Sixty Million Years in Evolution of Soft Grain Trait in Grasses: Emergence of the Softness Locus in the Common Ancestor of *Pooideae* and *Ehrhartoideae*, after their Divergence from *Panicoideae*

Mathieu Charles, Haibao Tang,† Harry Belcram,* Andrew Paterson,† Piotr Gornicki,‡ and Boulos Chalhoub**

*Unité de Recherches en Génomique Végétale (UMR INRA 1165–CNRS 8114UEVE), Organization and evolution of Plant Genomes, Evry, France; †Plant Genome Mapping Laboratory, University of Georgia; and ‡Department of Molecular Genetics and Cell Biology, University of Chicago

Together maize, Sorghum, rice, and wheat grass (*Poaceae*) species are the most important cereal crops in the world and exhibit different "grain endosperm texture." This trait has been studied extensively in wheat because of its pivotal role in determining quality of products obtained from wheat grain. Grain softness protein-1 and Puroindolines A and B (grain storage proteins), encoded by *Ha-like* genes: *Gsp-1*, *Pina*, and *Pinb*, of the *Hardness* (*Ha*) locus, are the main determinants of the grain softness/hardness trait in wheat. The origin and evolution of grain endosperm texture in grasses was addressed by comparing genomic sequences of the *Ha* orthologous region of wheat, *Brachypodium*, rice, and Sorghum. Results show that the *Ha-like* genes are present in wheat and *Brachypodium* but are absent from *Sorghum bicolor*. A truncated remnant of an *Ha-like* gene is present in rice. Synteny analysis of the genomes of these grass species shows that only one of the paralogous *Ha* regions, created 70 My by whole-genome duplication, contained *Ha-like* genes. The comparative genome analysis and evolutionary comparison with genes encoding grain reserve proteins of grasses suggest that an ancestral *Ha-like* gene emerged, as a new member of the prolamin gene family, in a common ancestor of the *Pooideae* (*Triticeae* and *Brachypoidieae* tribes) and *Ehrhartoideae* (rice), between 60 and 50 My, after their divergence from *Panicoideae* (Sorghum). It was subsequently lost in *Ehrhartoideae*. Recurring duplications, deletions, and/or truncations occurred independently and appear to characterize *Ha-like* gene evolution in the grass species. The *Ha-like* genes gained a new function in *Triticeae*, such as wheat, underlying the soft grain phenotype. Loss of these genes in some wheat species leads, in turn, to hard endosperm seeds.

Introduction

Grasses (*Poaceae*), with 10,000 species growing under diverse climates and latitudes, exceed all other plant families in ecological dominance and economic importance. Analysis of fossil records and phylogenetic data established that the grass subfamilies diverged from a common ancestor 50–80 My (for review, see Kellogg 2001; Gaut 2002; Prasad et al. 2005; Chalupska et al. 2008). Divergence time of several important grass lineages (*Triticum*, *Hordeum*, *Brachypodium*, *Oryza*, *Sorghum*, and *Zea*) has been recently reexamined based on sequence comparison of *Acc* and other genes, using 60 My for the divergence time of the *Panicoideae* (*Sorghum*, *Zea*) and *Ehrhartoideae* (rice) to calibrate the molecular clock (Chalupska et al. 2008). This and several earlier studies (Paterson et al. 2004; Bossolini et al. 2007; Faris et al. 2008) concluded that *Pooideae* (wheat, barley, and *Brachypodium*) and *Ehrhartoideae* (rice) diverged from each other 50 My, early after their divergence from *Panicoideae* (maize, Sorghum). Among the *Pooideae*, *Brachypoidieae* (*Brachypodium*), and *Triticeae* (wheat, barley), tribes diverged about 35 My. *Brachypodium*, with its small diploid genome, has become a model *Pooideae* grass with a potential to aid analysis of the large genomes of the *Triticeae* (Draper et al. 2001; Foote et al. 2004; Faris et al. 2008). A 4× draft version of the *Brachypodium distachyon* genomic sequence is already publicly available (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008), and the *Brachypodium* consortium is assembling an 8× genome sequence coverage (Vogel J, personal communication).

The *Hardness* (*Ha*) locus in wheat spans *Pina*, *Pinb*, and *Gsp-1* genes (called in this study *Ha-like* genes) and encodes Puroindolines A and B (PinA and PinB) and grain softness protein-1 (GSP-1) that determine the wheat grain hardness/softness or endosperm texture (for review, see Morris 2002). Because of the pivotal role of grain texture in determining quality of products obtained from wheat grain, this trait has been studied by geneticists (Law et al. 1978), chemists (Schofield 1986; Blochet et al. 1991, 1993), and molecular biologists (Gautier et al. 1994, 2000; Chantret et al. 2004, 2005, 2008; Li et al. 2008). At the genome organization level, the *Ha* locus is about 65 kb in the D genome of hexaploid wheat *Triticum aestivum* and contains three functional *Ha-like* genes: *Gsp-1*, *Pina*, and *Pinb*, as well as a *PseudoPinb*, a *Pinb-relic*, two other predicted genes (*Gene3* and *Gene5*), and several transposable elements (Chantret et al. 2005; fig. 1*A*). Upstream of *Gsp-1* gene, the *BGGP* (*Gene1*), encoding β-1-3-galactosyl-O-glycosyl-glycoprotein, delimits the 5′ boundary of the *Ha* locus. A *Nodulin* gene (*Gene8*) and a cluster of *ATPase* genes (*Genes7-1, 7-2, 7-3, 7-2′, and 7-3′*), located 20 kb downstream of *PseudoPinb*, delimit the 3′ boundary of the *Ha* locus (fig. 1*A*; Chantret et al. 2005, 2008). *Pina* and *Pinb* genes were also found in *Triticeae* species, in which soft endosperm is a dominant trait: in diploid and hexaploid wheat (*Triticum* and *Aegilops* species), barley (*Hordeum vulgare*), rye (*Secale cereale*), and oats (*Avena sativa*). Surprisingly, *Pina* and *Pinb* genes are absent from the A and B genomes of the tetraploid (*Triticum turgidum*) and hexaploid (*T. aestivum*) wheat species, although present in their progenitor species (Gautier et al. 2000). Comparative genomic analysis showed that *Pina*

Key words: *Poaceae*, evolution, comparative genomics, grain endosperm softness.

E-mail: chalhoub@evry.inra.fr.

Mol. Biol. Evol. 26(7):1651–1661. 2009
doi:10.1093/molbev/msp076
Advance Access publication April 24, 2009

