J Mol Evol (2008) 66:138–150
DOI 10.1007/s00239-008-9066-8

Contrasted Microcolinearity and Gene Evolution Within a Homoeologous Region of Wheat and Barley Species

Nathalie Chantret · Jérôme Salse · François Sabot · Arnaud Bellec · Bastien Laubin · Ivan Dubois · Carole Dossat · Pierre Sourdille · Philippe Joudrier · Marie-Françoise Gautier · Laurence Cattolico · Michel Beckert · Sébastien Aubourg · Jean Weissenbach · Michel Caboche · Philippe Leroy · Michel Bernard · Boulos Chalhoub

Received: 12 June 2007 / Accepted: 2 January 2008 / Published online: 15 February 2008

Abstract We study here the evolution of genes located in the same physical locus using the recently sequenced *Ha* locus in seven wheat genomes in diploid, tetraploid, and hexaploid species and compared them with barley and rice orthologous regions. We investigated both the conservation of microcolinearity and the molecular evolution of genes, including coding and noncoding sequences. Microcolinearity is restricted to two groups of genes (*Unknown gene-2*, *VAMP*, *BGGP*, *Gsp-1*, and *Unknown gene-8* surrounded by several copies of *ATPase*), almost conserved in rice and barley, but in a different relative position. Highly conserved genes between wheat and rice run along with genes harboring different copy numbers and highly variable sequences between close wheat genomes. The coding sequence evolution appeared to be submitted to heterogeneous selective pressure and intronic sequences analysis revealed that the molecular clock hypothesis is violated in most cases.

Keywords Comparative genomics · Gene evolution · Polyploidy · Wheat

Nathalie Chantret, Jérôme Salse, and François Sabot contributed equally to this work.

N. Chantret (✉)
Domaine de Melgueil, INRA–UMR DIAPC, 34130 Mauguio, France
e-mail: chantret@supagro.inra.fr

J. Salse · F. Sabot · B. Laubin · P. Sourdille · M. Beckert · P. Leroy · M. Bernard
Domaine de Crouël, UMR 1095 INRA-UBP Amélioration & Santé des Plantes, 234 Avenue du Brézet, F-63039 Clermont-Ferrand Cedex 2, France

A. Bellec · S. Aubourg · M. Caboche · B. Chalhoub
Laboratory of Genome Organization, Unité de Recherches en Génomique Végétale (INRA-URGV), 2 rue Gaston Crémieux, CP 5708, F-91057 Evry Cedex, France

I. Dubois · C. Dossat · L. Cattolico · J. Weissenbach
GENOSCOPE–CNRG, 2 rue Gaston Crémieux, CP 5706, F- 91057 Evry Cedex, France

P. Joudrier · M.-F. Gautier
Développement et Amélioration des Plantes, INRA–UMR1098, 2 Place Viala, F-34060 Montpellier Cedex 01, France

Present Address:
F. Sabot
Transposable Elements in Plant Genomes Group, Laboratoire Génome et Développement des Plantes, UMR 5096 - CNRS/UPVD/IRD, 52, Avenue P. Alduy, F-66860 Perpignan, France

Introduction

A small number of cereal species (wheat, rice, maize, sorghum and millet) form the core element of most human and domestic animal population diet worldwide. All these plants belong to the grass family (*Poaceae*), which includes more than 10,000 species, some of them polyploids, originating from more than 65 million years ago (Mya) (Kellogg 1998; Prasad et al. 2005) with a highly complex apparent evolutive story. A good example of this complexity may be the wheat species and their close relatives, belonging to the *Pooideae* subfamily and to the Triticeae tribe, which includes the genera *Triticum* (wheat), *Aegilops*, and *Hordeum* (barley). The *Triticum/Aegilops* and barley lineages diverged first, followed by the diploid *Triticum/Aegilops* species radiation. Polyploid species of the Triticeae tribe result from several independent interspecific hybridizations between the diploid species and the newly formed polyploid species, ensued by genome

doubling. Thus, *Triticum turgidum* ssp. *durum* (durum wheat), the most widely cultivated tetraploid wheat (domesticated from *Triticum turgidum* ssp. *dicoccoides*, AB genome), originated from the hybridization of *T. urartu* (A genome) and an unknown *Aegilops* species related to *Ae. speltoides* (B genome). Hexaploid bread wheat (*Triticum aestivum*) arose secondly from hybridization of the first domesticated form of *T. turgidum* (*T. turgidum* ssp. *dicoccum*) with the wild *Ae. tauschii* (reviewed by Feldman et al. 1995; Nesbitt and Samuel 1995; Nevo et al. 2002).

Still, primary comparative mapping has shown that the genomes of grass species may be composed of conserved genomic blocks, ordered differently on the chromosomes of different species (Gale and Devos 1998; Moore et al. 1995). This "synteny conservation" was considered potentially useful for enhancing extrapolation of the results obtained on simpler model species to other more complex species. However, more recent studies of microcolinearity (i.e., conservation of the nature and order of the genes between the orthologous region of two species) have revealed many exceptions, with, to date, none of the loci studied in detail displaying the complete conservation of microcolinearity (Bennetzen 2000; Bennetzen and Ramakrishna 2002; Brunner et al. 2003; Devos and Gale 2000; Feuillet and Keller 2002; Gu et al. 2004; Ilic et al. 2003; Keller and Feuillet 2000; Ramakrishna et al. 2002; Song et al. 2002; Wicker et al. 2003).

The *Ha* locus has been extensively studied for its agronomical importance, as it contains three paralogous genes implied in the *Hardness* character of the grain: *Pina*, *Pinb*, and *Gsp-1* (encoding puroindoline a, b, and "grain softness protein," respectively [Gautier et al. 1994; Rahman et al. 1994]). All diploid species tested possess all three paralogues, whereas the tetraploid wheat species *T. turgidum* lost the *Pina* and *Pinb* genes from both its genomes (Chantret et al. 2005; Gautier et al. 2000). Hexaploid wheat recovered a copy of each of these genes through the acquisition of the D genome during hexaploidization. The mechanisms underlying these specific deletions have already been analyzed by comparing the structure of the *Ha* locus among the three wheat genomes (Chantret et al. 2004, 2005) and barley (Caldwell et al. 2004), using a small set of the whole sequences of the region obtained at that time.

Using the entire set of sequence available from the *Ha* locus region, we examined the gene evolution, microcolinearity, and conservation between related species from the same genus (wheats), then from the same tribe (Triticeae), and, finally, from the same family (*Poaceae*). For that purpose, we compared the whole sequences, i.e., the exon and intron sequences, from seven genes (*Pina*, *Pinb*, *Gsp-1*, *CHS*, *BGGP*, *VAMP*, and *Unknown gene-8*). We showed, first, that microcolinearity was very limited at this locus and, second, that the genes present at this locus demonstrate contrasting modes of evolution, in particular, that intron evolution appeared unexpectedly unrelated to the species relationships.

