Theor Appl Genet (2009) 119:1397–1412
DOI 10.1007/s00122-009-1143-x

ORIGINAL PAPER

Amplification of prolamin storage protein genes in different subfamilies of the *Poaceae*

Jian-Hong Xu · Joachim Messing

Received: 1 June 2009 / Accepted: 12 August 2009 / Published online: 29 August 2009

Abstract Prolamins are seed storage proteins in cereals and represent an important source of essential amino acids for feed and food. Genes encoding these proteins resulted from dispersed and tandem amplification. While previous studies have concentrated on protein sequences from different grass species, we now can add a new perspective to their relationships by asking how their genes are shared by ancestry and copied in different lineages of the same family of species. These differences are derived from alignment of chromosomal regions, where collinearity is used to identify prolamin genes in syntenic positions, also called orthologous gene copies. New or paralogous gene copies are inserted in tandem or new locations of the same genome. More importantly, one can detect the loss of older genes. We analyzed chromosomal intervals containing prolamin genes from rice, sorghum, wheat, barley, and *Brachypodium*, representing different subfamilies of the *Poaceae*. The *Poaceae* commonly known as the grasses includes three major subfamilies, the *Ehrhartoideae* (rice), *Pooideae* (wheat, barley, and *Brachypodium*), and *Panicoideae* (millets, maize, sorghum, and switchgrass). Based on chromosomal position and sequence divergence, it becomes possible to infer the order of gene amplification events. Furthermore, the loss of older genes in different subfamilies seems to permit a faster pace of divergence of paralogous genes. Change in protein structure affects their physical properties, subcellular location, and amino acid composition. On the other hand, regulatory sequence elements and corresponding transcriptional activators of new gene copies are more conserved than coding sequences, consistent with the tissue-specific expression of these genes.

Communicated by J. Snape.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1143-x) contains supplementary material, which is available to authorized users.

J.-H. Xu · J. Messing (✉)
Waksman Institute of Microbiology,
Rutgers University, 190 Frelinghuysen Road,
Piscataway, NJ 08854, USA
e-mail: messing@waksman.rutgers.edu

Introduction

Cereals include well-known species like rice (*Oryza sativa*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), rye (*Secale cereale*), oats (*Avena sativa*), maize (*Zea mays*), and sorghum (*Sorghum bicolor*). They belong to the family of the *Poaceae*, commonly known as the grasses. Their seeds constitute a staple for the human and animal diet. A major protein nutrient of the seeds of cereals is a protein family called prolamins, rich in proline and amide nitrogen derived from glutamine and encoded by many gene copies. The proteins are soluble in aqueous alcohol in contrast to the globulins, which are water soluble (Shewry et al. 1999). We now have reached an interesting stage of genomic resources, where gene families can be investigated in their chromosomal context. Because two cereal genomes have been sequenced in their entirety (Matsumoto et al. 2005; Paterson et al. 2009), we can now use collinearity to delineate amplification of genes, their insertion in tandem and in dispersed locations, and deletions of older gene copies (Xu and Messing 2008b). Identification of paralogous gene copies is based on the disruption of collinearity between the alignments of at least two genomes. Once each paralogous copy can be discerned by position in the genome, it also becomes possible to determine the contribution of each gene copy to the pool of gene products. Furthermore, using nucleotide substitution rates, we can

re-create gene movement events and the evolution of protein structure. Although only the genomes of rice and sorghum have been completely sequenced, we also have contiguous chromosomal sequences of all prolamin genes in maize (Xu and Messing 2008b) and *Brachypodium* (Huo et al. 2007). There are also BAC, EST, and map resources available from other cereals (Gao et al. 2007; Gu et al. 2006; Wicker et al. 2003). Therefore, we can now apply an evolutionary zoom to analyze the origin and divergence of an entire gene family that is critical for the quality of our food supply.

With the exception of rice, the major cereals fall into two groups (Kellogg 2001). The temperate cereals comprise wheat, barley, rye, and oats and the tropical cereals include maize, sorghum, and millets. Maize, sorghum, and most millets belong to the *Panicoideae* subfamily and their prolamins are divided into alpha, beta, gamma, and delta prolamins (Gibbon and Larkins 2005). Out of those, the main protein component in maize and sorghum is the alpha prolamins, which, in contrast to the others, is encoded by relatively large gene families. There are 42 copies in maize inbred B73 and 23 in sorghum Btx623, while the others have only one or two copies (Xu and Messing 2008b; Miclaus et al. in preparation). Wheat, barley, and rye are closely related species of the tribe *Triticeae*, which belongs to the subfamily of *Pooideae*. These species contain prolamin proteins, which can be classified into three groups: sulfur-rich (S-rich), S-poor, and high molecular weight (HMW) prolamins. S-rich prolamin, the major component in the tribe of *Triticeae*, comprises alpha/beta-gliadin, gamma-gliadin, and low molecular weight (LMW) glutenin in wheat, B-hordein and gamma-hordein in barley. S-poor prolamins (including omega-gliadins and C-hordeins) and HMW prolamins represent a smaller component (Shewry et al. 1984). All three prolamin groups in *Triticeae* are encoded by highly conserved genes based on chromosomal locations, although they differ in copy number (Shewry et al. 1999). Rice, like oats, differs from most cereals in that prolamins are the minor storage protein. In rice, they can be divided into three groups by size, 10, 13, and 16 kDa (Muench et al. 1999). Immunological cross-reactivity of cereal prolamins showed that rice prolamins did not cross-react with wheat gliadin antisera (Okita et al. 1988), and only cross-react weakly with those of maize, sorghum, and barley (Shyur et al. 1994). Furthermore, sequence-similarity of prolamins is very low between rice, maize and sorghum, and wheat and barley.