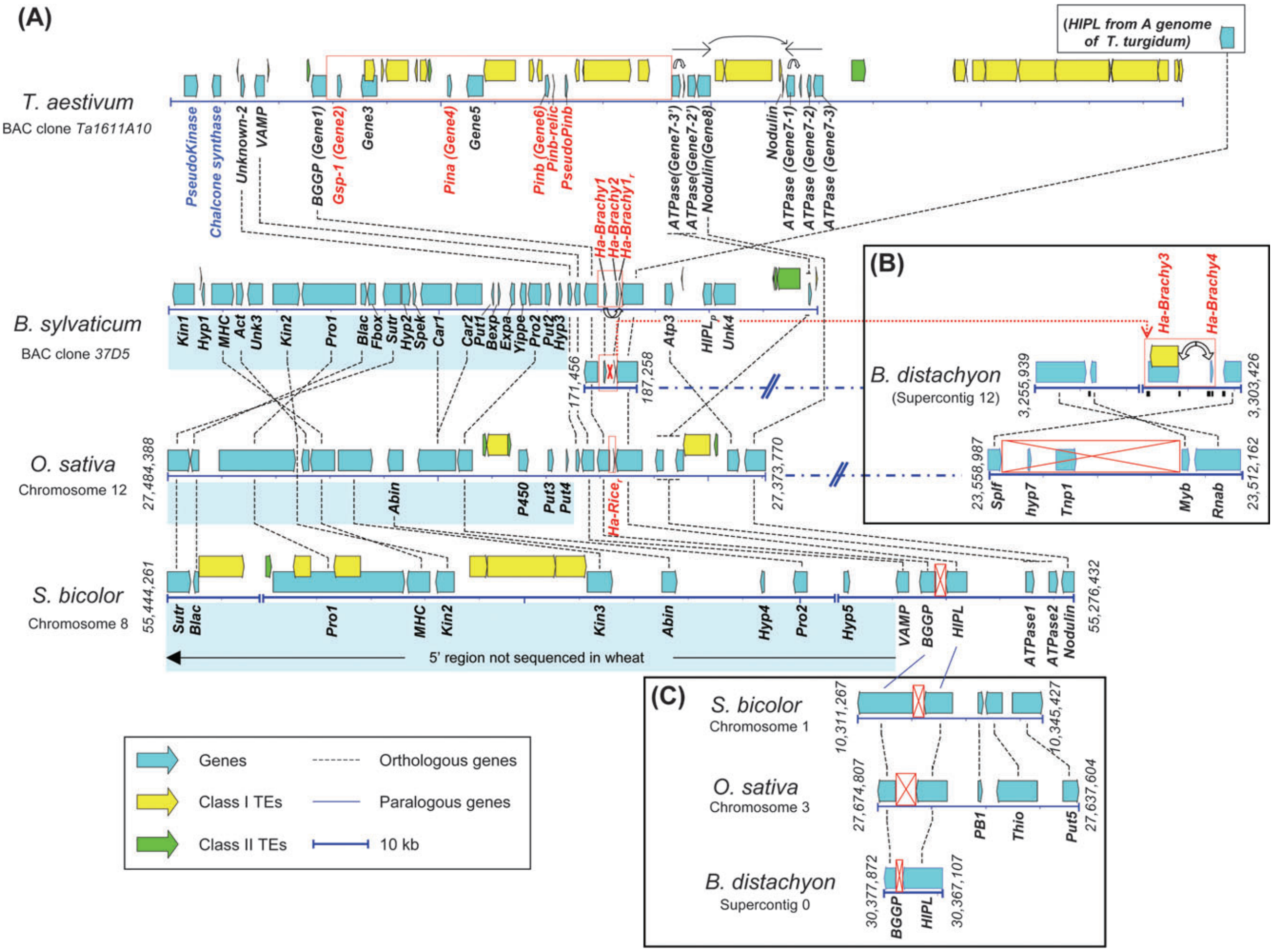

FIG. 1.—Comparison of orthologous and paralogous genomic regions including the *Ha* locus of wheat, *Brachypodium sylvaticum*, *Brachypodium distachyon*, rice, and *Sorghum bicolor*. (*A*) Comparison of orthologous regions between the five species. An overview of the 187,340 bp sequence (BAC clone Ta1611A10) of the D genome of hexaploid wheat *Triticum aestivum* (from Chantret et al. 2008). Relative position of *HIPL* gene as found in the sequence of the A genome of *Triticum turgidum* species (Chantret et al. 2008) is also shown. *Brachypodium sylvaticum* BAC clone (BAC37D5) of 120,033 bp was sequenced in this study. *Oryza sativa* and *S. bicolor* orthologous region sequences and annotations were, respectively, retrieved from the Michigan State University site (http://rice.plantbiology.msu.edu, release 6 January 2009) and from the Joint Genome Institute Web site (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html, release March 2008; Paterson et al. 2009). The *B. distachyon* 4X genome sequence is from http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/ (release October 2008). Sequence gaps at *Ha-like* genes were supplied by the international *Brachypodium* initiative (Vogel J, USDA, Albany, USA). The wheat genes were named in the same way as in previous studies (Chantret et al. 2005, 2008). *Ha-like* genes and related sequences are shown in red. *Ha-Brachy2* gene is deleted in *B. distachyon* (marked by a red cross). Genes present in wheat but not in other species are shown in blue. Genes conserved in multiple species are connected by dashed lines. Light blue boxes represent region, 5′ to the *Ha* locus, compared between *B. sylvaticum*, rice, and Sorghum but not sequenced in wheat. (*B*) Additional duplications of *Ha-like* genes (*Ha-Brachy3* and *Ha-Brachy4*) observed in *B. distachyon* at 3 Mb of the *Ha* locus (red dashed arrow). Flanked genes are also shown. Corresponding rice orthologous region is also presented and shows no *Ha-like* genes (red cross). Double arrow on *Ha-Brachy3* and *Ha-Brachy4* indicates that they are derived from recent tandem duplication (from each other's). Black bars, below the *B. distachyon* presented region, indicate location of sequences successfully used to derive PCR markers and confirm the presence of *Ha-Brachy3*, *Ha-Brachy4* as well as flanked *Myb* and *Splf* genes on same BAC clones of *B. sylvaticum*. Thickness of these bars is proportional to the length of the sequence used. (*C*) Synteny and collinearity of paralogous *Ha* regions (derived from last-shared ancestral whole-genome duplication) of rice, Sorghum, and *B. distachyon*. The predicted location of the *Ha-like* genes is between *BGGP* and *HIPL* genes. Absence of any *Ha-like* genes or related sequences is indicated by a red cross. Abbreviations of predicted gene names are detailed in supplementary figure 2 (Supplementary Material online). Nucleotide positions of analyzed regions of the *B. distachyon*, rice, and Sorghum are indicated.

and *Pinb* genes were deleted from the A and B genomes of polyploid wheat species (Chantret et al. 2005). A large deletion at this locus occurred independently not only in the A and B genomes but also in the G genome of another wheat allotetraploid (*Triticum timopheevii*; Li et al. 2008).

Homologs of the *Ha-like* genes have not been found in *Panicoideae* (maize and Sorghum) and *Ehrhartoideae* (rice), all with hard endosperm (Fabijanski et al. 1988; Gautier et al. 2000; Darlington et al. 2001; Morris 2002). Nevertheless, comparative genome analysis shows that a short genomic sequence of 105 bp, with 67% amino acids similarity to *Gsp-1* gene, is present in an otherwise orthologous rice locus (called *Ha-rice-relic*; Caldwell et al. 2004; Chantret et al. 2004, 2005). *Ha-rice-relic* is located between *BGGP* gene, orthologous to wheat *BGGP* and a gene called *HIPL*, encoding a Hedgehog-interacting–like protein, followed by a *Nodulin* gene and a cluster of *ATPase* genes (Chantret et al. 2004, 2008; fig. 1*A*). The situation is not clear for the *Panicoideae* (Sorghum, maize), which diverged earlier from *Pooideae* (wheat, barley) and *Ehrhartoideae* (rice).

In the present study, we used comparative genome analysis of orthologous *Ha* regions from wheat, *Brachypodium*, rice, and recently sequenced *Sorghum bicolor* (Paterson et al. 2009) to analyze the evolutionary origin and trace the relative time of emergence of *Ha-like* genes in grasses.

Materials and Methods

Brachypodium sylvaticum Bacterial Artificial Chromosome Library Screening

A six genome coverage bacterial artificial chromosome (BAC) library of *B. sylvaticum* (Foote et al. 2004) arrayed on high density filters was initially screened with probes prepared from separate, as well as mixture of, polymerase chain reaction (PCR) products, amplified from the hexaploid wheat *Gsp-1*, *Pina*, and *Pinb* genes using primers described in Chantret et al. (2005). Eighteen BAC clones were initially detected, indicating that homologs of these three genes are probably present in *Brachypodium*. Six of these BAC clones were retained in a second step after screening based on hybridization signal intensity, fingerprinting, and PCR confirmations. BAC clone (BAC37D5) was sequenced as described by Chantret et al. (2005).

Another round of PCR screening was also made to check the presence in *B. sylvaticum* of additional *Ha-like* gene duplicates, revealed from the analysis of the *B. distachyon* 4× genome sequence (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008; see Results). Primers were designed based on *B. distachyon* genome sequence, and the *B. sylvaticum* BAC library, organized into pools, was PCR screened.

Ha Genomic Regions from Grass Species Sequenced Genomes

Ha region from the *B. distachyon* was extracted from the available 4× coverage genome sequence (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/; release October 2008), that of rice from http://rice.plantbiology.msu.edu (release 6 January 2009), and that of *S. bicolor* from http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html (release March 2008; Paterson et al. 2009).

