotype, we amplified short genomic fragments (fragments A, B, and C, Fig. 6A) spanning the deletion and disruption breakpoints of the *RGA2-b* and *RGA2-a* fragments from 56 lines of haplotype H2 (based on hybridization data) using primers derived from the hexaploid sequence (Table 1, Supplemental material). Six diploid *T. monococcum* lines, two diploid *T. urartu* lines (of 10 tested lines, of which eight had haplotype H1), 10 tetraploid wheat, and 38 hexaploid *T. aestivum* lines were analyzed. Based on successful PCR amplification, two groups of lines could be distinguished. The first group amplified all three fragments A, B, and C, like cv. Renan. It comprises all of the 38 European hexaploid lines, nine tetraploid lines, and four of the six *T. monococcum* lines (e.g., Fig. 6B, lanes 1,2,4,5,7–11). The B fragments amplified from the *T. monococcum* line TRI 17434, as well as from seven tetraploid wheat lines (see Supplemental Table 1) were sequenced. This 331-bp B fragment was identical to the sequence of *T. aestivum* cv. Renan for all lines, demonstrating that the deletion breakpoint is identical. The finding of such a conserved molecular structure strongly suggests that in most lines with the H2 haplotype, the deletion derives from the same evolutionary event that occurred in the gene pool of the ancestor species of the wheat A genomes.

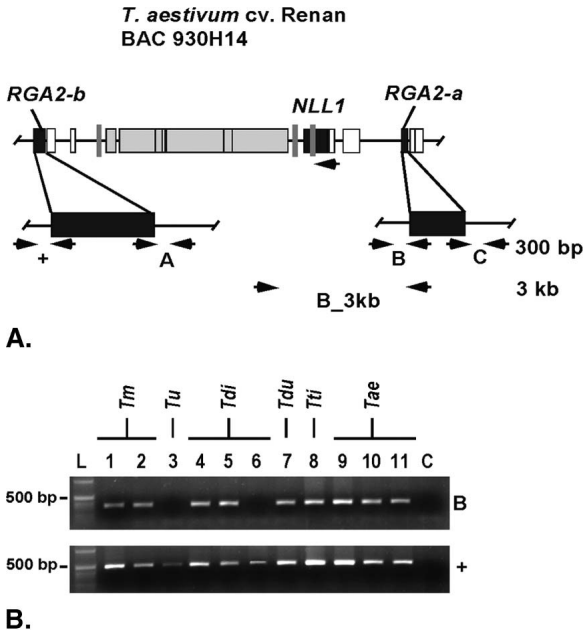

Figure 6. Analysis of H2 haplotype conservation in the wheat gene pool. (*A.*) Schematic representation of the amplified fragments spanning the *RGA2* fragments of *T. aestivum*. (*B.*) PCR amplification of 331-bp fragment (B) and the positive control for amplification (+). PCR fragments span the 3′ end of *RGA2* for the amplification control and the deletion breakpoint for the B fragment. Two *T. monococcum* lines (*Tm*, lanes *1*,*2*, TRI 2006 and TRI 17434), one *T. urartu* line (*Tu*, lane *3*, TRI 17413), three *T. turgidum* subsp. *dicoccoides* lines (*Tdi*, lanes *4–6*, TRI 3425, TTD22, IG 46516), one *T. turgidum* subsp. *durum* line (*Tdu*, lane *7*, TRI 673), one *T. timopheevii* subsp. *armeniacum* line (*Tti*, lane *8*, TA 952) and three *T. aestivum* lines (*Tae*, lanes *9–11*, cvs. Chinese spring, Arina, and Renan) of the haplotype H2 are shown. (L) 1 kb ladder; (C) water control.

We have compared at the molecular level, the organization and evolution of two haplotypes at the *Lr10* locus on the wheat A genomes of three different ploidy levels. The recent construction of two BAC libraries from hexaploid wheat lines belonging to the H2 haplotype (cv. Chinese Spring and Renan) (Allouis et al. 2003; B. Chalhoub, unpubl.) allowed the molecular comparison of the two haplotypes, including, for the first time, a sequence from hexaploid wheat. Conservation between the two haplotypes was limited to the genes located on both sides of a large 150-kb region, including the *Lr10* and *RGA2* genes, whereas extensive rearrangements occurred within this region.