Given the previous description of the syntenic relationships of grass genomes based on genetically mapped markers (Gale and Devos 1998; Moore et al. 1995), one might have expected that these proteins might be less divergent in structure than their biochemical classification suggests. Actually, previous alignments of amino acid sequences of prolamins have been used to find common motifs that would suggest common ancestry. Therefore, Kreis et al. (1985) divided the *C*-terminal domain of prolamin protein into three regions varying in length from 20 to 40 residues, which they called A, B, and C. These three regions are conserved in all prolamins except alpha prolamins in sorghum and maize (Shewry and Tatham 1990). Regions related to A, B, and C are also present in inhibitors of trypsin and alpha-amylase, and even in some of the 2S globulins, which are water soluble storage proteins. Furthermore, regions A, B, and C were also shown to be similar to each other, indicating that they might have originated from an ancient triplication of a short domain with about 30 residues (Shewry and Tatham 1999).

However, rather than depending on protein sequence alone in the interpretation of the origin of these proteins, we suggest that changes in gene structure are the result of gene amplification events and losses of older gene copies to counteract divergence by gene conversion via pairing of non-homologous chromosomes (Xu and Messing 2008a). This type of analysis was not possible prior to the availability of contiguous genomic segments from different species sharing gene collinearity, where it is possible to discern genes present from ancestral chromosomes (Song et al. 2002; Xu and Messing 2008b). For instance, if a gene is collinear in genomes of two species, but absent in the third, one can assume that it was lost by deletion of ancestral chromosomal DNA, or it was inserted in the ancestral genome of the other two species. On the other hand, a gene copy in a unique position of one genome, but absent in the collinear regions of the two other genomes indicate an insertion event, or deletion event in an ancient genome. These two different scenarios can be confirmed by phylogenetic analysis because new insertion events are younger than the split of the progenitors of the species under comparison. Using such a method of syntenic alignments of chromosomal regions of multiple genomes including rice, wheat, barley, *Brachypodium*, and sorghum, we suggest that seeds probably first made water-soluble storage proteins before water-insoluble ones. One would also suspect that non-homologous pairing of chromosomal regions might have played a role in the concerted evolution of the prolamin gene family because of the loss of orthologous gene copies in different subfamilies of the *Poaceae* after they were copied and dispersed into different chromosomal locations as has been described for segmental duplications in rice (Xu and Messing 2008a).

Methods

Prolamin identification

Oryzeins were identified by homology searches of 10, 13, and 16 kDa prolamins in the rice genome. Brachypodins were identified by homology searches of prolamins (gliadins

and glutenins of wheat) against a 4× draft *Brachypodium distachyon* genome sequence (http://blast.Brachypodium.org). BAC and EST resources of other species were used in similar searches.

DNA sequences

Sequences of BAC clones from all species were downloaded from GenBank except sorghum and *Brachypodium* from The US Department of Energy Joint Genome Institute (JGI) websites (http://www.phytozome.net/cgi-bin/gbrowse.sorghum for sorghum and http://www.Brachypodium.org for *Brachypodium*). Accession numbers that were used are rice BAC clones: AC099043, AC133248, AC137747, AC134343, AC105320, AP005719, AP003943, AL928779 and AC113332; maize BAC clone AC204581; wheat BAC clones: DQ537336, EF426565; and barley BAC clone: AY268139.

Sequence annotation

Rice BAC clones annotated data were obtained from the RAP-DB (Tanaka et al. 2008). Sequences from other species were annotated by software fgenesh (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind), and predicted genes were manually modified and confirmed by EST and protein sequence resources. Only conserved genes were used in this study to align chromosomal regions between species.