Sequence Annotation

Genomic sequences were annotated as described by Chantret et al. (2005). The first step of our annotation method is to detect transposable elements (TEs). Primarily, TEs were detected by a BlastN search against two databases of repetitive elements: TREP (Wicker et al. 2002, http://wheat.pw.usda.gov/ITMI/Repeats/index.shtml) and Repbase (Jurka 2000, http://www.girinst.org/Repbase_Update.html). Core domains (nucleic coordinates of known elements) were identified through BlastN alignments against TREPn. Long terminal repeats (LTRs) and limits were identified through BlastN and CENSOR (Jurka et al. 1996) alignments against Repbase and TREP databases. Putative polyproteins were identified by BlastX alignments against TREPprot. No a priori cutoff was imposed for BlastX and BlastN. We also used structural detection method using LTR_STRUC (McCarthy and McDonald 2003) and DOTTER program (Sonnhammer and Durbin 1995) for de novo identification of TEs. TE prediction and classification were performed as essentially suggested by the unified classification system for eukaryotic TEs, based on the 80–80–80 rule (Wicker et al. 2007). Retrotransposon insertion dates were estimated when necessary based on their LTR divergence as described (Charles et al. 2008).

The next step is the gene annotation. We used the gene prediction given by the program FGENESH (http://www.softberry.com; with the Monocot matrix) as well as BlastN and BlastX and TBlastX alignments against dbEST (http://www.ncbi.nlm.nih.gov/), SwissProt (http://expasy.org/sprot/), and synteny with characterized rice gene to precise gene structure and potential functions.

Finally, we systematically proceeded to a comparative annotation of genes common to several species, checking the coding sequence, and introns/exons transitions.

Gene Classification

Genes of known and unknown functions or putative genes were defined based on FGENESH predictions and the existence of rice or other *Triticeae* homologs. Hypothetical genes were identified based on FGENESH prediction only. Pseudogenes were not well predicted by FGENESH program, and frameshifts need to be introduced within the coding sequences (CDS) structure to better fit a putative function based on BlastX (mainly with rice). Large part of genes, truncated at one end (by TE insertion or unassigned DNA), potentially conversing coding capacity were qualified as “truncated.” Truncated pseudogenes (genes disrupted by large insertion or deletion) and highly degenerated CDS sequences were considered as gene relics.

Identification of Duplicated Paralogous Regions in *B. distachyon*, *Oryza*, and *Sorghum*

We used the gene annotation of *Oryza sativa* (http://rice.plantbiology.msu.edu, MSU rice genome annotation release 6 January 2009) and *S. bicolor* (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html, Sbi version 1.4, release March 2008, Paterson et al. 2009), along with analysis of the *B. distachyon* 4× genome sequence coverage (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008) to identify the syntenic blocks both within and among the three genomes. Identified syntenic blocks from *B. distachyon* were also annotated in the study (FGENESH predictions and Blast against the National Center for Biotechnology Information nonredundant databases). BlastP results ($E < 1 \times 10^{-5}$) among the predicted genes were used as input to feed the collinearity detection program MCscan with the default parameters (score >300, $E < 0.01$; Tang, Bowers, et al. 2008). MCscan generates a number of syntenic blocks, among which we selected the set of regions that are collinear to the identified *Ha* region.

Nucleotide and Protein (amino acid) Sequence Comparisons

We used MEGA3 (Kumar et al. 2004) to make all the nucleic/proteic multiple alignments. We manually enhance these alignments taking into account special feature conservations (such as cysteine skeleton and tryptophan-rich domain [TRD]) or other domain conservation. The pairwise similarity comparisons are based on multiple alignments.

Results

Sequence analysis of the *Ha* locus in the D genome of hexaploid wheat *T. aestivum* and the orthologous region of rice were previously described (Chantret et al. 2005, 2008). We isolated and sequenced in the present study the *Ha* orthologous region from *B. sylvaticum* and conduct comparative genome analysis between all three grass species, along with that from the *B. distachyon* genome (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008) as well as the recently sequenced *S. bicolor* genome (Paterson et al. 2009; fig. 1).

Isolation and Sequencing of the *Ha* Locus in *B. sylvaticum*

Six BAC clones were retained after screening of a six genome coverage BAC library of *B. sylvaticum* (Foote et al. 2004) arrayed on high density filters, using probes prepared from wheat *Gsp-1*, *Pina*, and *Pinb* genes and further characterization, based on hybridization signal intensity, fingerprinting, and PCR. The longest BAC clone (BAC37D5) of 120,033 bp was sequenced.

Two *Ha-like* genes, *Ha-Brachy1* and *Ha-Brachy2*, were found in a 120-kb fragment of the *B. sylvaticum* genome, flanked by a *BGGP* gene on one side and by an *HIPL* and an *ATPase* gene on the other (fig. 1*A*). The tandemly duplicated *Ha-Brachy1* and *Ha-Brachy2* genes contain a single exon each and show 62% amino acid similarity to each other (fig. 2; supplementary table 1, Supplementary Material online). Predicted products of these two genes show 48–54% sequence similarity to wheat GSP-1, PinA, and PinB proteins throughout their entire length (151 and 146 amino acids; fig. 2; supplementary table 1, Supplementary Material online), indicating that an *Ha-like* gene was present in a common ancestor of the *Triticeae* and the *Brachypoidieae* tribes.

Comparison with *Ha* Locus Region from *B. distachyon*

A sequence similarity search on the available 4× shotgun sequences of the *B. distachyon* genome (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008), completed with additional 751 bp sequence, kindly provided by Dr John Vogel (USDA, Albany, USA) to fill the sequence gap, identified only the *Ha-Brachy1* gene and the *Ha-Brachy1-relic* at the *Ha* locus region of this species. The *B. distachyon Ha* locus is located on a region (super_12 contig), which is orthologous to rice chromosome12 and Sorghum chromosome8 (fig. 1*A*). The *Ha-Brachy1* gene and the *Ha-Brachy1-relic* are both very close to their *B. sylvaticum* counterparts, showing 97% and 93% amino acid similarity, respectively (fig. 2; supplementary table 1, Supplementary Material online). Thus, *Ha-Brachy1-relic* was present prior to the two *Brachypodium* species divergence, estimated to 4.2 ± 0.78 My in this study (data not shown). Surprisingly, the *Ha-Brachy2* gene is absent from the *Ha* locus region of *B. distachyon* (fig. 1*A*). The relatively old tandem duplication of *Ha-Brachy1* and *Ha-Brachy2* genes, indicated by the low level of observed amino acid similarities in *B. sylvaticum* (fig. 2; supplementary table 1, Supplementary Material online), from one side, and precise sequence comparisons between the two *Brachypodium* species from the other side, suggest that *Ha-Brachy2* has been deleted from *B. distachyon*. This occurred apparently by an illegitimate DNA recombination, driven by 62–65 bp direct repeats that flank the 842 bp deleted segment (data not shown).

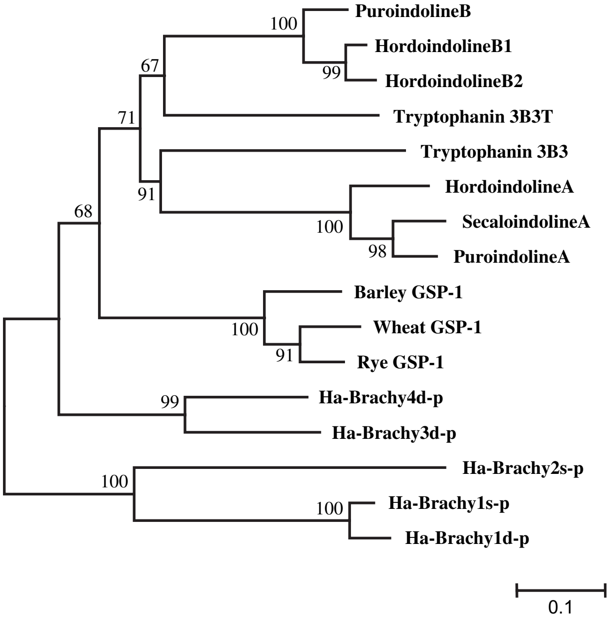

FIG. 2.—Neighbor-Joining tree, illustrating relationships between Ha-like proteins of the *Pooideae* family. Amino acid sequence alignment is shown in supplementary figure 1 (Supplementary Material online). The Ha-like proteins from *Brachypodium distachyon* are ended by (d), those of *Brachypodium sylvaticum* by (s).
Protein reference sequences:

Wheat GSP-1 (CAH10195.1), PuroindolineA (CAH10197.1), and PuroindolineB (CAH10199.1) from Chantret et al. (2005).
Rye GSP-1 (AAT76525.1) from Simeone and Lafiandra (2005).
SecaloindolineA (ABB88759.1) from Massa and Morris (2006).
Barley GSP-1 (AAV49992.1), hordoindolineA (AAV49987.1), hordoindolineB1 (AAV49986.1), and hordoindolineB2 (AAV49985.1) from Caldwell et al. (2004).
Tryptophanin 3B3 (ABU39829.1) and 3B3T (ABU39832.1) from Tanchak et al. (1998).
Ha-like proteins from *B. distachyon* from http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/ (release October 2008).
Ha-like proteins from *B. sylvaticum* were determined in this study.