Sequence comparison between H1 haplotypes derived from two different species revealed complete conservation of the structure and orientation of all of the genes, whereas the composition of the intergenic regions, which mainly consist of blocks of repetitive elements, is very different. Strikingly, despite these differences, the length of the large intergenic region of about 130 kb between the *RGA2* and *Lr10* genes is very similar. Conservation of the distances between blocks of genes has also been observed at the *Glu-3A* locus (Wicker et al. 2003a), as well as in two maize cultivars at the orthologous *bz*1 loci (Fu and Dooner 2002), although the intergenic regions were extensively rearranged. This suggests that the specific length (and not necessarily content) of

A second group of H2 lines comprising two *T. urartu*, two *T. monococcum*, and one tetraploid line, did not amplify any of the fragments (Supplemental Table 1, lanes 3,6; Fig. 6B). Thus, whereas the sequence organization observed in Renan is conserved in all but one of the tetraploid and hexaploid lines of the H2 haplotype, this is not the case for all diploid lines of this haplotype. The presence of this second group of lines suggests either sequence divergence after the formation of the haplotype from a common ancestor or completely independent deletion events leading to similar haplotype structures.

To further investigate conservation among the 51 H2 lines that had the same PCR amplification pattern as Renan, we amplified a 3-kb fragment spanning the deletion breakpoint of the *RGA2-a* fragment with a primer common to the B fragment previously described and a primer located 3-kb upstream in the sequence (Fig. 6A). Successful PCR amplification was observed for 40 lines (three *T. monococcum*, seven tetraploid lines, and 30 *T. aestivum* lines), whereas no amplification was observed in 11 lines (one *T. monococcum*, two tetraploid, and eight *T. aestivum* lines, Supplemental Table 1) suggesting sequence divergence in these lines. In conclusion, we could classify 56 H2 lines into three sub-H2 haplotypes. The presence of a main haplotype, represented by Renan, including diploid *T. monococcum*, tetraploid, and hexaploid lines, indicates high conservation and stability of the H2 haplotype during evolution and throughout polyploidization events in the last 1–3 million years of A genome divergence.

Discussion

an LTR retrotransposon cluster is conserved during evolution of orthologous loci, possibly acting as a determinant of chromatin and chromosome structure (Bennetzen and Ramakrishna 2002a).

A large deletion/inversion event is at the origin of the H2 haplotype

In the grass genomes, gene loss plays an important role in genome evolution and is the basis of the mosaic conservation of orthologous sequences (e.g., Song et al. 2002; Illic et al. 2003). The *Lr10* haplotype evolution provides an interesting example, where the molecular events leading to such a gene loss could be studied in detail. The extensive rearrangements found at the origin of haplotype H2 are due to a large deletion that includes the disease resistance gene *Lr10* and part of *RGA2*. The molecular basis of this deletion is possibly an illegitimate recombination event, a mechanism that is at the origin of other rearrangements in wheat (Wicker et al. 2003a; Ma et al. 2004). So far, the only mechanism responsible for gene disruption that has been described in wheat was the insertion of retroelements into genes (Gu et al. 2004). In the H2 haplotype, the deletion was followed by a large inversion that led to the disruption of the remaining *RGA2* gene fragment. The molecular mechanism of this sequence inversion could not be determined, but the TA repeats found on both sides of the inverted fragment might be relevant. It is likely that during the estimated 4 million years since the deletion/inversion events, all tracks of the original sequences have been covered by other elements. Sequence inversions have already been described in other comparative studies in grasses, but they specifically concerned complete genes in the distantly related barley and rice species (Dubcovsky et al. 2001) or groups of genes in rice compared with sorghum and maize (Bennetzen and Ramakrishna 2002b). Interestingly, Dubcovsky et al. (2001) have also identified two sequences of inverted homology in the vicinity of gene 2, which was found inverted in barley compared with rice. In this case, the inversion did not cause disruption of the gene.