Phylogenetic analysis and age estimation

Multiple nucleotide and amino acid sequences were aligned by the ClusterW program and MAFFT program (Katoh and Toh 2008). As can be seen in supplemental Figs. 2 and 3, amino acid repeat-blocks are easily recognized and would significantly interfere with a phylogenetic analysis. Therefore, these alignments were manually adjusted by removing these sequences before phylogenetic analysis. Phylograms were then drawn with the MEGA4 program using the NJ and UPGMA method (Tamura et al. 2007). The bootstrap was calculated with 1,000 replications. The synonymous substitution rates (r) for beta, gamma, and delta zeins and kafirins and the synonymous substitutions (Ks value) were calculated with Nei-Gojobori method (Nei and Gojobori 1986) as used before (Xu and Messing 2006). The average r and Ks were used to estimate the evolution time (t) for oryzeins, $t = Ks/r_{ave}$ (average synonymous substitution rate $r_{ave} = 6.18 \times 10^{-9}$).

Expression and promoter analysis

As prolamin genes do not have introns, all coding regions of genomic DNA can be directly aligned with cDNAs to determine whether the copy is expressed or not. All oryzein copies were compared to a collection of 28,000 cDNAs from Nipponbare that are deposited in KOME (http://cdna01.dna.affrc.go.jp/cDNA/). The expression of zeins and kafirins (sorghum and maize) was described previously (Xu and Messing 2008b), and all other prolamins were compared to sequences of the NCBI EST database (http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi). Regions of 1,000 bp upstream of the start codon were used for promoter analysis; P-box including endosperm motif and GCN4-like motif were identified by PLACE Web Signal Scan (http://www.dna.affrc.go.jp/PLACE/signalscan.html) (Higo et al. 1999).

Results

Prolamin nomenclature

Some prolamins have common names, such as gliadins and glutenins for wheat, kafirins for sorghum, but most of them are derived from their Latin generic names: zein from *Zea* (maize), hordein from *Hordeum* (barley), secalin from *Secale* (rye), avenin from *Avena* (oats). In some cases, no specific names have been applied to the prolamins of other species, such as *Brachypodium*. In rice, the name oryzin has been used, but it has also been used for *Aspergillus* alkaline proteinase. According to standard nomenclature, all new prolamins should be named by their Latin generic names. To avoid confusion, we re-named rice prolamins oryzeins and named *Brachypodium* prolamins brachypodins, which are derived from the rice Latin tribal name *Oryzeae* and *Brachypodium* Latin generic name *Brachypodium*. Rice prolamin genes that differ in their relative molecular weight, 10, 13, and 16-kDa, were named as *Ory10*, *Ory13*, and *Ory16*, respectively. We also wanted to point out that evolutionary and molecular biologists have previously applied the terms of family and subfamily for different classifications (Fig. 1); among species, the *Pooideae* are a subfamily of the *Poaceae* and, among proteins, the alpha zeins are a subfamily of the zeins or a group of proteins.

Genome-wide identification and analysis of oryzeins

As the rice genome has been completely sequenced (Matsumoto et al. 2005), we can conduct for the first time a systematic genome-wide analysis of all prolamins in rice. We could identify a total of 34 oryzein gene copies in the rice cultivar Nipponbare, which are located on chromosomes 3, 5, 6, 7, 11, and 12 (Table 1) and can be divided in three size classes, the 10, 13, and 16-kDa prolamins (Muench et al. 1999). There are four copies of 10-kDa, 24 copies of 13-kDa, and 6 copies of 16-kDa oryzeins. All

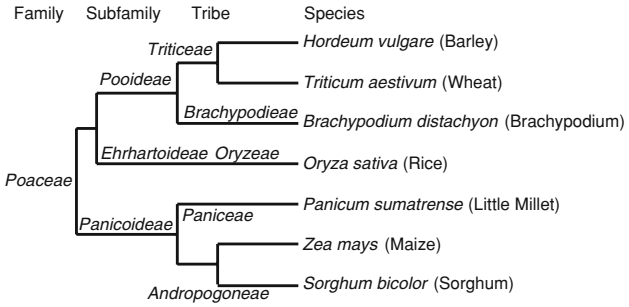

Fig. 1 The phylogenetic relationships of cereal's species, names of family, subfamily, tribe and species are shown. It is a generalized tree of relationships, with no computational support

oryzeins are divided into four clusters, *Ory10*, *Ory13a*, *Ory13b*, and *Ory16* (Fig. 2). When these chromosomal regions from rice chromosomes are aligned with orthologous regions from maize and sorghum, collinearity between all three genomes clearly indicates that oryzeins were new gene copies that arose after the *Ehrhartoideae* and *Panicoideae* split. The original prolamin gene shared between both, however, appears to be lost in the *Ehrhartoideae* and perhaps also in the *Panicoideae* subfamily (Table 2). On the other hand, all prolamin gene copies in sorghum were also collinear with maize as previously described (Xu and Messing 2008b).

Three copies of *Ory10* were tandemly duplicated on chromosome 3, and one copy *Ory10.4* was inserted into chromosome 11. The major prolamin group in rice is *Ory13* (24/34, 70%), which can be further subdivided into two subgroups *Ory13a* and *Ory13b*. There are only two copies of *Ory13a*, which were tandemly duplicated on chromosome 6, while 22 copies of *Ory13b* were tandemly duplicated on chromosome 5; a copy of *Ory13b* was also inserted into chromosome 7, following additional tandem duplication in its new location. *Ory16* has six gene copies on chromosome 12 and 7, where tandem duplications arose in both locations subsequently (Table 1, Fig. 2).