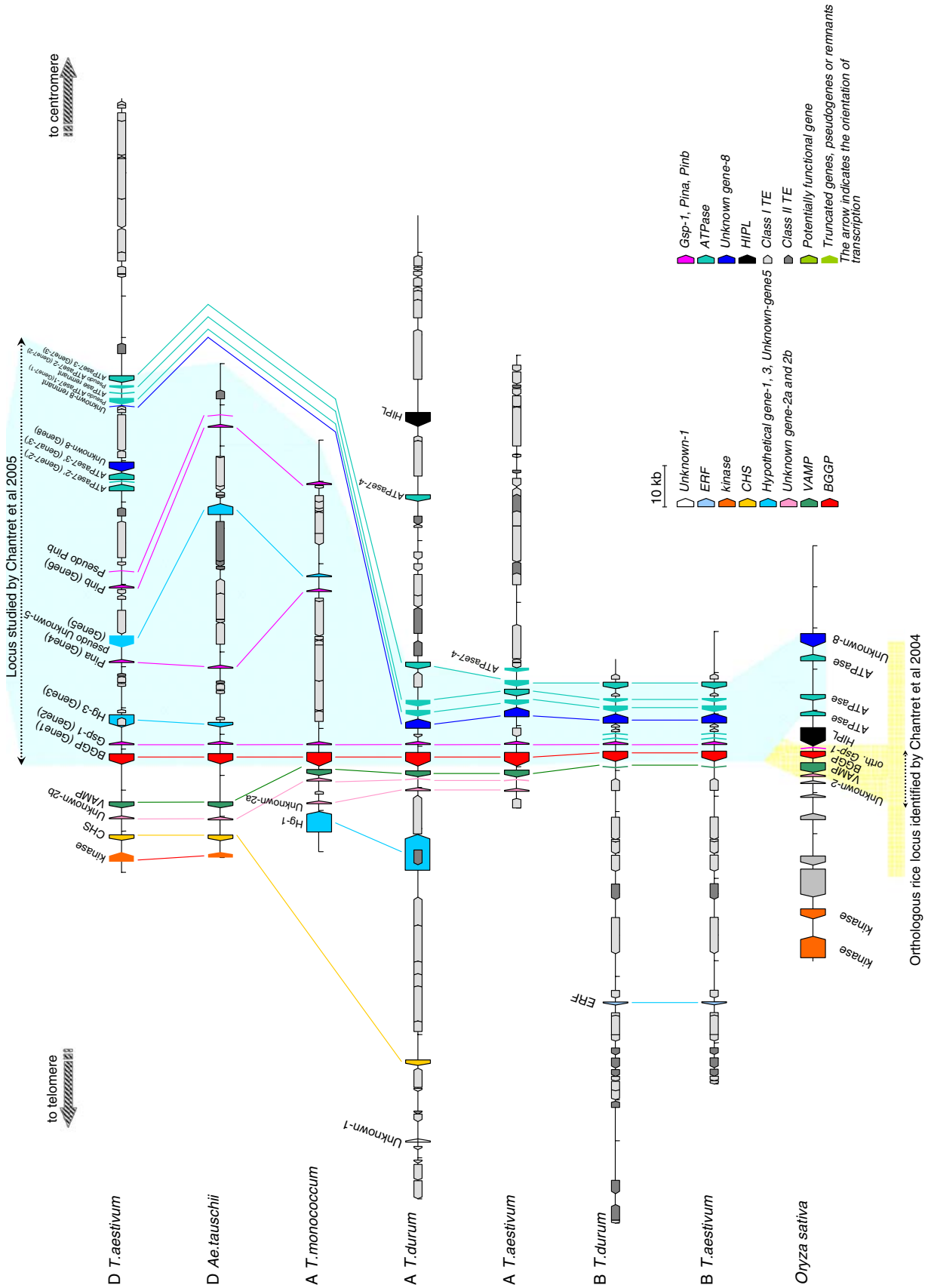
Materials and Methods

BAC Sequencing and Annotation

Wheat BAC clones were the same as in Chantret et al. (2004, 2005). Briefly, they were selected from four different Triticeae species BAC libraries (*T. monococcum* accession DV92 [Lijavetzky et al. 1999], *Ae. tauschii* accession Aus18913 [Moullet et al. 1999], *T. turgidum.* ssp. *durum* cultivar Langdon 65 [Cenci et al. 2003], and *T. aestivum* ssp. *aestivum* cv Renan [Chalhoub et al., unpublished results]), sequenced, and their genes annotated as described by Chantret et al. (2004, 2005). Transposable elements (TEs) were annotated and named as described by Sabot et al. (2005). Complete sequences and annotations are deposited in GenBank under accession numbers CT009586, CT009585, and CT009735 for the BAC sequences of the A, B, and D genomes of *T. aestivum*; CT009587 and CT009588 for the BAC sequences of the A and B genome of *T. durum*; and CT009625 for the BAC sequence of the D genome of *Ae. tauschii*. The *T. monococcum* BAC 109N23 sequence is available from GenBank under accession no. AY491681.

The sequences presented here stand for three times the size of the *Ha* locus (*stricto sensu*) originally described by Chantret et al. (2005) (Fig. 1; blue highlighting), as the "new" sequences represent 69% of the region currently analyzed. The sequences of the *Ha* locus in barley were kindly provided by Dr C. Caldwell and are available from GenBank under the accession numbers AY643842 to AY643844.

Fig. 1 Schematic scaled representations of seven Triticeae BAC ▶ clones and 80 kb located in the distal position of the rice chromosome 12 (inverted complement of coordinates 27369197…27448138 of rice chromosome 12 pseudomolecule release 4.0, TIGR, January 2006; N.B., end of pseudomolecule at 27492551; sequence accessible at http://orygenesdb.cirad.fr, or as GenBank accession AL732378). The region studied by Chantret et al. (2005) is highlighted in light blue and indicated by the dotted black horizontal arrow. The rice region studied by Chantret et al. (2004) is highlighted in yellow. Potentially functional genes are represented as large full arrows outlined in black (indicating the transcription orientation). Truncated genes, pseudogenes, or remnants are represented as large full arrows with no black outline. Each gene has its own color code and putative orthologous relationships are represented by full bars of the same color connecting genes. The number in brackets after gene name corresponds to the gene number from Chantret et al. (2005). Class I and Class II transposable elements are represented as narrow light- and dark-gray arrows, respectively


to centromere
to telomere
Locus studied by Chantret et al 2005
Orthologous rice locus identified by Chantret et al 2004
10 kb
Unknown-1
ERF
kinase
CHS
Hypothetical gene-1, 3, Unknown-gene5
Unknown gene-2a and 2b
VAMP
BGGP
Gsp-1, Pina, Pinb
ATPase
Unknown gene-8
HIPL
Class I TE
Class II TE
Potentially functional gene
Truncated genes, pseudogenes or remnants
The arrow indicates the orientation of transcription
D T.aestivum
D Ae.tauschii
A T.monococcum
A T.durum
A T.aestivum
B T.durum
B T.aestivum
Onyza sativa