The deletion/inversion events in the H2 haplotype had several consequences for the further evolution of the *Lr10-RGA2* region in the wheat gene pool. In fact, it may have effectively inhibited recombination between the different haplotypes. If recombination occurs within the inverted sequence (e.g., in or near the *NLL1* gene), the recombinant gametes will carry chromosomes that are partially deleted or duplicated. Such gametes probably have a greatly reduced fitness, and recombination between the haplotypes in the *Lr10-RGA2* interval is effectively suppressed. This provides a molecular explanation for two earlier observations; first, there are no recombinant haplotypes between H1 and H2 in the gene pool, and the *Lr10* gene was always found together with an intact *RGA2* (Scherrer et al. 2002). Second, there was no recombination between the two genes in a segregating population of more than 6000 gametes originating from a cross between the cultivars Frisal and Thatcher*Lr10* (Stein et al. 2000), which belong to the H2 and H1 haplotypes, respectively. Thus, our data indicate that recombination in hexaploid wheat cannot only be blocked by chromosomal segments derived from recent introgressions from wild relatives, but also by inversions that originate from ancient haplotypes.

Common themes of evolution in both haplotypes

Insertions of retroelements followed by deletions, most likely through illegitimate or unequal recombination, played a prominent role in the evolution of both haplotypes, similar to observations in other wheat comparative studies (Wicker et al. 2003a; Gu et al. 2004). These two mechanisms were also found to be responsible for LTR retrotransposon removal in the *Arabidopsis* and rice genomes (Devos et al. 2002; Ma et al. 2004) and seem to play a key role in plant genome evolution in general. Dating of intact elements in *T. monococcum* and *T. turgidum* revealed two distinct groups of insertion times for LTR retrotransposons. It is possible that different types of retroelements evolve at different rates, and that dates of retroelement insertions may, in general, be overestimated (San Miguel et al. 1998; Ma et al. 2004). However, "old" and "recent" *Angela* retroelements, which are likely to evolve at the same rate, are found at the same locus, supporting the idea that two series of insertions of repetitive elements have occurred at the *Lr10* locus in both diploid and tetraploid species. This suggests that at least two waves of retrotransposon insertions have occurred during the evolution of the A ancestral wheat genome.

It is very rare to find conserved sequences of transposable elements in colinear regions of cereal species, and so far, only partial sequences of conserved elements have been reported (Wicker et al. 2003a; Gu et al. 2004; Kong et al. 2004). The half-life of LTR retrotransposons was recently estimated to be ~6 million years in rice (Ma et al. 2004). Therefore, the conservation of nine complete elements (including one CACTA transposon, one LTR retrotransposon, one LITE, and six MITES) at the *Lr10* locus is exceptional and might be due to a lower tolerance of random rearrangements in this region of the genome. These 25 kb of conserved sequences from repetitive elements allowed us to estimate a divergence date between the wheat A genomes that had previously been estimated to 0.5–3 Mya (Huang et al. 2002; Wicker et al. 2003a). Our estimate (2.4 Mya) confirms, at a larger scale, what had previously been published by Wicker et al. (2003a) based on 8 kb of conserved sequence. However, it differs from the results of Huang at al. (2002). This is probably due to variations in rates of sequence divergence (e.g., study of genes vs. repetitive elements). Such variations in substitution rates have been found at different levels. Repetitive elements may diverge more rapidly than genes, as shown in rice (Ma and Bennetzen 2004). Ma and Bennetzen (2004) have also shown that the two rice subspecies *indica* and *japonica* have diverged at different rates, and they suggest that, even within a genome, separate regions may evolve at different rates. Finally, genes have very different divergence rates (Ma and Bennetzen 2004). All of these variations in rates of sequence divergence between subspecies, within a genome at different loci, or even among genes of the same locus, are very intriguing and seem to be common in grasses. Further sequence comparisons between and within grass species, or even wheat varieties, are needed at a larger scale to deepen our understanding of these variations.

The conservation of two ancient haplotypes in the wheat species of three ploidy levels suggests that no significant rearrangements have affected the *Lr10* locus during the polyploidization events that are at the origin of tetraploid and hexaploid wheat. Thus, there is no evidence in this region for rapid genomic changes following polyploidization, as observed for a substantial proportion of noncoding regions in newly synthetized allopolyploids (Liu et al. 1998; Ozkan et al. 2001). The stability of the two haplotypes at the *Lr10* locus throughout evolution and polyploidization is striking and contrasts with the assumption that disease resistance loci are generally more variable than other loci. This might only be true for some *R* gene clusters, but not for loci under balancing selection (see below).