Additional Duplications of *Ha-Like* Gene in the *Brachypoidieae*

Blast similarity searches of *Ha-like* genes against *B. distachyon* genome sequence (http://www.modelcrop.org/cgi-bin/gbrowse/brachy4x/, release October 2008) allow identification of two other *Ha-like* genes that we called *Ha-Brachy3* and *Ha-Brachy4*. They are located on the same region (super_12 contig), separated by approximately 3 Mb (3,015,111 bp) from *Ha-Brachy1* gene of the *Ha* locus (fig. 1*B*). *Ha-Brachy3* and *Ha-Brachy4* genes are separated by 5.8 kb and show 83% amino acid similarity (fig. 2; supplementary table 1, Supplementary Material online) and 81% nucleotide sequence identity, indicating that they are more likely derived from recent tandem duplication between each other. *Ha-Brachy3* is inserted in this turn by an LTR retrotransposon for which we estimate insertion date to 1.2 ± 0.36 My. These additional *Ha-Brachy* gene copies show between 50% and 60% amino acid similarity to the other *Ha-like* genes (fig. 2; supplementary table 1, Supplementary Material online).

PCR-derived markers (fig. 1*B*) and BAC library screening confirm the presence of both *Ha-Brachy3 and Ha-Brachy4* as well as flanked *Myb* and *Splf* genes (fig. 1*B*) on common BAC clones of *B. sylvaticum* (data not shown). As expected, PCR analysis confirms that the retrotransposon insertion in the *Ha-Brachy3* gene of *B. distachyon* (fig. 1*B*) is not common to that of *B. sylvaticum* (data not shown). Thus, the *Ha-Brachy3* gene is not interrupted in *B. sylvaticum*.

Comparison of *B. distachyon Ha-Brachy3* and *Ha-Brachy4* genomic region (super_12 contig) with corresponding orthologous regions from rice chromosome12 (fig. 1*B*) and Sorghum chromosome8 (data not shown), identified based on flanking conserved genes, did not show any traces of *Ha-like* genes in these two later grass species. These comparisons suggest that *Ha-Brachy3* and *Ha-Brachy4* genes were generated in the *Brachypoidieae* through duplication from an *Ha-like* gene of the *Ha* locus, after divergence from *Ehrhartoideae* (rice).

The situation is not clear for *Triticeae* (wheat and barley) as their genomes are not entirely sequenced yet. Nevertheless, no *Ha-like* genes, other than *puroindolines* or *Gsp-1* genes, were so far described in these *Triticeae* species (reviewed by Morris 2002). Moreover, physical characterization of BAC clones from these species, identified as harboring *Ha-like* genes, revealed one single *Ha-like* region (Caldwell et al. 2004; Chantret et al. 2004, 2005). Further characterizations would better confirm whether this additional *Ha-like* gene duplication is specific to *Brachypoidieae*.

Thus, recurring gene duplications and/or deletions occurred independently at different stages of the grass species evolution, as indicated by the number of *Ha-like* gene copies as well as related gene fragments (partially deleted or incompletely duplicated genes) and pseudogenes found in the *Triticeae* and *Brachypoidieae Ha* locus (fig. 1*A* and *B*; supplementary table 1, Supplementary Material online; Caldwell et al. 2004; Chantret et al. 2005, 2008; discussed also hereafter).

The Orthologous *Ha* Locus Region in *S. bicolor*

A 168-kb fragment of *S. bicolor* genome (coordinates 55,276,432–55,444,261 on chromosome8; Paterson et al. 2009) was identified as containing a region orthologous to that spanning the *Ha* locus sequenced from *B. sylvaticum* (fig. 1*A*). The arguments supporting the orthologous relationship are presented hereafter. DNA sequence of this fragment is available in three contigs and includes 18 genes and putative genes (33% of the sequence), class I TEs (23%), and class II TEs (0.6%). All three numbers are substantially lower than the corresponding genome-wide averages (Paterson et al. 2009).

We found no evidence of any *Ha-like* genes or their relics, such as those found in *Pooideae* and rice, in the *Ha* orthologous region or anywhere in the sequenced Sorghum genome.

Collinearity of the Orthologous *Ha* Region in Wheat and Three Other Grasses

We compared the gene order of the 187-kb region including the D genome *Ha* locus of hexaploid wheat and amino acid sequences they encode to those of the orthologous region of rice, Sorghum, and *B. sylvaticum* (fig. 1*A*). The wheat region is larger because of the expansion of repetitive elements (fig. 1*A*)—it has a higher TE content (47%) and a lower gene content (12%). The corresponding orthologous regions of the other three species are of similar sizes and have comparable gene content (40%). Nevertheless, gene content is higher in *B. sylvaticum* than in sorghum when we extend comparison to the entire sequenced region (detailed hereafter). We detected fewer TEs in *Brachypodium* than in rice (and wheat). Although some *Brachypodium* TEs may have escaped detection because a comprehensive library of TE sequences for this species is not yet available, there is limited remaining space to detect an important proportion of TEs because of the high gene content. However, wheat, rice, and Sorghum also contain fewer TEs in this region than predicted from the genome-wide averages (Charles et al. 2008; Charles H, unpublished data).

Six single-copy genes and a cluster of *ATPase* genes are found in at least three of the four species (fig. 1*A*). The *HIPL* gene is not present in the sequenced fragment of the D genome of hexaploid wheat but instead we used the *HIPL* gene from the *Ha* region of the A genome of *T. turgidum* (fig. 1*A*; Chantret et al. 2008) for comparisons. Different levels of conservation at the amino acid level are observed for the genes when the four species are considered (fig. 3; supplementary table 2, Supplementary Material online). In Sorghum, we have not found any sequences related to the gene *Unknown-2* (fig. 1*A*).

The level of amino acid sequence similarity is consistent with closer evolutionary relationship between *Brachypodium* and wheat (*Triticeae*) than between these two species, rice and Sorghum (fig. 3; supplementary table 2, Supplementary Material online), with the exception of the *ATPase* genes. These genes are often found in clusters of complete and truncated genes, as well as pseudogenes

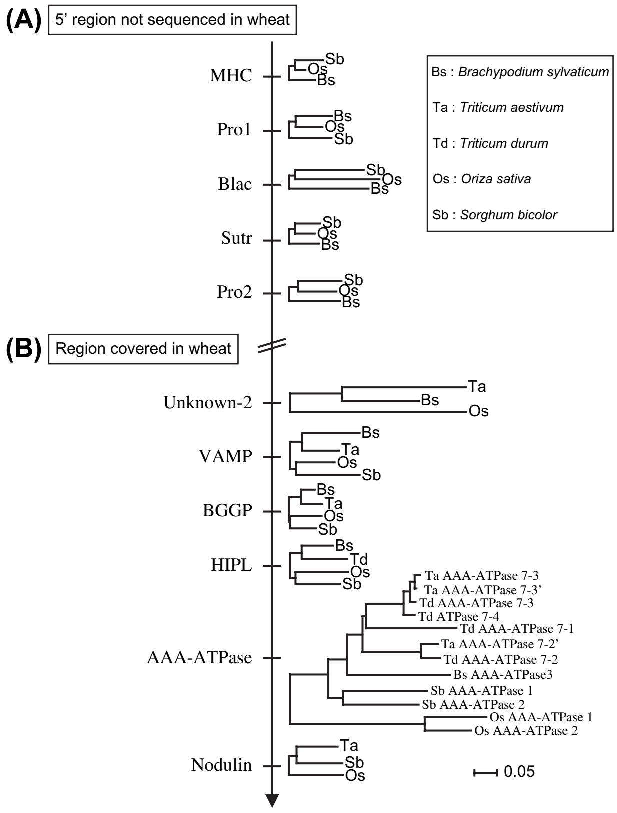

FIG. 3.—Amino acid sequence comparisons of proteins encoded by the predicted genes at the *Ha* locus regions of wheat, *Brachypodium sylvaticum*, rice, and Sorghum, shown in the order, the genes are found in *B. sylvaticum*. (*A*) Genes not present in the sequenced region of the wheat genome. (*B*) Genes present in the sequenced region of the wheat genome. Neighbor-Joining trees for genes, present in at least three of the four grass species, are shown. Pairwise sequence identities for all predicted proteins are listed in supplementary table 2 (Supplementary Material online).