Coding Sequence Analysis

Multiple alignments of orthologous complete coding domain sequences (CDSs) were generated with *CLUSTALX* (Thompson et al. 1997). Flanking sequences (3' and/or 5') and parts showing ambiguous alignments were manually removed from the multiple alignments, maintaining perfectly aligned coding phase. These CDS alignments are shown in online Supplemental Data 1. Sequences were then analyzed with codon substitution models (Goldman and Yang 1994). The ratio of nonsynonymous-to-synonymous mutations (dN/dS = ω) across amino acid sites was investigated, using the various models of the *codeml* program in the *PAML* package (v3.14; Yang 1997). The models were M0, assuming one category of ω ($0 < \omega_0 < 1$); M1, assuming two categories of ω ($0 < \omega_0 < 1$ and $\omega_1 = 1$); M2, assuming three categories of ω ($0 < \omega_0 < 1$, $\omega_1 = 1$, and $\omega_S > 1$ accounting for sites under positive selection); M7, assuming that ω is β distributed in 10 categories of the same frequency ($0 < \omega_{1-10} < 1$); and M8, assuming that ω is β distributed as in M7, plus one category of $\omega_S > 1$, accounting for sites under positive selection. Models were compared using the likelihood ratio test when nested within each other (M0 versus M1, M1 versus M2, and M7 versus M8), with the significance threshold set at $p = 0.05$ and according to Akaike's Informative Criterion [AIC = (2×number of parameters) – (2×log*L*)], with 1.5 as the minimum ΔAIC for M1 versus M7. The initial phylogenic trees used to test the codon substitution models for each gene were built using the maximum likelihood method, with the *PHYML* program (Guindon and Gascuel 2003). The *HKY* model (Hasegawa et al. 1985) was used as the nucleotide substitution model. Base frequencies, the ratio between transitions and transversions, and the proportion of invariable sites were estimated. The substitution rate for the remaining sites was estimated assuming a gamma distribution.

Intron Sequence Analysis

Multiple alignments of intron sequences were extracted from the multiple alignments of the complete genomic sequences of *VAMP*, *BGGP*, and *Unknown gene-8*, generated by *CLUSTALX*. The molecular evolutionary clock hypothesis was tested using Tajima's (1993) relative rate test, as implemented in *MEGA* v2.1 (Molecular Evolutionary Genetic Analysis [Kumar et al. 2001]).

For each multiple alignment of introns, "global identity" was defined as the percentage of sites conserved among all aligned sequences — i.e., the sites for which all sequences shared the same base, even null (see below). Global identity along the aligned sequence was assessed using a 50-bp sliding window incremented by 5 bp at a time. For a given multiple alignment, global identity was also calculated and compared for various genome combinations extracted from the multiple alignment. For example, only a subset from the A, B, or D genome sequences could be considered, or only wheat genome sequences involved in the whole alignment. In those alignment subsets, if all the involved sequences presented a gap at the same site (i.e., if the whole alignment included a longer sequence, thus creating a gap for the other involved sequences), that site was counted as conserved for the considered subset. These calculations were performed with homemade Python scripts, available upon request.

Results

The complete annotation of the analyzed region is presented in Fig. 1 and Table 1. Regardless of ploidy, sequence conservation between the A, B, and D genomes was restricted to the CDS and their immediate 5′ and 3′ untranslated regions (UTRs), whereas most of the intergenic regions were composed of multiple TE insertions (Fig. 1), as expected.

Analysis of Microcolinearity

Analysis of the conservation of microcolinearity between wheats, barley, and rice in this region suggests different possible hypotheses concerning the series of events shaping the orthologous region in these species; one of these is presented in Fig. 2. Comparison among those species showed that two groups of genes were conserved and, therefore, presumably present in the common ancestor of these species.

The first group includes the common ancestors of *Unknown gene-2*, *VAMP* (encoding a putative vesicle-associated membrane protein), *BGGP*, and *Gsp-1* (Fig. 2 A). Like that of *T. monococcum* (Chantret et al. 2004), the A genomes of *T. durum* and *T. aestivum* contain two copies of *Unknown gene-2*. One of these copies (*Unknown gene-2b*) has an internal deletion of ~380 bp in *T. durum* and ~200 bp (the larger copy) in *T. aestivum*. The *Unknown gene-2b* sequence of *T. durum*, even carrying this large deletion, maintains an open reading frame which may potentially be translated into an 86-amino acid protein. However, the initial function of *Unknown gene-2* is probably not conserved here. Likewise, the *T. aestivum Unknown gene-2b* copy has a stop codon in the middle of the potential coding sequence. The same event was probably responsible for generating these two paralogues in the A^m genome and in the A genome of polyploid wheat

Table 1 Description of the known, unknown, and hypothetical genes and pseudogenes (including truncated or remnant genes) identified in the extended *Ha* locus

Name	Putative function	BlastN EST best hits Acc. number; E-value; species	BlastP hit (SwissProt) Acc. number; E-value; species	Genome	Description	CDS size (bp) (exon number)	Total size (bp)
ERF	Ethylene-responsive element binding factor 5	—	O80341; 2e-08; *A. thaliana*	B *T. durum*[a]	Complete	585 (1)	585
				B *T. aestivum*			
Unknown-1	Unknown gene	BE422636; e^{-125}; *T. aestivum*	—	A *T. durum*[a]	Complete	744 (1)	744
		BQ167197; e^{-113}; *T. aestivum*					
Kinase	Serine/threonine-protein kinase receptor	CD908771; $2e^{-19}$; *T. aestivum*	Q09092; $6e^{-40}$; Brassica oleracea	D *T. aestivum*[a]	Complete	1893 (5)	2471
				D *Ae. tauschii*	Truncated at 5′ BAC extremity	—	
CHS	Chalcone synthase	CK207874; 0.0; *T. aestivum*	P51081; $6e^{-76}$; *Pisum sativum*	D *T. aestivum*	Pseudogene	1266 (2)	1359
		CA502438; 0.0; *T. aestivum*		D *Ae. tauschii*[a]	Complete	1263 (2)	1356
				A *T. durum*	Complete	1257 (2)	1338
Hg-1	Hypothetical gene	—	—	A^m *T. monococcum*[a]	Complete	1872 (6)	9048
				A *T. durum*	Pseudogene		
Unknown-2a	Unknown gene	BE495630; $2e^{-44}$; *Secale cereale*	—	A^m *T. monococcum*[a]	Complete		
		BQ160244; $2e^{-23}$; *Secale cereal*		A *T. durum*	Complete		
				A *T. aestivum*	Complete		
Unknown-2b	Unknown gene	BQ160244; e^{-137}; *Secale cereal*	—	D *T. aestivum*	Complete		
		BE495630; $4e^{-77}$; *Secale cereal*		D *Ae. tauschii*	Complete		
				A^m *T. monococcum*[a]	Complete		
				A *T. durum*	Truncated		
				A *T. aestivum*	Truncated		
VAMP	Vesicle-associated membrane protein	CA668237; 0.0; *T. aestivum*	Q9MAS5; $6e^{-80}$; *A. thaliana*	D *T. aestivum*[a]	Complete	648 (4)	1530
		CB667109; $2e^{-72}$; *Oryza sativa*		D *Ae. tauschii*	Complete	648 (4)	1523
				A^m *T. monococcum*	Complete	657 (4)	1470
				A *T. durum*	Complete	657 (4)	1543
				A *T. aestivum*	Complete	657 (4)	1758
				B *T. durum*	Truncated	—	
				B *T. aestivum*	Truncated	—	
ATPase (-4)	Cell division protein ftsH homologues	CD875876; 0.0; *T. aestivum*	O32617; $4e^{-12}$; *Helicobacter felis*	A *T. durum*[a]	Complete		
		CK204402; 0.0; *T. aestivum*		A *T. aestivum*	Truncated		
		BQ753090; 0.0; *T. aestivum*					