Old and stable haplotypes at the *Lr10* resistance locus

The arms-race model for host-pathogen interactions predicts resistance loci to be young and monomorphic. Neither is true for the *Lr10* locus, making this wheat resistance locus a new member of a group of resistance loci described in *Arabidopsis* and tomato that existed before speciation and are also old and polymorphic (Grant et al. 1998; Stahl et al. 1999; Bergelson et al. 2001; Riely et al. 2001). Similar to *Arabidopsis RPM1*, which confers resistance to *Pseudomonas syringae*, there is a presence/absence polymorphism for *Lr10* in the wheat gene pool. This H2 haplotype caused by the deletion event is identical in lines of all three ploidy levels, as determined by sequencing the deletion breakpoint in the *RGA2* gene. Thus, the polymorphism represented by the two haplotypes is evolutionarily stable, and the two haplotypes have coexisted for 4 million years, suggesting a balanced polymorphism. The *Lr10–RGA2* complex might confer a fitness advantage under certain environmental conditions, whereas fitness costs are associated with it under other conditions, resulting in natural selection for both haplotypes and their maintenance in the wheat gene pool. In the absence of the pathogen, fitness costs associated with a resistance gene have been shown for the *RPM1* gene (Tian et al. 2003). Evidence for balancing selection at a resistance locus was described for the *Arabidopsis RPS5* gene (Tian et al. 2002). There, an island of enhanced sequence variability could be detected around the gene, indicative of an old polymorphism under selection.

Our study of the *Lr10* locus demonstrates the power of comparative studies between the A genomes of different ploidy levels in wheat to unravel molecular mechanisms involved in genome evolution. It has also provided a molecular explanation for previous observations at the genetic and phenotypic levels. The detailed molecular understanding of polymorphism at this resistance locus, including its consequence on recombination, provides a solid basis for further evolutionary studies. The presence of both haplotypes in the large collection of accessions of wild diploid and tetraploid wheat in the gene banks will allow future studies correlating ecogeographical and molecular data for a better understanding of the environmental factors underlying balancing selection and haplotype stability.

Methods

Genomic DNA isolation and PCR

The plant material used in this study is listed in Supplemental Table 1A. Genomic DNA isolation was performed as described by Graner et al. (1990). The primers used for PCR amplification are listed in Supplemental Table 1B and PCR amplifications were performed according to standard procedures (Sambrook et al. 1989), and using the annealing temperatures specific for the different primer pairs.

BAC clone isolation and sequencing

The *T. turgidum* subsp. *durum* Langdon BAC library (Cenci et al. 2003) was screened by hybridization as described by Stein et al. (2000) using probes derived from the *T. monococcum* DV92 sequence (Wicker et al. 2001). The *T. aestivum* Renan BAC library was screened by PCR with the primer pair specifically amplifying the *RGA2* copy on chromosome 1D (primers DF: 5′-GTGGTGTTGTGTTGCCAG-3′ and DR: 5′-GCAAGCTGGGTCCGGAAC-3′) and with a primer pair amplifying *RGA2* from the 1A and 1D chromosomes (primers ADF: 5′-GAAGCCGGATTATAGTGTC-3′ and ADR: 5′-CTGCCCAGCTAAGTTCTCT-3′) according to standard procedures (Sambrook et al. 1989). BAC DNA preparation for fingerprint analysis and BAC end sequencing were performed as previously described (Stein et al. 2000). Resulting BAC clones from both BAC libraries were assembled into individual contigs based on HindIII, NotI, and SalI restriction patterns and hybridization with BAC end probes as well as probes derived from the *T. monococcum* DV92 orthologous sequence (Fig. 1) (Stein et al. 2000; Scherrer et al. 2002).

Shotgun library construction of the BAC clones 1156G16 and 930H14 and sequencing of shotgun clones for each BAC were performed as follows: The BAC DNA was isolated using the large construct kit (Qiagen). A total of 20 µg of BAC DNA were sonicated, and the resulting fragments were purified by agarose gel electrophoresis. The fractions of 1200–1800 bp and 2500–3000 bp were eluted from the gel, and the fragments were subcloned into a SmaI-digested pUC19 sequencing vector. The subclones were sequenced using Big-Dye terminator chemistry (Applied Biosystems). Data were collected using ABI 3730 automated sequencers (Applied Biosystems). A total of 700–1200 clones were sequenced for each BAC clone for seven- to ninefold coverage. Gaps between subcontigs were filled by primer walking on the spanning shotgun clones or direct sequencing of BAC DNA. The assembled BAC sequences were finally checked by comparison of BamHI, EcoRI, HindIII, NotI, and PacI BAC fingerprints with theoretical restriction digests of the assembled sequences.