Reciprocal sequence alignments of the seven chromosomal oryzein regions with sorghum exhibit the same pattern of gene amplification. We could not find any traces of orthologous copies of oryzeins in sorghum (Table 2, Suppl Fig. 1). Therefore, the prolamins in sorghum represent insertion events that must have occurred after the progenitors of the *Ehrhartoideae* and *Panicoideae* subfamily split. On the other hand, analysis of two genomes of the *Panicoideae*, sorghum and maize from the tribe *Andropogoneae* have examples of both types of dispersed gene copies, those that pre-existed in their progenitors that split about 12 mya (Swigonova et al. 2004) and those that arose later. Interestingly, these findings are further corroborated by data from sugarcane and millets that contain highly conserved prolamin sequences homologous to the ones in sorghum and maize, but not to prolamins in the *Ehrhartoideae* subfamily (Xu and Messing 2008b).

To date the oryzein copying events, we needed to calculate their substitution rates (r). Because prolamins have the same function, we assumed that the average substitution rate of beta, gamma, and delta zeins and kafirins (6.18×10^{-9}) is the same for oryzeins. Based on this rate, the oldest oryzeins *Ory10*/*Ory13* arose about 47.5 million years ago (mya), which is just after rice, wheat, sorghum, and maize split 50 mya (Kellogg 2001). *Ory16* and *Ory13b* arose at about 28.5 and 12.2 mya, respectively. All oryzeins are gene copies that arose after the progenitors of the different subfamilies split and were inserted into chromosomal locations different to their donor copy that existed in a common progenitor of the grass family. Moreover, the *Panicoideae* subfamily exhibits similar dispersal modes, where prolamin copies were produced and inserted into unlinked positions. Nucleotide substitution rates also suggest another similarity of gene amplification among the different subfamilies of the grasses. In each tandem cluster the oldest copy is younger than the donor from the unlinked position, indicating that tandem duplication is followed by the insertion of the first paralogous copy into a new chromosomal location. While dispersal to new chromosomal locations can occur in long intervals, tandem duplications are spread over much shorter time periods. Therefore, it is not surprising that tandem copies form clusters distinct for each chromosomal location. Another interesting aspect of gene amplification is that tandem clusters seems to have arisen more frequently from recent dispersals than ancient copying events.

Chromosomal organization of prolamins in *Ehrhartoideae* (*Oryzeae*), *Panicoideae* (*Andropogoneae*), and *Pooideae* (*Triticeae*, *Brachypodieae*)

In search of a common denominator of prolamins between *Ehrhartoideae* and *Panicoideae*, we needed chromosomal information on prolamin genes from a third subfamily of the grasses. The best-suited examples are the prolamins from wheat and barley, which belong to the tribe *Triticeae* of the *Pooideae* subfamily of the grasses. Although no physical map of any *Triticeae* (wheat and barley) is available yet nor are their genomes sequenced, there are a few examples of contiguous genomic sequences containing prolamin genes, including gliadin and glutenin gene copies (Gao et al. 2007; Gu et al. 2006; Wicker et al. 2003). As these are selected regions of the genome, we cannot ask whether rice, maize, or sorghum genes were orthologous to wheat or barley. Although there is no biochemical evidence or ESTs that would suggest so, it could be that wheat and sorghum still had pseudogenes in these orthologous positions. On the other hand, we can pose the question, whether