Table 1 continued

Name	Putative function	BlastN EST best hits Acc. number; E-value; species	BlastP hit (SwissProt) Acc. number; E-value; species	Genome	Description	CDS size (bp) (exon number)	Total size (bp)
HIPL	Hedgehog-interacting protein-like	BU100503; 0.0; *T. aestivum* BJ278101; 0.0; *T. aestivum* BQ483550; e^{-160}; *T. aestivum*	Q9SSG3; 0.0; *A. thaliana*	A *T. durum*[a]	Complete		

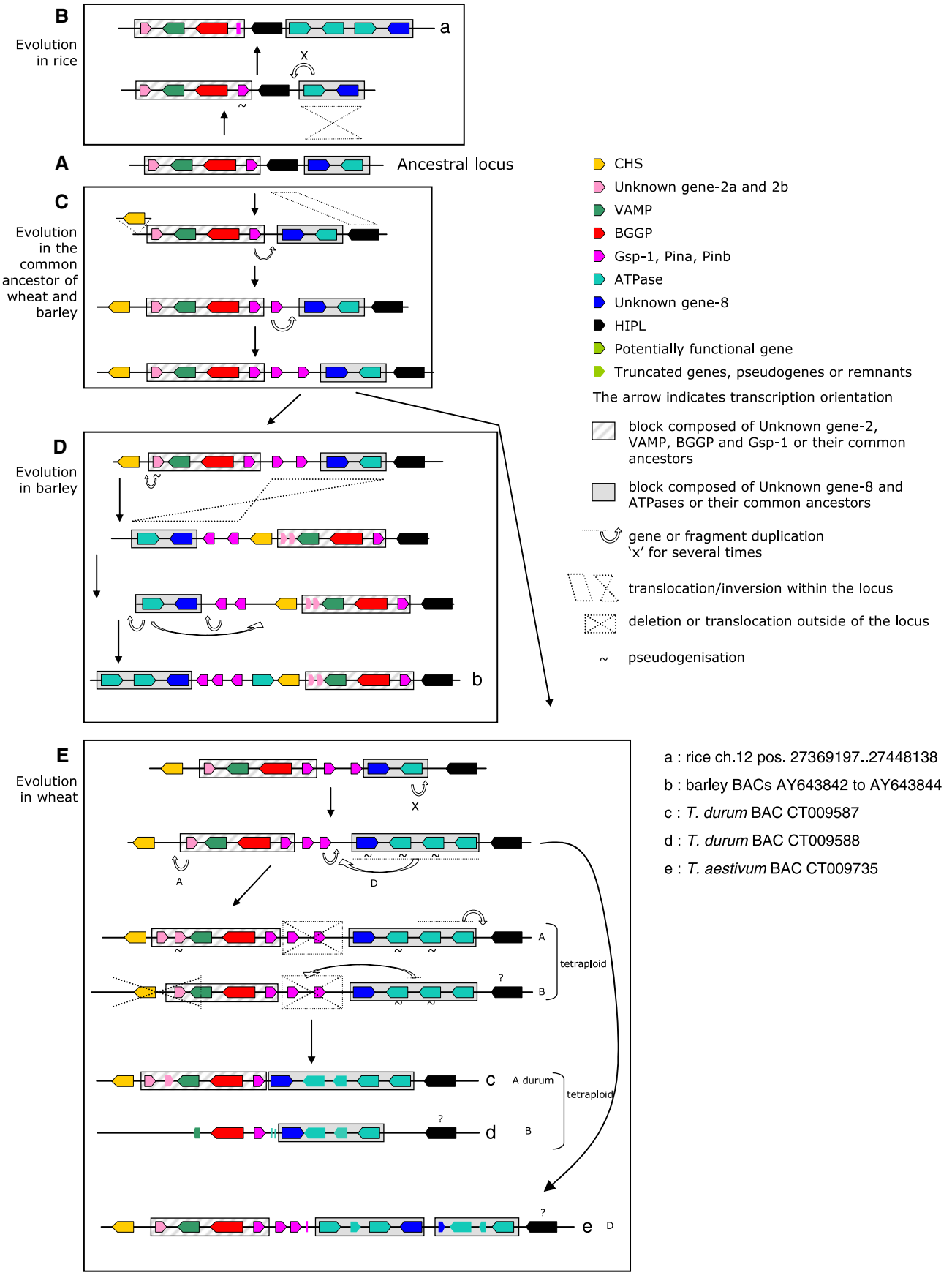
Note. The genes present in the restricted *Ha* locus are not detailed here but are available in Chantret et al. (2005)

[a] The genome from which the sequence was used for the Blast results presented here

species and, therefore, must have occurred before the separation of the different A genomes. Only one copy of *Unknown gene-2* is present in D genomes, suggesting this duplication to be specific to the A genome lineage (Fig. 2E). However, the D genome copy is more similar to the *T. monococcum 2b* (95%) gene than to the *2a* gene (86%), i.e., more similar to the most mutated copy. This suggests that duplication may also have occurred before divergence of the A, B, and D genomes, thus implying that the B and D genomes underwent deletions and/or translocations. This succession of event cannot be excluded but appeared less parsimonious. Finally, in barley, two small remnants of the *Unknown gene-2* (140 and 428 bp, respectively), separated by 40 bp, are present in orthologous positions; the longest remnant harbors a large number of point mutations and insertions/deletions, including an internal duplication of 74 bp. These remnants display between 79% and 86% identity to *Unknown gene-2a* and *Unknown gene-2b* of *T. monococcum*. They are 99% identical (one single nucleotide polymorphism [SNP] over 140 bp), indicating that the duplication that generated them was recent, and probably different from that known to have occurred in wheat, as *Unknown gene-2a* and *2b* are only 86% identical.