For the Langdon BAC 1156G16, a total of 2040 shotgun sequences resulted in 10 contigs with nine gaps. The preassembly procedures could not close those gaps due to the presence of G- and C-rich regions or aberrant assemblies caused by multiple Long Terminal Repeats (LTRs) of similar retroelements nested into each other. Despite the presence of these GC-rich regions, all of the gaps could be closed by primer walking. Finally, a total of 187,054 bp of sequence was assembled with a 9.18-fold redundancy for the Langdon 1156G16 BAC clone. For the cv. Renan BAC 930H14, a total of 1235 shotgun sequences resulted in 12 contigs with 11 gaps caused by similar features to the ones found in the cv. Langdon BAC. A total of 10 of the 11 gaps could be closed by primer walking, leaving one gap of 1 kb as determined from the sizes of the spanning shotgun clones. This 1 kb of missing sequence was located at base 121,554 of the BAC sequence, which is between the *AH5Y* and *NLL2* genes and downsteam of the sequence used for the comparative study. We therefore filled it artificially with 1000 Ns, indicating in the annotation that 1 kb of sequence are missing at the position 121,554 bp of this BAC clone. Finally, a total of 154,778 bp of sequence was assembled. The final error rate for the two BAC sequences was <1 base per 10 kb, and all finished sequences were of high quality (PHRED scores >25).

Sequence analysis

Base calling and quality of the shotgun sequences were processed using Phred (Ewing and Green 1998) and initial assembly of each BAC clone was performed using the Phrap assembly engine (version 0.990319; provided by P. Green and available at http://www.phrap.org). Subcontigs and singlet DNA sequences were analyzed using BLAST (Basic Local Alignment Search Tool) algorithms (Altschul et al. 1997) against public DNA and protein databases. Shotgun clones overlapping between subcontigs were identified using the Staden Package program GAP4 version 2003.0b (Bonfield et al. 1995). Annotation of the genes was performed using a mixture of BLASTX, BLASTN, and RiceGAAS annotation system (Sakata et al. 2002). Known repetitive elements were identified by BLAST against TREP (http://www.wheat.pw.usda.gov/ITMI/

Repeats; Wicker et al. 2002) database (nucleotide and protein) and dot plot analysis (DOTTER) (Sonnhammer and Durbin 1995). New repetitive elements were identified by dot plot analysis and BLAST against public protein databases. Comparisons of the assembled sequences were performed using dot plot and GCG sequence Analysis Software. Rates of nonsynonymous (*Ka*) vs. synonymous (*Ks*) nucleotide substitutions (*Ka/Ks*) per 100 sites were calculated for the exon 3 of *RGA2* in *T. monococcum* DV92, cv. Langdon, and cv. Renan with MEGA2 (Kumar et al. 2001) based on the algorithm of Li (1993). The same method was applied to calculate the synonymous substitution rates of the other genes. Dating of retrotransposon insertions was performed as described by SanMiguel et al. (1998) using the MEGA2 software (Kumar et al. 2001). The average substitution rate of repetitive sequences was calculated using the complete conserved repetitive elements between *T. monococcum* and *T. turgidum* and the divergence time of the two loci was estimated as described in SanMiguel et al. (1998).

Acknowledgments

We thank Romain Guyot, Thomas Wicker, and Edith Schlagenhauf for their help and advice in bioinformatic analysis. We also thank Genoplante (French Plant Genome consortium, http://www.genoplante.com) for access to the hexaploid Renan wheat BAC clones and Jorge Dubcovsky for providing the tetraploid Langdon BAC library. We are very thankful to Jacques David, Moshe Feldman, and Jon Raupp for sharing with us the accessions of tetraploid wheat used in this study. This work was supported by grant 3100-065114 of the Swiss National Science Foundation.

References

Allouis, S., Moore, G., Bellec, A., Sharp, R., Faivre Rampant, P., Mortimer, K., Pateyron, S., Foote, T.N., Griffiths, S., Caboche, M., et al. 2003. Construction and characterisation of a hexaploid wheat (*Triticum aestivum* L.) BAC library from the reference germplasm 'Chinese Sprin'. *Cereal Res. Commun.* **31:** 331–338.