Table 1 Summary of oryzein genes and their expression

Gene Copy	Accession	Chr	Position	Length (bp)	Comment	mRNA[a]	Expression
Ory10.1	AC099043	3	63543–63947	405	Intact	100%	Yes
Ory10.2	AC099043	3	66701–67108	408	Prestop	100%	Yes
Ory10.3	AC099043	3	69579–69983	405	Intact	100%	Yes
Ory10.4	AC133248	11	108057–108464	408	Intact	100%	Yes
Ory13a1	AP003618	6	53523–53074	450	No start codon	100%	Yes
Ory13a2	AP003618	6	59870–59421	450	Intact	100%	Yes
Ory13b1	AC137747	5	65486-65938	453	Prestop	95.6%	No
Ory13b2	AC137747	5	71336–71785	450	Prestop	94.7%	No
Ory13b3	AC137747	5	76664–77116	453	Prestop	95.1%	No
Ory13b4	AC137747	5	129657–130109	453	Intact	100%	Yes
Ory13b5	AC137747	5	132218–132670	453	Prestop	1SNP	No
Ory13b6	AC137747	5	134779–135231	453	Prestop	1SNP	No
Ory13b7	AC137747	5	137340–137792	453	Intact	100%	Yes
Ory13b8	AC137747	5	139901–140353	453	Intact	100%	Yes
Ory13b9	AC137747	5	142462–142914	453	Intact	100%	Yes
Ory13b10	AC134343	5	59425–69586	453	Insertion[c]	3SNPs	No
Ory13b11	AC134343	5	71692–72144	453	Intact	100%	Yes
Ory13b12	AC134343	5	85324–85776	453	Prestop	2SNPs	No
Ory13b13[b]	AC134343	5	87884–120440				
	AC105320	5	4242–29163	453	Insertion[c]	98.9%	No
Ory13b14	AC105320	5	35110–35562	453	Intact	4SNPs	No
Ory13b15	AC105320	5	73551–73090	462	Prestop	97.4%	No
Ory13b16	AC105320	5	83700–83239	462	Intact	100%	Yes
Ory13b17	AC105320	5	95533–95072	462	Intact	100%	Yes
Ory13b18	AC105320	5	101403–100942	462	Intact	100%	Yes
Ory13b19	AP005719	7	34844–34389	456	Intact	2SNPs	No
Ory13b20	AP005719	7	36850–36395	456	Intact	100%	Yes
Ory13b21	AP005719	7	38856–38401	456	Intact	6SNPs	No
Ory13b22	AP005719	7	46025–45570	456	Intact	100%	Yes
Ory16.1	AP003943	7	76117–76587	471	Intact	100%	Yes
Ory16.2	AP003943	7	81087–81557	471	Intact	100%	Yes
Ory16.3	AL928779	12	11946–12416	471	Intact	100%	Yes
Ory16.4	AL928779	12	14019–14489	471	Intact	100%	Yes
Ory16.5	AL928779	12	55488–59097	471	Insertion[c]	97.2%	No
Ory16.6	AL928779	12	65741–66211	471	Prestop	100%	Yes

[a] Sequence identity between genomic coding and cDNA sequences

[b] Retroelements insertion spans two BACs, AC134343 and AC105320

[c] Sequence after removal of insertion

the genomes of rice, sorghum, and maize had any sequence traces left that would indicate that the wheat and barley genes existed already in a common progenitor. However, when orthologous regions of wheat and barley were aligned via linked non-prolamin genes with rice, sorghum, and maize no traces of gliadin and glutenin gene copies were found (Fig. 3). Therefore, the wheat and barley genes were also likely copies of older genes that were inserted into their respective positions after the progenitors of the *Ehrhartoideae*, *Panicoideae*, and *Pooideae* subfamilies split. Having a second example now, where paralogous copies of prolamins were dispersed to unlinked positions of the genome, one can suggest that this type of amplification pattern occurred in parallel rather than in serial order.

Although no complete genome sequence within the tribe of the *Triticeae* is available yet, a member of the *Brachypodieae* tribe of the same *Pooideae* subfamily, *Brachypodium distachyon*, has recently been sequenced by whole genome shotgun