The long-term evolution of the other genes in this first group is harder to describe. It is not completely conserved in the B genome of polyploid wheat species, for which a large microcolinearity break is observed after the 55 first codons of the *VAMP* gene, and to what extent has yet to be characterized. Moreover, in the absence of a sequence from diploid species closely related to the B genome progenitor, we cannot determine if the series of events leading to such a genomic structure occurred before or after polyploidization. Nevertheless, as *VAMP* is perfectly conserved over a

Fig. 2 Schematic representation of one possible evolutionary scenario for the *Ha* locus in rice, barley, and wheat. The color codes are as in Fig. 1. (**A**) The hypothetical ancestral locus is thought to possess at least one ancestral copy of *Unknown gene-2*, *VAMP*, *BGGP*, *Gsp*, *HIPL*, *Unknown gene-8*, and *ATPase*. (**B**) Possible evolution for rice, involving an inversion of the block composed of *Unknown gene-8* and *ATPase*, tandem duplication of *ATPase*, and pseudogenization of the ancestral copy of *Gsp*. (**C**) Possible evolution for the last common ancestor of wheat and barley, in which *HIPL* was translocated, *CHS* inserted, and the ancestral copy of *Gsp* underwent two successive tandem duplications, leading to *Gsp-1*, *Pina* (*hinda*), and *Pinb* (*hindb*). (**D**) Possible evolution for barley, involving the translocation/inversion of a fragment including *hinda*, *hindb*, *Unknown gene-8* (*HvPG1*), and at least one copy of *ATPase*. (**E**) Possible evolution for the different wheat genomes, involving duplications of *ATPases*, duplications of *Unknown gene-2* in A genomes and *Pinb* in the D genome, deletion or translocation of a fragment from the 5′ end up to *VAMP* in the B genome, and deletions of *Pina* and *Pinb* in tetraploid A and B genomes. a, rice chromosome 12 pseudomolecule from position 27369197 to position 27448138; b, barley BACs AY643842 to AY643844; c, *T. durum* BAC CT009587; d, *T. durum* BAC CT009588; e, *T. aestivum* BAC CT009735



a : rice ch.12 pos. 27369197..27448138
b : barley BACs AY643842 to AY643844
c : *T. durum* BAC CT009587
d : *T. durum* BAC CT009588
e : *T. aestivum* BAC CT009735

large evolutionary timescale (from rice to Triticeae; Fig. 2), we may hypothesize that deletion or translocation occurred *after* polyploidization (Fig. 2E). In the same way, *Gsp-1* is present in rice only as a small sequence, potentially derived from a common ancestral gene of rice and Triticeae species (Chantret et al. 2004).

The second group of genes is composed of the *Unknown gene-8* surrounded by several copies of *ATPase*. The evolution of *ATPase* genes can hardly be reconstructed over the timescale of rice/Triticeae evolution, as many tandem duplications, inversions, and/or deletions occurred over this time frame and as quite a number of copies degenerated. Still, the presence of several copies downstream from the 3′ end of *Unknown gene-8* suggests that at least one copy of *ATPase* and *Unknown gene-8* was present in the common ancestor of rice, barley, and wheat (Fig. 2A). Tandem duplications and/or inversion of *ATPase* occurred in each lineage, but it is tricky to determine their occurrence exactly with respect to lineage separation. In wheat, tandem duplication may have generated at least three copies in the diploid ancestral genome (Fig. 2E). The B genome harbors three copies and an additional segment originating from a recent duplication, involving only fragments of the genes (Chantret et al. 2005). In D genomes, *Unknown gene-8* and three *ATPase* copies seem to have undergone tandem duplication (Chantret et al. 2005). Finally, the A genomes of *T. aestivum* and *T. durum* each contain a fourth copy of *ATPase* (*ATPase7-4*).

Conservation of Genic and Intergenic Regions

The B genomes of *T. durum* and *T. aestivum* display a very high level of sequence identity, differing by only 58 SNPs (1 per 1800 bp on average) and a few small insertions/deletions, accounting for 111 bp of the distal part of the common 102 kb. Only one of these SNPs is located in a coding region (last exon of *BGGP*). For the A genomes, 239 SNPs are present in the distal region of *T. durum* and *T. aestivum* (from the 5′ end of the *T. aestivum* BAC to the 3′ end of *ATPase7-3*). Fifteen of these are found in common functional genes (*Unknown gene-2a*, *VAMP*, *BGGP*, *Gsp-1*, and *Unknown gene-8*)—seven in CDS and eight in introns. Nine SNPs exist in *ATPase7-2* (pseudogenic in *T. durum*), 44 in *ATPase7-3*, and 16 in *Unknown gene2b*, all of them being pseudogenic in *T. aestivum*. Finally, 97 SNPs are found in *ATPase3-1*, pseudogenic in both *T. durum* and *T. aestivum*. There are 58 SNPs in the intergenic space stricto sensu, leading to 1 SNP per 470 bp. If the *ATPase3-1* pseudogene is included in the intergenic space (as it is no longer a coding sequence), this frequency rises up to one SNP per 180 bp. This difference in polymorphism frequency between the A and the B genomes may be due to allele sampling; investigation of a larger sample of *T. durum* and *T. aestivum* accessions should be used to confirm this. Finally, as previously reported, comparisons between the D genomes of *T. aestivum* and *Ae. tauschii* revealed several shuffling events (Chantret et al. 2005); however, the remaining distal region included in this study showed no evidence of additional ones.

Evolution of Coding Sequences in this Region

We applied different models of codon evolution to our data to determine whether the different genes of the same locus were subjected to different selection pressures and to quantify these differences (see Table 2). *CHS*, *VAMP*, and *BGGP* appeared to be strongly constrained, with a mean ω of 0.054, 0.055, and 0.037, respectively, and with an extremely low frequency of more relaxed sites ($p_1 = 0.046$ for $\omega_1 = 1$ for *CHS*). *Unknown gene8* and *Pinb* were found to be slightly less constrained, with a mean ω of 0.100 and 0.268 respectively. *Gsp-1* and *Pina* were clearly the least constrained genes of the locus, with a mean ω of 0.429 and 0.487, respectively. The most likely model (M0) included no sites under positive selection, in contrast to the results of Massa and Morris (2006). This difference is probably due to the small size of our sample, as we used only the available BAC sequences. However, the ω value obtained for *Pina* with the M0 model was similar in our study and that by Massa and Morris (2006).

Evolution of Introns in this Region

We analyzed intron evolution, by focusing on the four largest introns in this region: one in *Unknown gene-8*, two in *VAMP*, and one in *BGGP*. Analyses of intron sequences (their alignments are available in Supplemental Data 2) revealed that (i) the molecular clock is rarely respected in all species, (ii) this violation differs between introns, and (iii) sequence conservation is highly heterogeneous within introns.