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

Bennetzen, J.L. and Ramakrishna, W. 2002a. Exceptional haplotype variation in maize. *Proc. Natl. Acad. Sci.* **99:** 9093–9095.

———. 2002b. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. *Plant Mol. Biol.* **48:** 821–827.

Bergelson, J., Kreitman, M., Stahl, E.A., and Tian, D. 2001. Evolutionary dynamics of plant *R*-genes. *Science* **292:** 2281–2285.

Bonfield, J.K., Smith, K.F., and Staden, R. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.* **23:** 4992–4999.

Brunner, S., Keller, B., and Feuillet, C. 2003. A large rearrangement involving genes and low-copy DNA interrupts the microcollinearity between rice and barley at the *Rph7* locus. *Genetics* **164:** 673–683.

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* s0sp *durum*). *Theor. Appl. Genet.* **107:** 931–939.

Charlesworth, D., Bartolome, C., Schierup, M.H., and Mable, B.K. 2003. Haplotype structure of the stigmatic self-incompatibility gene in natural populations of *Arabidopsis lyrata*. *Mol. Biol. Evol.* **20:** 1741–1753.

Collins, N.C., Lahaye, T., Peterhansel, C., Freialdenhoven, A., Corbitt, M., and Schulze-Lefert, P. 2001. Sequence haplotypes revealed by sequence-tagged site fine mapping of the *Ror1* gene in the centromeric region of barley chromosome 1H. *Plant Physiol.* **125:** 1236–1247.

Devos, K.M. and Gale, M.D. 2000. Genome relationships: The grass model in current research. *Plant Cell* **12:** 637–646.

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

Dubcovsky, J., Ramakrishna, W., SanMiguel, P.J., Busso, C.S., Yan, L.L., Shiloff, B.A., and Bennetzen, J.L. 2001. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol.* **125:** 1342–1353.

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

Feldman, M. 2001. Origin of cultivated wheat. In *The world wheat book: A history of wheat breeding* (eds. A.P. Bonjean and W.J. Angus). Lavoisier, Paris, France.

Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A., and Keller, B. 2003. Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proc. Natl. Acad. Sci.* **100:** 15253–15258.

Fu, H.H. and Dooner, H.K. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci.* **99:** 9573–9578.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. 1996. Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci.* **93:** 10274–10279.

Graner, A., Siedler, H., Jahoor, A., Herrmann, R.G., and Wenzel, G. 1990. Assessment of the degree and the type of restriction fragment length polymorphism in barley (*Hordeum vulgare*). *Theor. Appl. Genet.* **80:** 826–832.

Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W., and Dangl, J.L. 1995. Structure of the *Arabidopsis Rpm1* gene enabling dual-specificity disease resistance. *Science* **269:** 843–846.

Grant, M.R., McDowell, J.M., Sharpe, A.G., Zabala, M.D.T., Lydiate, D.J., and Dangl, J.L. 1998. Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. *Proc. Natl. Acad. Sci.* **95:** 15843–15848.

Gu, Y.Q., Anderson, O.D., Londeore, C.F., Kong, X.Y., Chibbar, R.N., and Lazo, G.R. 2003. Structural organization of the barley D-hordein locus in comparison with its orthologous regions of wheat genomes. *Genome* **46:** 1084–1097.

Gu, Y.Q., Coleman-Derr, D., Kong, X.Y., and Anderson, O.D. 2004. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four Triticeae genomes. *Plant Physiol.* **135:** 459–470.

Guyot, R., Yahiaoui, N., Feuillet, C., and Keller, B. 2004. In silico comparative analysis reveals a mosaic conservation of genes within a novel colinear region in wheat chromosome 1AS and rice chromosome 5S. *Funct. Integr. Genomics* **4:** 47–58.

Huang, S.X., 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.* **99:** 8133–8138.

Ilic, K., San Miguel, P.J., and Bennetzen, J.L. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc. Natl. Acad. Sci.* **100:** 12265–12270.

Keller, B. and Feuillet, C. 2000. Colinearity and gene density in grass genomes. *Trends Plant Sci.* **5:** 246–251.

Kong, X.Y., Gu, Y.Q., You, F.M., Dubcovsky, J., and Anderson, O.D. 2004. Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. *Plant Mol. Biol.* **54:** 55–69.