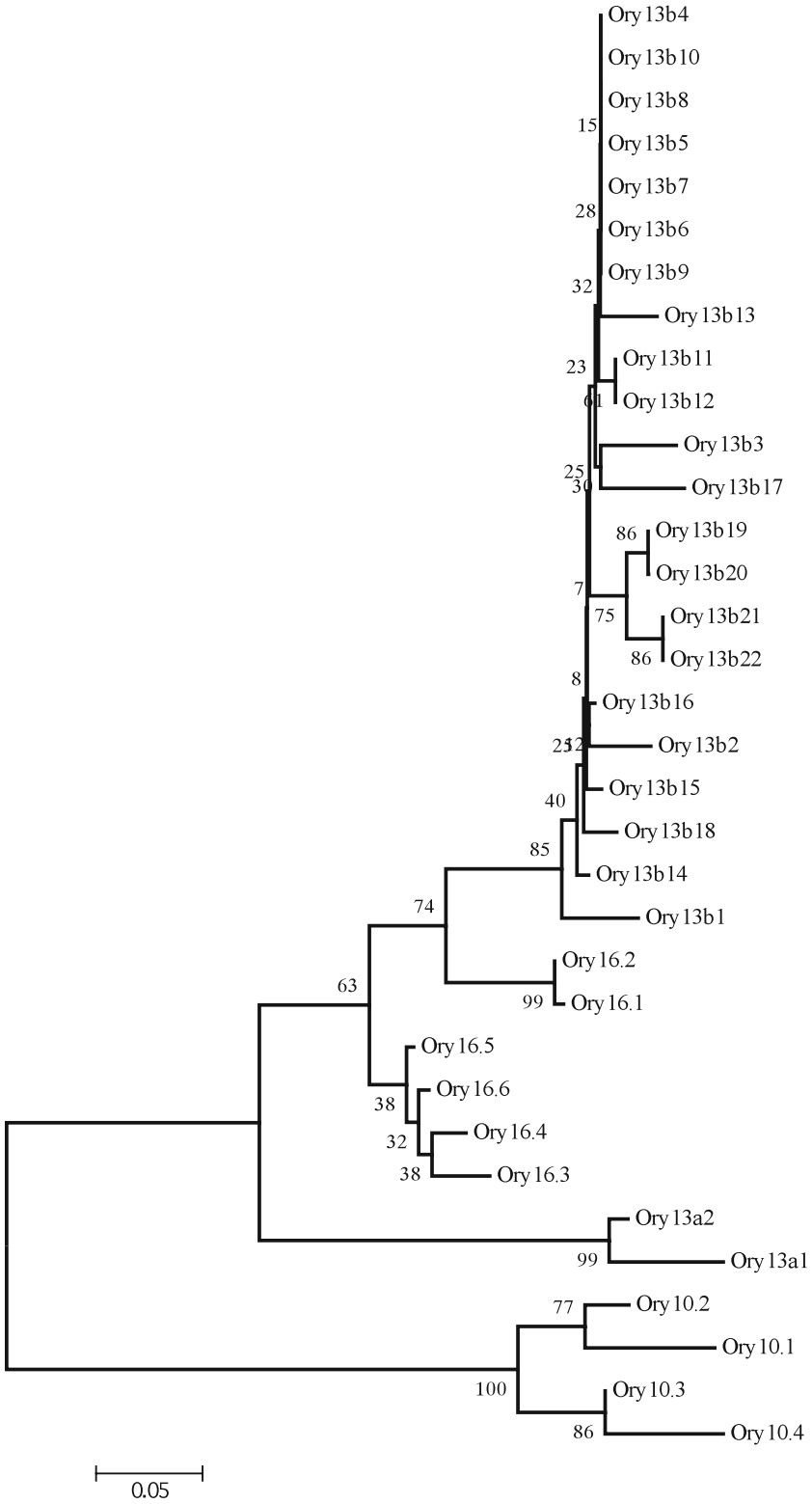

Fig. 2 Phylogenetic analysis of the rice oryzein genes. Genomic DNA sequences of all oryzein genes were obtained as described under "Methods" and the names are listed in Table 1. A phylogenetic tree was drawn using MEGA4 program with NJ method (Tamura et al. 2007)

sequencing at a level of 4× and assembled into supercontigs (www.Brachypodium.org). Because of its small genome size, *Brachypodium* has gained interest as a model system for the grasses similar like Arabidopsis (Huo et al. 2009). Belonging to the *Pooideae* subfamily, its low content of repeat sequences has permitted the use of collinearity to locate the *ph1* locus in wheat (Griffiths et al. 2006). We first tested its collinearity with wheat and barley prolamin gene copies. Indeed, it became possible to align two BAC (DQ537336 and EF426565) sequences from wheat with orthologous regions

Table 2 Oryzein gene copy number in rice and its orthology in sorghum

Locus	Rice	#	Sorghum	#
Ory10	Chr3	3	Chr1	0
	Chr11	1	Chr5	0
Ory13a	Chr6	2	Chr4	0
Ory13b	Chr5	18	Chr9	0
	Chr7	4	Chr2	0
Ory16	Chr7	2	Chr2	0
	Chr12	4	Chr8	0

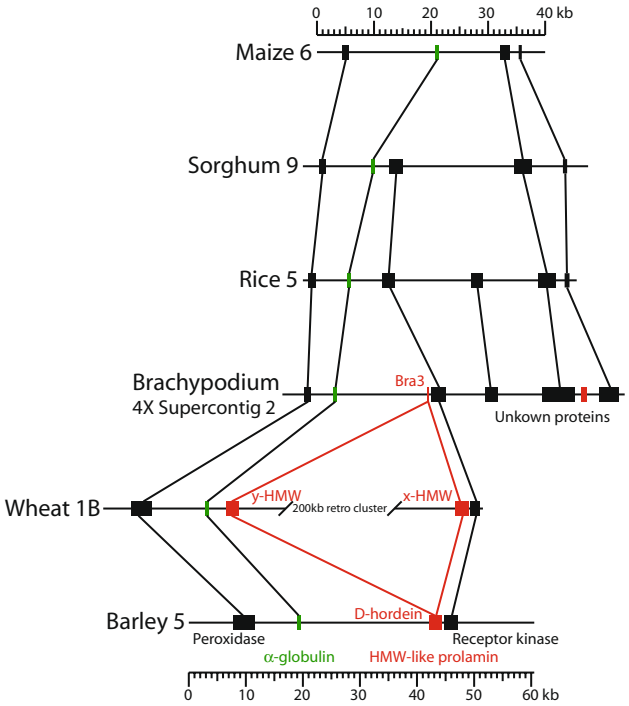

Fig. 3 Sequence alignment of orthologous regions of HMW prolamins. *Vertical bars* connect conserved genes. HMW prolamin genes are shown in *red*, alpha-globulin genes in *green*, and other conserved genes in *black*. A 200-kb long retrotransposon cluster in the collinear region of the wheat genome is deleted to allow easier alignment of conserved genes. The orthologous regions are AC204581.3 from maize chromosome 6 (Maize 6); 54,125,001-54,075,000 from Sorghum 9; AC113332 from Rice 5; 13,700,000-13,640,001 from 4× *Brachypodium* supercontig 2; DQ537336 from Wheat 1B; and AY268139 from Barley 5