Tajima's relative rate tests were performed on different sets of three sequences for the four introns (Table 3). The molecular clock hypothesis was rejected for three introns (*BGGP*-intron3, *VAMP*-intron1, and *VAMP*-intron3). Moreover, the sequences because of which the molecular clock hypothesis was rejected were not the same. For example, in *BGGP*-intron3, the molecular clock hypothesis was rejected for comparisons between A genome sequences and B or D genome sequences, whereas for *VAMP*-intron1 and *VAMP*-intron3, any combination involving the sequence from the D genome of *T. aestivum* and the A genome of *T. monococcum*, respectively, led to rejection of the molecular clock hypothesis (Table 3).

Table 2 Codon substitution models tested for the genes of the *Ha* locus; the model described is the most probable[a] for each gene

Gene	Model[a]	Parameters	log*L*	ω, averaged[b]
CHS	M1	$p_0 = 0.954$; $\omega_0 = 0.008$; $p_1 = 0.046$; $\omega_1 = 1.000$	−1701.70	0.054
VAMP	M0	$p_0 = 1$; $\omega_0 = 0.055$	−1074.64	0.055
BGGP	M7	β parameters = 0.07; 1.42; $p = 0.100$; $\omega_{1-6} = 0.000$; $\omega_{6-10} = 0.001$; 0.008; 0.052; 0.313	−2039.39	0.037
Gsp-1	M0	$p_0 = 1$; $\omega_0 = 0.429$	−1118.54	0.429
Pina	M0	$p_0 = 1$; $\omega_0 = 0.487$	−823.24	0.487
Pinb	M0	$p_0 = 1$; $\omega_0 = 0.268$	−1005.04	0.268
Unknown gene-8	M1	$p_0 = 0.942$; $\omega_0 = 0.045$; $p_1 = 0.059$; $\omega_1 = 1.000$	−3080.93	0.100

[a] Most probable model according to likelihood ratio test (M0 vs. M1, M1 vs. M2, M7 vs. M8), with $p = 0.05$ as the significance threshold, using Akaike's Informative Criterion [AIC = (2×number of parameters) – (2×log*L*)], with 1.5 as a minimum ΔAIC (M1 vs. M7). M0, one category of ω ($0 < \omega_0 < 1$); M1, two categories of ω ($0 < \omega_0 < 1$ and $\omega_1 = 1$); M2, three categories of ω ($0 < \omega_0 < 1$, $\omega_1 = 1$, and $\omega_S > 1$, accounting for sites under positive selection); M7, ω following a β distribution discretized into 10 categories of similar frequency ($0 < \omega_{1-10} < 1$); M8, ω following a β distribution as in M7; plus one category of $\omega_S > 1$, accounting for sites under positive selection

[b] ω obtained by adding the values of ω for each category, balanced by their frequency

Table 3 Tajima's (1993) relative rate test between different pairs of sequences (sequence 1 and sequence 2) with respect to different outgroups, for four introns

Sequence1	Sequence2	Out-group	*BGGP* intron 3			*VAMP* intron 1			*VAMP* intron 3			*UG-8*		
			Length	χ^2	*p*	Length	χ^2	*p*	Length	χ^2	*p*	Length	χ^2	*p*
A[a]	Amono	D[b]	1009	0.20	0.655	351	0.22	0.637	126	26.06	**0.000**	—[c]	—	—
A	Amono	Barley	875	0.00	1.000	289	0.00	1.000	129	0.71	0.398	—	—	—
Daes	Dtau	A	NE[d]	NE	NE	348	4.57	**0.033**[e]	NE	NE	NE	—	—	—
Daes	Dtau	Barley	NE	NE	NE	280	2.27	0.132	NE	NE	NE	—	—	—
A	B	Barley	815	70.72	**0.000**	—	—	—	—	—	—	536	0.00	1.000
Amono	B	Barley	819	70.72	**0.000**	—	—	—	—	—	—	—	—	—
Daes	B	Barley	826	0.73	0.392	—	—	—	—	—	—	456	0.36	0.546
Dtau	B	Barley	826	0.73	0.392	—	—	—	—	—	—	—	—	—
A	Daes	Barley	859	70.47	**0.000**	283	4.74	**0.029**	129	0.61	0.435	603	0.03	0.857
A	Dtau	Barley	860	70.47	**0.000**	276	0.69	0.406	129	0.61	0.435	—	—	—
Amono	Daes	Barley	863	70.47	**0.000**	284	4.74	**0.029**	130	4.57	**0.033**	—	—	—
Amono	Dtau	Barley	864	70.47	**0.000**	277	0.89	0.346	130	4.57	**0.033**	—	—	—

Note. *UG-8, Unknown gene-8*. The length of the common sequence for each triplet (sequence 1, sequence 2, and outgroup) is given as base pairs. *p* is the probability associated with the χ^2 distribution, with 1 degree of freedom, in boldface when significant at the 5% level. Sequences originate from the BACs of *T. monococcum* (Amono), *Ae. tauschii* (Dtau), and the A, B, and D genomes of *T. aestivum* (A, B, and Daes, respectively) and of barley

[a] The A genome sequence from *T. aestivum* BAC was used, as similar results were obtained in terms of significance of the χ^2 test as with the A sequence from *T. durum* BAC, unless otherwise specified

[b] The D genome sequence from *T. aestivum* BAC was used, as similar results were obtained in terms of significance of the χ^2 test as with the D sequence from *Ae. tauschii* BAC

[c] Missing data

[d] Not estimated, as sequence 1 and sequence 2 did not differ

[e] 0.061, with Adur as the outgroup

Global identity was determined as the percentage of conserved sites over the aligned sequences for *BGGP*-intron3 (Fig. 3; the results for the other analyzed introns are shown in Supplemental Data 3). The different curves correspond to different genome combinations extracted from the same multiple alignment. Two large insertions/deletions (488 and 50 bp) decrease global identity to zero when sequences harboring the insertion (A genomes) are aligned with sequences lacking it (B, D, and H genomes), i.e., for the ABDH, AH, and ABD combinations in Fig. 3.