(fig. 1*A*) making orthology assignments difficult (fig. 3), suggesting possible gene conversion as previously reported for rice genes (Wang et al. 2007) and further complicates the analysis.

Synteny and Collinearity Perturbation

The *chalcone synthase* gene and a *kinase* pseudogene (names shown in blue in fig. 1*A*) are not present at a corresponding site in the other grass species. Duplication of the *ATPase* gene and *Nodulin* gene cluster in inverse orientation is also specific to wheat (fig. 1*A*). In previous papers (Chantret et al. 2005, 2008), we suggested that the second cluster was ancestral, based on orientation of the *ATPase* genes in wheat, rice, and barley. However, the order and orientation of the *ATPase* and *Nodulin* genes in *B. sylvaticum* and *S. bicolor* are the same as the order and orientation of the first cluster of the genes in wheat (fig. 1*A*), suggesting that it is ancestral. *ATPase3* gene, conserved between rice and *B. sylvaticum*, has no orthologs in sorghum and was probably not covered in the wheat sequenced region.

In *B. sylvaticum*, Sorghum, and rice, the *HIPL* gene is located between the *Ha* locus and the *ATPase* genes, but in the A genome of tetraploid, wheat is located at a noncollinear position separated from the *Ha* locus by 50 kb (fig. 1*A*; Chantret et al. 2008).

Collinearity between *Brachypodium*, Rice, and Sorghum outside of the Sequenced Wheat Region

We extended comparison between the 120-kb BAC clone of *B. sylvaticum* and the corresponding regions in Sorghum and rice (fig. 1*A*). Comparisons of the additional sequence in *Brachypodium*, rice, and Sorghum confirmed a high level of collinearity and similarity between the three grass species: 10 genes (of known or unknown functions, putative genes, pseudogenes, and gene relics) out of 21 in *B. sylvaticum*, 12 in rice, and 10 in Sorghum are orthologous in at least two of the species (figs. 1*A* and 3; supplementary table 2, Supplementary Material online). A 34-kb large inversion, including eight of the genes, differentiates *B. sylvaticum* from rice and Sorghum (fig. 1*A*).

Time of Emergence and Evolutionary Origin of the *Ha-Lik*e Genes

We searched the available genomic sequences of *B. sylvaticum*, rice, and Sorghum to determine whether ancestral *Ha-like* genes existed before the whole-genome duplication (ancient polyploidy) of the cereal genome, which occurred ~70 My, before the radiation of the major subfamilies (Paterson et al. 2004; Salse et al. 2008; Tang, Wang, et al. 2008). The paralogous regions resulting from the ancestral duplication and collinear to the *Ha* region, based on the overall content of conserved genes, were identified for rice, Sorghum, and *B. sylvaticum* (fig. 1*C*) using MCscan search (see Materials and Methods). The two genes flanking the *Ha* locus, *BGGP* and *HIPL*, are preserved in the three collinear paralogous genomic segments from *B. distachyon*, rice, and Sorghum separated by less then 10 kb (fig. 1*C*). These intergenic sequences were searched extensively, and no *Ha-like* genes or related sequences were identified. We concluded that the *Ha-like* genes emerged after the whole-genome duplication and after the divergence of *Pooideae* and *Ehrhartoideae* from *Panicoideae*.

Homologs of *Ha-like* genes (*Gsp-1*, *Pina* and *Pinb*) encoding GSP-1 and Puroindolines were previously identified in the *Triticeae* (wheat [*Triticum* and *Aegilops* species], barley [*H. vulgare*], and rye [*S. cereale*]) and *Aveneae* (oats: *A. sativa*) tribes (Tanchak et al. 1998; Gautier et al. 2000; Kan et al. 2006; Gollan et al. 2007; Mohammadi et al. 2007; reviewed by Bhave and Morris 2008).

Puroindoline-like proteins (products of *Ha-like* genes) from wheat endosperm and several other grasses (Blochet et al. 1993; Tanchak et al. 1998; Gautier et al. 2000; Kan et al. 2006; Gollan et al. 2007; Mohammadi et al. 2007; reviewed by Bhave and Morris 2008) are characterized by a cysteine skeleton and a unique TRD. Products of *Ha-Brachy1*, *Ha-Brachy2*, *Ha-Brachy3*, and *Ha-Brachy4* genes have the cysteine skeleton and one and two conserved residues of the TRD (fig. 4*A* and *B*; supplementary fig. 1, Supplementary Material online). The N-terminal 19-amino acid signal peptide and 100-amino acid domain also found in alpha amylase inhibitor and seed storage proteins