Kumar, S., Tamura, K., Jakobsen, I.A., and Nei, M. 2001. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244–1245.

Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36:** 96–99.

Lijavetzky, D., Muzzi, G., Wicker, T., Keller, B., Wing, R., and Dubcovsky, J. 1999. Construction and characterization of a bacterial artificial chromosome (BAC) library for the A genome of wheat. *Genome* **42:** 1176–1182.

Liu, B., Vega, J.M., and Feldman, M. 1998. Rapid genomic changes in newly synthesized amphiploids of *Triticum* and *Aegilops*. II. Changes in low-copy coding DNA sequences. *Genome* **41:** 535–542.

Ma, J.X. and Bennetzen, J.L. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci.* **101:** 12404–12410.

Ma, J.X., 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.

Moullet, O., Zhang, H.B., and Lagudah, E.S. 1999. Construction and characterization of a large DNA insert library from the D genome of

wheat. *Theor. Appl. Genet.* **99:** 305–313.
Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **30:** 190–193.
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.
Rafalski, A. and Morgante, M. 2004. Corn and humans: Recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet.* **20:** 103–111.
Ramakrishna, W., Dubcovsky, J., Park, Y.J., Busso, C.S., Emberton, J., SanMiguel, P., and Bennetzen, J.L. 2002. Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* **162:** 1389–1400.
Riely, B.K. and Martin, G.B. 2001. Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene *Pto*. *Proc. Natl. Acad. Sci.* **98:** 2059–2064.
Sakata, K., Nagamura, Y., Numa, H., Antonio, B.A., Nagasaki, H., Idonuma, A., Watanabe, W., Shimizu, Y., Horiuchi, I., Matsumoto, T., et al. 2002. RiceGAAS: An automated annotation system and database for rice genome sequence. *Nucleic Acids Res.* **30:** 98–102.
Sambrook J., Fritsch, E.F., and Maniatis, T. 1989. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. 1998. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20:** 43–45.
Scherrer, B., Keller, B., and Feuillet, C. 2002. Two haplotypes of resistance gene analogs have been conserved during evolution at the leaf rust resistance locus *Lr10* in wild and cultivated wheat. *Funct. Integr. Genomics* **2:** 40–50.
Song, R.T. and Messing, J. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc. Natl. Acad. Sci.* **100:** 9055–9060.
Song, R., Llaca, V., and Messing, J. 2002. Mosaic organization of orthologous sequences in grass genomes. *Genome Res.* **12:** 1549–1555.
Sonnhammer, E.L.L. and Durbin, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene-Combis* **167:** 1–10.
Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M., and Bergelson, J. 1999. Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. *Nat. Genet.* **400:** 667–671.
Stein, N., Feuillet, C., Wicker, T., Schlagenhauf, E., and Keller, B. 2000. Subgenome chromosome walking in wheat: A 450-kb physical contig in *Triticum monococcum* L. spans the *Lr10* resistance locus in hexaploid wheat (*Triticum aestivum* L.). *Proc. Natl. Acad. Sci.* **97:** 13436–13441.
Tian, D., Araki, H., Stahl, E., Bergelson, J., and Kreitman, M. 2002. Signature of balancing selection in *Arabidopsis*. *Proc. Natl. Acad. Sci.* **99:** 11525–11530.
Tian, D., Traw, M.B., Chen, J.Q., Kreitman, M., and Bergelson, J. 2003. Fitness costs of *R*-gene-mediated resistance in *Arabidopsis thaliana*. *Nature* **423:** 74–78.
Wicker, T., Stein, N., Albar, L., Feuillet, C., Schlagenhauf, E., and Keller, B. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J.* **26:** 307–316.
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., Dubcovsky, J., and Keller, B. 2003a. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and A^m genomes of wheat. *Plant Cell* **15:** 1186–1197.
Wicker, T., Guyot, R., Yahiaoui, N., and Keller, B. 2003b. CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol.* **132:** 52–63.
Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., and Cregan, P.B. 2003. Single-nucleotide polymorphisms in soybean. *Genetics* **163:** 1123–1134.

Web site references

http://www.phrap.org; PHRAP assembly engine.
http://www.wheat.pw.usda.gov/ITMI/Repeats; TREP (Triticeae Repeat database).
http://www.genoplante.com; French Plant Genome consortium.

Received August 10, 2004; accepted in revised form January 10, 2005.