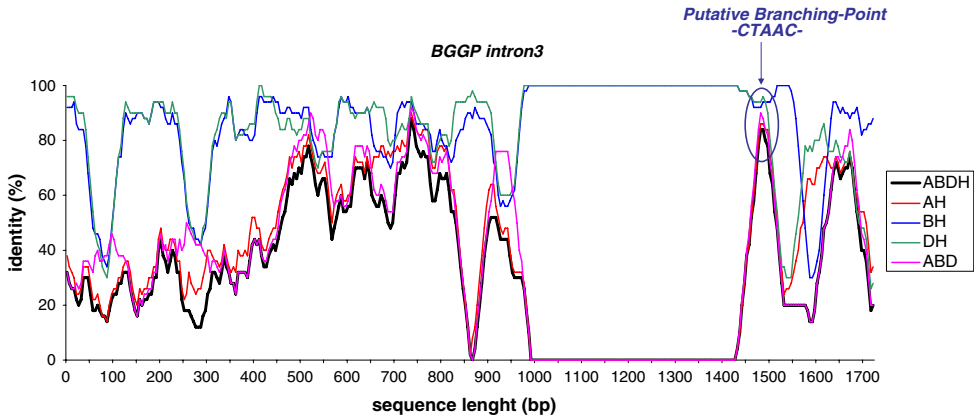


Fig. 3 Percentage of global identity for intron 3 of *BGGP*, based on the complete multiple alignment of wheat A, B, and D genomes and barley sequences. Global identity was calculated using a sliding window of 50 bp, advanced in 5-bp increments, as the percentage of sites conserved in all the aligned sequences. The different curves correspond to different genome subsets; A, A genomes of *T. monococcum*, *T. durum*, and *T. aestivum*; B, B genomes of *T. durum* and *T. aestivum*; D, D genomes of *T. aestivum* and *Ae. tauschii*; H, genome of barley. The putative branching point position is shown on the curves

The multiple alignments display a mosaic pattern of regions with high and low levels of conservation. For the first 500 bp, major differences were observed between curves including the A genomes (AH, ABD and ABDH) and the others (BH and DH). Unexpectedly, the loss of global identity in this region was found to be due to considerable divergence of the A genomes from the other genomes (the three A genomes being 98.8% identical to each other), rather than barley divergence. This result is nevertheless consistent with rejection of the molecular clock hypothesis when A genomes are considered (Table 3). The putative branching point, located at position ~1500 in the complete alignment, may partly account for the high percentage global identity observed between position 1450 and position 1550. This branching point sequence is required for correct elimination of the intron sequence in Lariat structures. In higher eukaryotes, the minimal consensus sequence is generally YTRAY, with Y = C/T and R = A/G, with the T and the A as the principal bases required. Moreover, the consensus must itself be located in a specific "sequence environment" (Lim and Burge 2001) and in an AT-rich sequence: here, the branching point is CTAAC (CTAcC for barley) and is followed by an AT-rich sequence (> 60%; see alignment in Supplemental Data 2). If all introns are considered, the general consensus for the branching point in this region of the Triticeae genome is YTGAT. The 5′- and 3′-splicing sites (SSs) also conform to the classical consensus for higher plants—i.e., GTAMMT and WRCAG versus GTAAGT and TKCAG (with M = A/C, W = A/T, K = G/T), for 5′-SS and 3′-SS in Triticeae (this study) and *Arabidopsis thaliana* (Lim and Burge 2001), respectively.

Discussion

The available data for the *Ha* locus made it possible to infer a possible sequence of events leading to the currently observed highly variable pattern of gene conservation among different species including wheat, rice, and barley (Fig. 2). Microcolinearity conservation appeared to be restricted to two groups of genes. The first group (*Unknown gene-2*, *VAMP*, *BGGP*, and *Gsp-1*) is conserved in wheat and barley species, and is also found in rice, which bears a putative gene relic similar to *Gsp-1* (Chantret et al. 2004). The second group of genes, *Unknown gene-8* surrounded by several copies of *ATPase*, is found in wheat, rice, and barley. *Pina* and *Pinb* can be added to this group in barley and the nondeleted genomes of wheat.

The positions of these two clusters of genes differ between barley and wheat (Caldwell et al. 2004). Probable chromosomal rearrangement in barley may have led to this inversion, as suggested by Caldwell et al. (2004), but we suggest here an alternative sequence of events (Fig. 2). The duplication generating the three paralogous genes *Gsp-1*, *Pina*, and *Pinb* was probably the first event, occurring in the ancestor of wheat and barley. Separation of barley and wheat lineages was followed by a chromosomal rearrangement in barley leading to the translocation and inversion of a fragment containing at least four genes (*hina*; one copy or both copies of *hinb*; *PG1—Unknown gene-8* in wheat; and at least one copy of *ATPase*).

This chromosomal region provides a good illustration of modifications tolerated by the genomes of polyploid species. The wheat B genomes present a clear rupture of microcolinearity after the first bases of the *VAMP* gene. The extent of the region in which the B genome sequence

is not related to that of other Triticeae genomes suggests a large genomic rearrangement, involving either deletion or translocation, but probably including the loss of *VAMP*, as the first bases of its CDS remained truncated. The presence of homeo-alleles provided by other genomes in polyploids may explain the greater tolerance of polyploid plants to genome and/or gene rearrangements (Blanc and Wolfe 2004; Levy and Feldman 2002; Wendel 2000). Microcolinearity conservation thus depends on the considered locus and polyploidy may decrease the chances of finding conserved genes between homoeologous genomes.

This study highlights that several neighboring genes within the same locus evolved *via* rounds of gene duplication and loss and that their CDS were subjected to contrasting selection pressures. The *ATPase* genes constitute a typical example of genes undergoing a "birth and death" process, with rounds of tandem duplication, followed by inactivation because of deletions or accumulation of point mutations: only 8 of the 25 copies (genes, pseudogenes, or remnants) identified are potentially functional. *Unknown gene-2* also provides a good example of duplication followed by a release of selection pressure on one copy in polyploid context. For *Unknown gene-2b*, a large deletion and a total loss of initial function are tolerated in *T. aestivum* and at least a partial loss of function is tolerated in *T. durum*.

The genes present in the *Ha* locus are subject to different selection pressures. *CHS*, *VAMP*, and *BGGP* are the most highly constrained genes, but different models can be retained for these genes, as only *CHS* contains some sites under neutral evolution. Finally, the three paralogous genes *Gsp-1*, *Pina*, and *Pinb* are not subject to the same constraints, with *Pinb* specifically under more intense purifying selection. A recent study revealed the presence of a positive selection signature in the *Pina* sequence, with certain amino acid changes in the corresponding protein resulting from evolution under adaptation, but no such positive selection was detected for *Pinb* (Massa and Morris 2006). We detect no positive selection in *Pina*, probably because our dataset is smaller, but we obtain similar ω values when equivalent models are compared.

Alignments of intron sequences reveal that the molecular clock hypothesis is violated in most cases. Moreover, this violation involves different sequences from different introns, and sequence conservation is highly heterogeneous within introns. Intron size and maintenance are thought to be subject to natural selection, whereas intron sequence conservation (except for the 5′SS, 3′SS, and branching point) is less well understood. Recent studies have reported elsewhere such an unexpected high degree of conservation in intron sequence blocks under partial selective constraints (Bergman and Kreitman 2001; Hare and Palumbi 2003). Our results are consistent with the heterogeneity of sequence conservation, as this heterogeneity is observed in short and long introns. The regions with the lowest sequence conservation are found only in certain genomes, and not necessarily the least related (e.g., the lower level of divergence between barley and a specific wheat subgenome than between two wheat subgenomes for the third intron of *BGGP*; Fig. 3). Additional sequencing will be required to determine whether conservation of particular regions can be explained by selective constraint (over relatively short phylogenic distances) and whether a higher than expected mutation rate results in regions with a lower level of conservation.