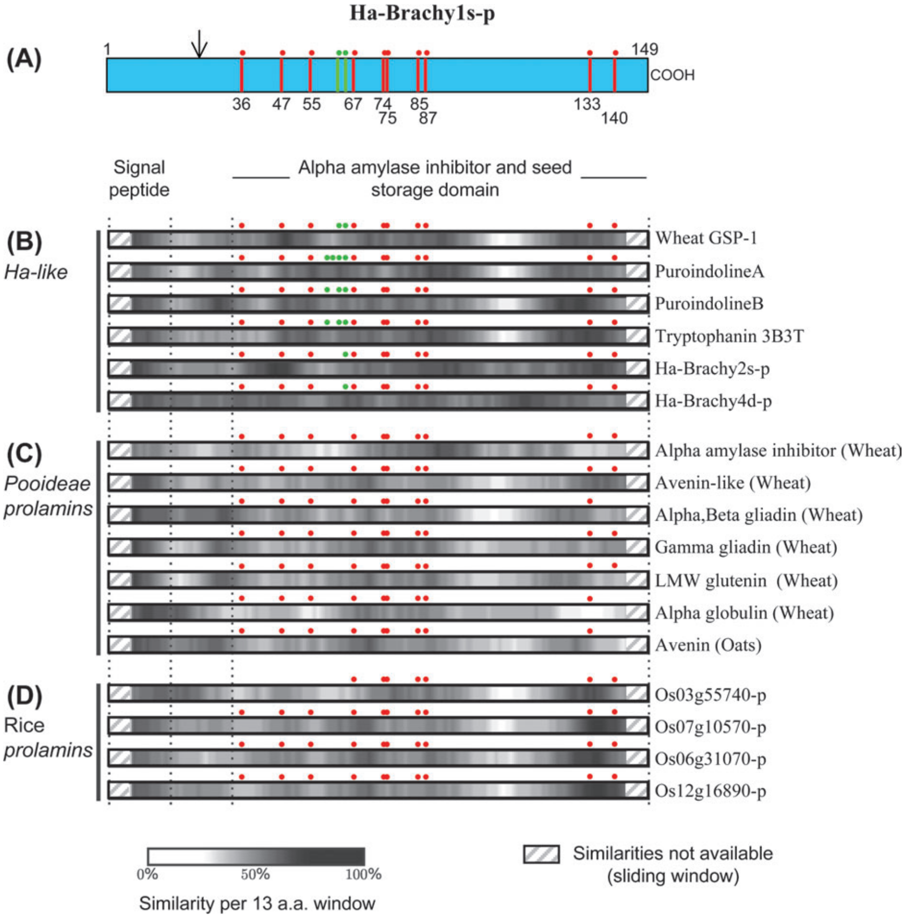

Fig. 4.—Similarity comparisons (Heatmap) between a putative protein encoded by *Ha-Brachy1s* gene of *Brachypodium sylvaticum* used as reference, GSP-1, Puroindolines, other Ha-like and Prolamins proteins. (*A*) An overview of the primary structure of the Ha-Brachy1s-p protein with its characteristic features: A 10-residue cysteine skeleton (red dots and vertical lines) and tryptophan residues (green dots and vertical lines) of the TRD. The N-terminal position of the processed protein is indicated by black arrow. (*B*) Comparison with representative Ha-like proteins from *Pooideae*. The *Brachypodium distachyon* Ha-like proteins are ended by (d), those of *B. sylvaticum* by (s). (*C*) Comparison with representative known *Pooideae* prolamins. (*D*) Comparison with representative cysteine-rich prolamins from rice (for an overview of all the rice prolamins, see supplementary fig. 2, Supplementary Material online). The 13-amino acid window sliding from the left to the right boundaries shown in gray was used for Heatmap comparisons. For the clarity of the illustration, comparisons of only representative sequences per each group of proteins showing a very high level of similarity are shown. The signal peptide is strongly conserved in all proteins, the cysteine skeleton is found in all proteins, whereas tryptophan residues of the TRD are found only in proteins encoded by *Ha-like* genes.

Uniprot reference of wheat and oats prolamin representative sequences:
Alpha/beta gliadin (P04721, P04722, P04723, P04724, and P04725) are represented by P04721.
Gamma gliadin (P04729, P04730, and P08453) are represented by P04729.
LMW glutenin (P10385 and P10386) are represented by P10386.
Avenin-like: A5A4L4.
Alpha amylase inhibitor (A4ZIZ0, Q4U195, and P01085) are represented by A4ZIZ0.
Alpha globulin: Q0Q5D4.
Avenin: DQ370180.

Reference of rice prolamin representative sequences:
Os03g55740-p represents Os11g0535100-p/Os03g55740-p/Os03g55734-p.
Os07g10570-p represents Os07g10570-p/Os07g10580-p.
Os06g31070-p represents Os06g31060-p/Os06g31070-p.
Os12g16890-p represents Os12g16880-p/Os12g16890-p/Os12g17010-p.

domain (IPR006106; http://www.ebi.ac.uk/interpro/IEntry?ac=IPR006106) are highly conserved in Ha-like proteins (fig. 4*B* and *C*; supplementary fig. 1, Supplementary Material online).

The cysteine skeleton of Puroindolines and GSP-1 proteins is also present in seed storage proteins of the prolamin superfamily (Kan et al. 2006; Bhave and Morris 2008). Prolamins encoded by *Alpha*, *Beta*, and *Gamma gliadin* and *low molecular weight (LMW)-glutenin* genes (Gao et al. 2007) as well as the *avenin* and *avenin-like* genes from oats and wheat show significant sequence similarities with Ha-like proteins (38–49%, depending on the domain, detailed in fig. 4). One gene from each of these prolamins was chosen as a reference for the subsequent sequence comparisons with *Ha-Brachy1* gene (fig. 4*C*). Although the cysteine skeleton is also generally well conserved (at least 7 out of 10 cysteine residues found in orthologous position), no tryptophan residues of the TRD found in Ha-like proteins are found in *Pooideae* prolamins. The peptide signal domain is still strongly conserved with wheat prolamins and avenins, whereas lower conservations were observed between IPR006106 domain of the *Ha-like* encoded proteins and the corresponding *Pooideae* prolamins (fig. 4*C*).

None of the 31 prolamin genes (annotated as prolamin or putative prolamin genes) found in the rice genome (http://rice.plantbiology.msu.edu/cgi-bin/putative_function_search.pl) contains the TRD characteristic of puroindolines. The 29 complete copies of these genes group in six clades (supplementary fig. 2, Supplementary Material online), four of which encode proteins with the cysteine skeleton and the IPR001954 domain (a "child" domain of IPR006106 found in gliadins and LMW glutenins). The coding sequence of the *Ha-rice-relic* is most similar to prolamin encoding genes belonging to these four groups. Interestingly, Ha-like proteins show higher amino acid sequence similarity to these rice prolamins than to *Triticeae* prolamins: gliadins and LMW glutenins (fig. 4*C* and *D*).

Finally, our analysis shows that several prolamins of *Panicoideae*, such as beta and gamma zeins (Woo et al. 2001), exhibit cysteine skeleton. None of these could be compared (aligned) with prolamins of *Ehrhartoideae* and *Triticeae* analyzed above (data not shown) because sequences are too divergent.

Discussion

Our study shows that *Ha-like* genes are present in *Brachypoidieae* (*B. sylvaticum and B. distachyon*), tribe sister to the *Triticeae,* and *Aveneae* tribes of the *Pooideae* subfamily of grasses. Although *Ha-like* genes were not initially found in *Ehrhartoideae* (rice: *O. sativa*) and *Panicoideae* (maize: *Zea mays*, and sorghum: *S. bicolor*; Gautier et al. 2000), genome sequence analysis of the *Ha* orthologous region from rice showed a short sequence related to *Ha-like* genes (Caldwell et al. 2004; Chantret et al. 2004, 2005) that is probably a nonfunctional truncated gene remnant (*Ha-rice-relic*). Similarly, *Ha-like* genes, with specific deletions, duplications and/or truncations, were identified at the *Ha* locus region of the *Brachypoidieae* tribe and additional *Ha-like* gene duplications (*Ha-Brachy3* and *Ha-Brachy4* genes) also occurred at 3 Mb from the *Ha* locus region. Thus, it was important to analyze and confirm the absence of *Ha-like*–related sequences at the *Ha* orthologous region of recently sequenced *S. bicolor* (*Panicoideae* subfamily of grasses; Paterson et al. 2009), which diverged earlier from *Pooideae* (wheat, barley, *Brachypodium*) and *Ehrhartoideae* (rice). Overall, comparative genome analysis of orthologous *Ha* regions as well as comparison with sequences of genes encoding prolamins from wheat, *Brachypodium*, rice, and Sorghum allow elucidation of evolutionary origin and time of emergence of *Ha-like* genes in grasses.

Evolutionary Origin of *Ha-Like* Genes

As the Ha-like proteins of *Triticeae* and *Aveneae*, the *Brachypoidieae* Ha-like proteins contain one and two conserved tryptophan residues of the TRD and a conserved cysteine skeleton (fig. 4*A*; Blochet et al. 1993; Gautier et al. 2000). These conserved features suggest that the *Brachypodium* Ha-like proteins may also play a role in determining endosperm hardness/softness, although this trait has not yet been investigated in this model species.

Our sequence comparisons confirmed previous observations of a common evolutionary origin of GSP-1 and Puroindolines encoded by the *Ha-like* genes and proteins of the prolamin superfamily (Kan et al. 2006; Bhave and Morris 2008). The prolamin superfamily was defined by Kreis et al. (1985) and initially comprised three groups of seed proteins rich in prolines and glutamines: the major prolamin storage proteins of *Triticeae* (alpha, beta and gamma gliadins; LMW glutenins), the alpha amylase/trypsin inhibitors of cereal seeds, and the 2S storage albumins from oilseed rape and other dicotyledonous plants. An expanded family includes also, among others, the major prolamins of *Panicoideae* and the alpha globulins of cereals (Shewry et al. 2004; Kan et al. 2006). All these prolamins are seed-specific proteins found only in the Plant Kingdom. It has been postulated that addition of a repetitive domain in grass *prolamin* genes accelerated their divergence and drastically limited their sequence homology with prolamins from other species (Shewry et al. 2002; Nagy et al. 2005). The TRD motif is specific to GSP-1 and Puroindolines encoded by *Ha-like* genes and is not shared with other prolamins (see Results and fig. 4).

None of the rice prolamin genes are conserved at orthologous position in Sorghum, confirming previously reported highly divergent and dynamic evolution of grass prolamins, similar to other seed storage proteins, not syntenic, often clustered and known to generate recombinant copies by gene fusion, duplication, or other types of genomic rearrangements (recombination, frameshifts; Kreis et al. 1985; Shewry et al. 2002; Nagy et al. 2005; Gao et al. 2007).