of *Brachypodium*. Furthermore, using wheat prolamin's protein sequences, we identified additional copies of prolamin-like sequences in *Brachypodium*. Three copies of brachypodins [super2(13690441-13691139; 13674316-13674615; and 13646667-13647551)] are positioned orthologous to the HMW glutenins in wheat and D-hordein in barley, which we called HMW-like brachypodins (Fig. 3). We also found gliadin-like brachypodin copies syntenic to gliadins in wheat (data not shown). These results confirm previous assumptions that the *Brachypodium* genome has reasonable collinearity with the wheat and barley genomes and that the dispersal of these prolamin gene copies occurred in an ancestor to the *Brachypodieae* and *Triticeae* tribes, but after the *Pooideae* subfamily split from the rest of the *Poaceae* (Fig. 1). On the other hand, we could not find any prolamin loci from either the *Ehrhartoideae* (*Oryzeae*) or *Panicoideae* (*Andropogoneae*) in orthologous positions of the *Brachypodium* genome. Therefore, in all three grass subfamilies, new loci of prolamin gene copies arose in parallel rather than in a common ancestor. If this were the case, one would expect that a common gene existed in the *Poaceae* shared by genomes of different subfamilies. Alternatively, one would have to assume the older shared copies were deleted.

Evolutionary relationship of prolamins

A major challenge to the question above is the change that occurred at the nucleotide sequence level after speciation and to detect related sequences by homology. With more and more plant genome sequences becoming available, a clearer picture will emerge (Messing 2009). However, we argued that at this stage we could cluster protein sequences because they probably are more conserved than nucleotide sequences due to the function of prolamins as storage proteins and the molecular architecture of seeds. Although the seed biochemistry of species like *Brachypodium* or millets is lagging behind those of wheat and maize, we still can gain preliminary insight into the evolution of structure by assuming a conserved role of all prolamins in seed development. We therefore reasoned that a phylogenetic analysis of prolamins from the *Ehrhartoideae*, *Pooideae*, and *Panicoideae* subfamilies of the grasses could be conducted at the protein sequence level using prolamin sequences from rice, wheat, barley, rye, *Brachypodium*, sorghum, and maize (Suppl Fig. 2). The phylogenetic tree in Fig. 4 was drawn based on such multi-alignments of protein sequences.

Based on amino acid sequence homology prolamins essentially fall into three groups. Group I includes alpha and delta prolamins. The alpha prolamins are only present in *Panicoideae*, but not in the *Ehrhartoideae* and *Pooideae*. One prolamin sequence from little millet (*Panicum sumatrense*) clustered with alpha zeins and kafirins, indicating that little millet shares with maize and sorghum an alpha prolamin gene copy. Because alpha prolamin genes represent the youngest group among the prolamins, it is very likely that little millet also shares with maize and sorghum the other three prolamin groups (beta, gamma, and delta). Previous analysis of nucleotide substitution rates of prolamin genes indicated that the oldest alpha prolamin gene arose 22–26 mya, long after the progenitors of the *Panicoideae*, *Ehrhartoideae*, and *Pooideae* subfamilies split