The *Ha* locus is located in a very distal position on the short arm of group 5 homoeologous chromosomes. In general, colinearity conservation is supposed to be stronger in proximal than in distal chromosomic regions (Akhunov et al. 2003a). The fragmental microcolinearity conservation observed here is consistent with this hypothesis. Recombination rate may affect microcolinearity conservation, as it is known to be higher in distal than in proximal regions in wheat chromosomes (Akhunov et al. 2003b). Regions of the genome with high rates of recombination are supposed to have an intrinsically high rate of sequence divergence (Perry and Ashworth 1999). However, according to Yi et al. (2004), recombination may have a weaker mutagenic effect than previously thought.

Concluding Remarks

Using the large number of homoeologous sequences available for the *Ha* locus, we showed that microcolinearity is restricted to a small part of the gene content, even on a short evolutionary time. This study also suggests a possible evolutionary scenario for this locus in these species, involving rounds of duplication/deletions and release of pressure because of polyploidy. Moreover, our study of this highly plastic region shows that the conserved genes (exon as well as intron sequences) in a short chromosomic region are not subjected to homogeneous selective pressure, and that sequence conservation is not necessarily linked to the species relationship.

Acknowledgments We sincerely thank Evans Lagudah for supplying the *Ae. tauschii* 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 with development of the annotation platform at Institut National de la Recherche Agronomique (Clermont-Ferrand, France), and Frank Samson for BAC annotation visualization-tool development. We also thank Alberto Cenci for his critical comments on the manuscript. Finally, we sincerely thank Stephane De Mita for his help with the development of Python scripts.

References

Akhunov ED, Akhunova AR, Linkiewicz AM, Dubcovsky J, Hummel D, Lazo G, Chao S, Anderson OD, David J, Qi L, Echalier B, Gill BS, Miftahudin, Gustafson JP, La Rota M, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, Lapitan NL, Wennerlind EJ, Nduati V, Anderson JA, Sidhu D, Gill KS, McGuire PE, Qualset CO, Dvorak J (2003a) Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc Natl Acad Sci USA 100:10836–10841

Akhunov ED, Goodyear AW, Geng S, Qi LL, Echalier B, Gill BS, Miftahudin, Gustafson JP, Lazo G, Chao S, Anderson OD, Linkiewicz AM, Dubcovsky J, Rota ML, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, Lapitan NL, Gonzalez-Hernandez JL, Anderson JA, Choi DW, Close TJ, Dilbirligi M, Gill KS, Walker-Simmons MK, Steber C, McGuire PE, Qualset CO, Dvorak J (2003b) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res 13:753–763

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12:1021–1029

Bennetzen JL, Ramakrishna W (2002) Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. Plant Mol Biol 48:821–827

Bergman CM, Kreitman M (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Res 11:1335–1345

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16:1679–1691

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

Caldwell KS, Langridge P, 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 OD, Fahima T, Distelfeld A, 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, Dubcovsky J (2004) Sequencing of the *Triticum monococcum Hardness* locus reveals good microcolinearity with rice. Mol Genet Genomics 271:377–386

Chantret N, Salse J, Sabot F, Rahman S, Bellec A, Laubin B, Dubois I, Dossat C, Sourdille P, Joudrier P, Gautier MF, Cattolico L, Beckert M, Aubourg S, Weissenbach J, Caboche M, Bernard M, Leroy P, Chalhoub B (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). Plant Cell 17:1033–1045

Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. Plant Cell 12:637–646

Feldman M, Lupton FGH, Miller TE (1995) Wheats. In: Smartt J, Simmonds N (eds) Evolution of crops, 2nd ed. Longman Scientific, London, pp 184–192

Feuillet C, Keller B (2002) Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. Ann Bot 89:3–10

Gale MD, Devos KM (1998) Comparative genetics in the grasses. Proc Natl Acad Sci USA 95:1971–1974

Gautier MF, Aleman ME, Guirao A, Marion D, 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 MF, Cosson P, Guirao A, Alary R, 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

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Gu YQ, Coleman-Derr D, Kong X, Anderson OD (2004) Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four triticeae genomes. Plant Physiol 135:459–470

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Hare MP, Palumbi SR (2003) High intron sequence conservation across three mammalian orders suggests functional constraints. Mol Biol Evol 20:969–978

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Ilic K, SanMiguel PJ, Bennetzen JL (2003) A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc Natl Acad Sci USA 100:12265–12270

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

Kellogg EA (1998) Relationships of cereal crops and other grasses. Proc Natl Acad Sci USA 95:2005–2010

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 17:1244–1245

Levy AA, Feldman M (2002) The impact of polyploidy on grass genome evolution. Plant Physiol 130:1587–1593

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

Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci USA 98:11193–11198

Massa AN, Morris CF (2006) Molecular evolution of the puroindoline-a, puroindoline-b, and grain softness protein-1 genes in the tribe Triticeae. J Mol Evol 63:526–536

Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal genome evolution. Grasses, line up and form a circle. Curr Biol 5:737–739

Moullet O, Zhang HB, Lagudah ES (1999) Construction and characterisation of a large DNA insert library from the D genome of wheat. Theor Appl Genet 99:305–313

Nesbitt M, Samuel D (1995) From staple crop to extinction? The archeology and history of the hulled wheats. In: Proceedings of the First International Workshop on Hulled Wheats. Castelvecchio Pascoli, Italy, 21–22 July, pp 40–99

Nevo E, Korol AB, Beiles A, Fahima T (2002) Origin and evolution of wheat. In: Evolution of wild emmer and wheat improvement. Springer-Verlag, Berlin, pp 1–22

Perry J, Ashworth A (1999) Evolutionary rate of a gene affected by chromosomal position. Curr Biol 9:987–989

Prasad V, Stromberg CA, Alimohammadian H, Sahni A (2005) Dinosaur coprolites and the early evolution of grasses and grazers. Science 310:1177–1180

Rahman S, Jolly JC, Skerritt JH, 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, SanMiguel P, Ogden M, Llaca V, Messing J, Bennetzen JL (2002) Comparative sequence analysis of the sorghum *Rph* region and the maize *Rp1* resistance gene complex. Plant Physiol 130:1728–1738

Sabot F, Guyot R, Wicker T, Chantret N, Laubin B, Chalhoub B, Leroy P, Sourdille P, Bernard M (2005) Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. Mol Genet Genomics 274:119–130

Song R, Llaca V, Messing J (2002) Mosaic organization of orthologous sequences in grass genomes. Genome Res 12:1549–1555

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol 42:225–249

Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu ZD, Dubcovsky J, 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

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yi S, Summers TJ, Pearson NM, Li WH (2004) Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. Genome Res 14:37–43