Evolution of *Ha-Like* Genes by Recurring and Independent Duplications and/or Deletions

The present study supplies further insights about dynamic evolution of the *Ha-like* genes through independent duplications and/or deletions, which appear to occur recurrently at different stages of the grass species evolution. *Gsp-1*, *Pina*/*Hina*, and *Pinb*/*Hinb* genes of *Triticeae* (wheat/barley) were most likely formed by duplication of an ancestral *Ha-like* gene (Caldwell et al. 2004; Chantret et al. 2005, 2008), closely after the divergence of the three tribes (*Triticeae*, *Aveneae*, and *Brachypoidieae*). Our study also shows that independent duplications and deletions of *Ha-like* genes (*Ha-Brachy1*, *Ha-Brachy2*, *Ha-Brachy1-relic*, *Ha-Brachy3*, and *Ha-Brachy4*) have also occurred in the *Brachypoidieae* lineage (figs. 1*A*, 1*B*, and 2). Another recent duplication occurred independently in barley (*Hinb-1 and Hinb-2*) (Caldwell et al. 2004). Deletions of *Ha-like* gene occurred also independently in the A and B genomes of *T. turgidum* (*Pina and Pinb*; Chantret et al. 2005), the G genome of in *T. timopheevi* (*Pinb*; Li et al. 2008), and in *B. distachyon* (*Ha-Brachy2*) as revealed in the present study.

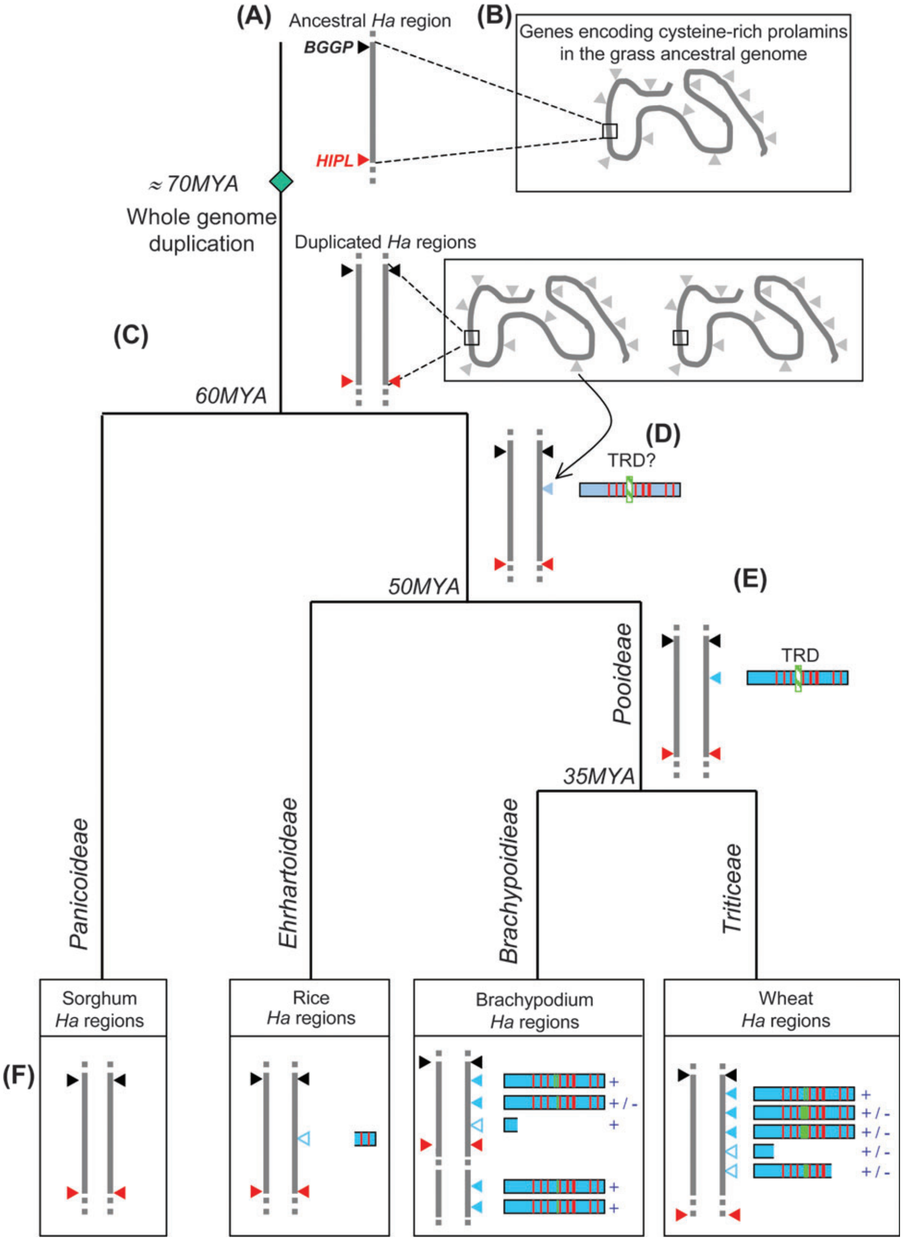

Fig. 5.—Suggested model for origin, time of emergence, and evolution of the *Ha-like* genes and locus in grasses. (*A*) Ancestral *Ha* region from the *BGGP* gene (black arrow) to the *HIPL* gene (red arrow) before the whole-genome duplication (polyploidization) predating divergence of grass subfamilies, with no *Ha-like* genes or related sequences. (*B*) Genes of the cysteine-rich prolamin superfamily were present in the ancestral grass genome. (*C*) The whole-genome duplication occurred in grasses 70 My (Paterson et al. 2004; Salse et al. 2008). (*D*) Emergence of an *Ha-like* gene in an ancestor of *Ehrhartoideae* and *Pooideae*, at one paralogous *Ha* region, after their divergence from *Panicoideae* by gene duplication, translocation, and divergence of a member of the prolamin superfamily. The cysteine residues are shown as red vertical lines in a rectangle representing the protein. (*E*) The TRD characteristic of GSP-1 and Puroindolines encoded by *Ha-like* genes (green vertical bar) appeared in an *Ha-like* gene ancestor either before or shortly after *Pooideae* and *Ehrhartoideae* diverged. (*F*) Evolution of *Ha* locus and genes by duplications, deletions, and/or truncations, occurring independently in each of the *Pooideae* and *Ehrhartoideae* families and tribes. "+" indicates *Ha-like* gene copies observed in all studies species, "+/−" those that were found deleted in specific species of the indicated grass family or tribe. *Ha-like* genes and prolamin encoding genes positions are marked with small blue triangles (filled for complete; empty for truncated or "pseudoized" copies). Rectangles represent primary structures of putative encoded proteins where the cysteine residues and tryptophan residues of the TRD are represented by, respectively, red and green vertical bars.

Time of Emergence of the *Ha* Locus in Grasses

There are two possible explanations of the presence of *Ha-like* genes on only one duplicated region in wheat, *Brachypodium*, and rice and their absence from both duplicated regions of Sorghum (the whole-genome duplication predating radiation of the major grass subfamilies is considered here) 1) The *Ha* genes emerged in this locus in a common ancestor of *Pooideae* and *Ehrhartoideae* after the duplication and after their divergence from *Panicoideae* or 2) an *Ha-like* gene was present in the ancestral grass genome but survived in only one of the two paralogous regions and only survived in some lineages, *Pooideae* and *Ehrhartoideae*, but not *Panicoideae*. Current evidence on the evolutionary origin of *Ha-like* genes—their closer relatedness to genes encoding prolamins of *Pooideae* and *Ehrhartoideae* than to those of Sorghum—favors the first explanation.

Concluding Remarks

As summarized in figure 5, the present study allows retracing of emergence, origin, and specific evolution of the *Ha-like* genes and locus. This locus emerged, in the ancestor of the *Pooideae* and *Ehrhartoideae*, between 60 and 50 My, as a new member of the prolamin gene family. The genes were subsequently lost in *Ehrhartoideae*. After independent duplications and divergent evolution, illustrating their rapid dynamic, *Ha-like* genes gained a new function in *Pooideae*, such as wheat, underlying the soft grain phenotype. Loss of these genes in some wheats leads, in turn, to hard endosperm seeds.

Supplementary Material

Supplementary figures 1 and 2 and tables 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). Sequence of *B. sylvaticum* BAC clone 37D5 was deposited at EMBL/GenBank under the accession number FJ234838.

Acknowledgments

We sincerely thank Dr Graham Moore and Dr Simon Griffiths (John Innes Center, Norwich, UK) for the *B. sylvaticum* BAC library (Foote et al. 2004) and for screening BAC clones with *Gsp-1*, *Pina*, and *Pinb* gene-specific probes; Dr Nathalie Boudet (URGV) for advices in gene annotation, Mrs Cécile Huneau (URGV) for technical assistance, Dr John Vogel (USDA, Albany, USA), and PI of the international *Brachypodium* initiative (http://www.brachypodium.org) for authorizing using the 4× available genome sequence of *Brachypodium distachyon* and supplying complementary sequences for gaps filling at *Ha-like* genes.

Literature Cited

Bhave M, Morris CF. 2008. Molecular genetics of puroindolines and related genes: allelic diversity in wheat and other grasses. Plant Mol Biol. 66:205–219.

Blochet JE, Chevalier C, Forest E, Pebay-Peyroula E, Gautier MF, Joudrier P, Pezolet M, Marion D. 1993. Complete amino acid sequence of puroindoline, a new basic and cystine-rich protein with a unique tryptophan-rich domain, isolated from wheat endosperm by Triton X-114 phase partitioning. FEBS Lett. 329:336–340.

Blochet JE, Kaboulou A, Compoint JP, Marion D. 1991. Gluten proteins. In: Bushuk W, Tkachuk R, editors. Gluten Proteins 1990. St Paul (MN): American Association of Cereal Chemists. p. 314–325.

Bossolini E, Wicker T, Knobel PA, Keller B. 2007. Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. Plant J. 49:704–717.

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.

Chalupska D, Lee HY, Faris JD, Evrard A, Chalhoub B, Haselkorn R, Gornicki P. 2008. Acc homoeoloci and the evolution of wheat genomes. Proc Natl Acad Sci USA. 105: 9691–9696.

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, et al. (19 co-authors). 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.

Chantret N, Salse J, Sabot F, et al. (17 co-authors). 2008. Contrasted microcolinearity and gene evolution within a homoeologous region of wheat and barley species. J Mol Evol. 66:138–150.

Charles M, Belcram H, Just J, et al. (13 co-authors). 2008. Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics. 180:1071–1086.

Darlington HF, Rouster J, Hoffmann L, Halford NG, Shewry PR, Simpson DJ. 2001. Identification and molecular characterisation of hordoindolines from barley grain. Plant Mol Biol. 47:785–794.

Draper J, Mur LA, Jenkins G, Ghosh-Biswas GC, Bablak P, Hasterok R, Routledge AP. 2001. *Brachypodium distachyon*. A new model system for functional genomics in grasses. Plant Physiol. 127:1539–1555.

Fabijanski S, Chang S-C, Dukiandjiev S, Bahramian MB, Ferrara P. 1988. The nucleotide sequence of a cDNA for a major prolamin (avenin) in oat (*Avena sativa* L. cultivar Hinoat) which reveals homology with oat globulin. Biochem Physiol Pflanzen. 183:143–152.

Faris JD, Zhang Z, Fellers JP, Gill BS. 2008. Micro-colinearity between rice, Brachypodium, and Triticum monococcum at the wheat domestication locus Q. Funct Integr Genomics. 8: 149–164.

Foote TN, Griffiths S, Allouis S, Moore G. 2004. Construction and analysis of a BAC library in the grass Brachypodium sylvaticum: its use as a tool to bridge the gap between rice and wheat in elucidating gene content. Funct Integr Genomics. 4: 26–33.

Gao S, Gu YQ, Wu J, et al. (11 co-authors). 2007. Rapid evolution and complex structural organization in genomic regions harboring multiple prolamin genes in the polyploid wheat genome. Plant Mol Biol. 65:189–203.

Gaut BS. 2002. Evolutionary dynamics of grass genomes. New Phytol. 154:15–28.

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 M, 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.

Gollan P, Smith K, Bhave M. 2007. Gsp-1 genes comprise a multigene family in wheat that exhibits a unique combination of sequence diversity yet conservation. J Cereal Sci. 45: 184–198.

Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. 16: 418–420.

Jurka J, Klonowski P, Dagman V, Pelton P. 1996. CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem. 20:119–121.

Kan Y, Wan Y, Beaudoin F, Leader DJ, Edwards K, Poole R, Wang D, Mitchell RAC, Shewry PR. 2006. Transcriptome analysis reveals differentially expressed storage protein transcripts in seeds of Aegilops and wheat. J Cereal Sci. 44:75–85.

Kellogg EA. 2001. Evolutionary history of the grasses. Plant Physiol. 125:1198–1205.

Kreis M, Forde BG, Rahman S, Miflin BJ, Shewry PR. 1985. Molecular evolution of the seed storage proteins of barley, rye and wheat. J Mol Biol. 183:499–502.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Law CN, Young CF, Brown JWS, Snape JW, Worland AJ. 1978. The study of grain protein control in wheat using whole chromosomes substitution lines. In: I.A.E. Agency. 1978. Seed protein improvement by nuclear techniques. Vienna (Austria): I.A.E. Agency. p. 483–502.

Li W, Huang L, Gill BS. 2008. Recurrent deletions of puroindoline genes at the grain hardness locus in four independent lineages of polyploid wheat. Plant Physiol. 146:200–212.

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.

McCarthy EM, McDonald JF. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 19:362–367.

Mohammadi M, Zaidi MA, Ochalski A, Tanchak MA, Altosaar I. 2007. Immunodetection and immunolocalization of tryptophanins in oat (*Avena sativa* L.) seeds. Plant Sci. 172:579–587.

Morris CF. 2002. Puroindolines: the molecular genetic basis of wheat grain hardness. Plant Mol Biol. 48:633–647.

Nagy IJ, Takacs I, Juhasz A, Tamas L, Bedo Z. 2005. Identification of a new class of recombinant prolamin genes in wheat. Genome. 48:840–847.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA. 101:9903–9908.

Paterson AH, Bowers JE, Bruggmann R, et al. (45 co-authors). 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 457:551–556.

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

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C. 2008. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell. 20:11–24.

Schofield JD. 1986. Flour proteins: structure and functionality in baked products. In: Blanshard JMV, Frazier PJ, Galliard T, editors. Chemistry and physics of baking. London: The Royal Society of Chemistry. p. 14–29.

Shewry PR, Beaudoin F, Jenkins J, Griffiths-Jones S, Mills EN. 2002. Plant protein families and their relationships to food allergy. Biochem Soc Trans. 30:906–910.

Shewry PR, Jenkins J, Beaudoin F, Mills ENC. 2004. The classification, functions and evolutionary relationships of plant proteins in relation to food allergens. In: Mills ENC, Shewry PR, editors. Plant food allergens. Oxford (UK): Blackwell Science. p. 24–41.

Simeone MC, Lafiandra D. 2005. Isolation and characterisation of friabilin genes in rye. J Cereal Sci. 41:115–122.

Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene. 167:GC1–GC10.

Tanchak MA, Schernthaner JP, Giband M, Altosaar I. 1998. Tryptophanins: isolation and molecular characterization of oat cDNA clones encoding proteins structurally related to puroindoline and wheat grain softness proteins. Plant Sci. 137:173–184.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. Science. 320:486–488.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 18:1944–1954.

Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. Genetics. 177:1753–1763.

Wicker T, Matthews D, Keller B. 2002. TREP: a database for Triticeae repetitive elements. Trends Plant Sci. 7: 561–562.

Wicker T, Sabot F, Hua-Van A, et al. (13 co-authors). 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.

Woo YM, Hu DW, Larkins BA, Jung R. 2001. Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. Plant Cell. 13:2297–2317.

Neelima Sinha, Associate Editor

Accepted April 13, 2009