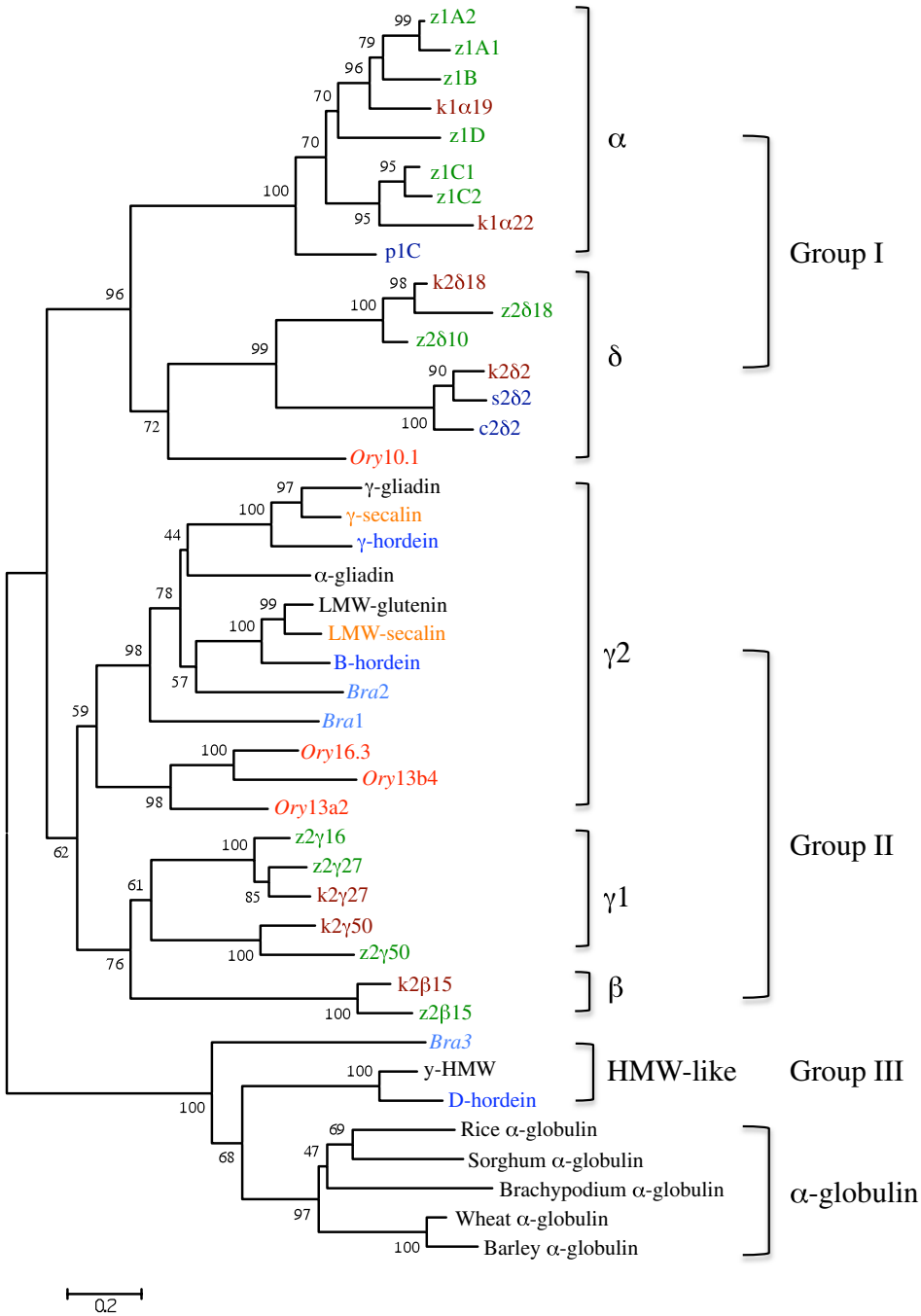

Fig. 4 Phylogenetic analysis of seed prolamin storage proteins. Kafirins and zeins in sorghum and maize have been described recently (Xu and Messing 2008b) and AAW82166 represents the prolamin p1C from little millet. Amino acid sequences for other prolamin genes from wheat, barley, rye, *Brachypodium*, and rice include: α-gliadin, ABS72143; γ-gliadin, ABO37961; LMW-glutenin, ABO37957; ω-gliadin, BAE20328; HMW glutenin, ABG68035; γ-hordein, AAA32955; B-hordein, CAA37729; C-hordein, AAA92333; D-hordein, AAP31051; γ-secalin, ABO32294; LMW-secalin, AAV86085; ω-secalin, AAB58043; *Bra1*, supercontig 1 (9503031-9503531); *Bra2*, supercontig 4 (9868110-9868655); *Bra3*, supercontig 2 (13462281-13462751); *Ory16*, *Ory16.3*; *Ory13b*, *Ory13b4*; *Ory13a*, *Ory13a2*; and *Ory10*, *Ory10.1*. Protein sequences were aligned with MAFFT program (Katoh and Toh 2008), and the phylogenetic tree was drawn using MEGA4 program with NJ method (Tamura et al. 2007)

(Xu and Messing 2008b). The closest related prolamins to alpha prolamins are the delta prolamins, suggesting that alpha prolamins have originated from delta prolamins.

The largest collection is group II, with copies in all grass species, include gamma and beta zeins/kafirins in maize and sorghum, *Ory13* and *Ory16* in rice, *Bra1* and *Bra2* in *Brachypodium*, S-rich prolamins (alpha-gliadins, gamma-gliadins, gamma-hordein, gamma-secalin, B-hordein, and LMW prolamins) in wheat, barley, and rye. Within the group II prolamins, those of the *Pooideae* (wheat, barley, rye, and *Brachypodium*) formed one cluster, indicating their close relationships, and those of the *Ehrhartoideae* (rice) and the *Panicoideae* (maize and sorghum) cluster together, respectively, indicating that the group II prolamins are conserved not only in their chromosome locations but also in their protein sequences within the grass subfamilies. Group III comprises only HMW-like prolamins, which are only present in *Pooideae*, include y-HMW from wheat, D-hordein from barley, and *Bra*3 from *Brachypodium*. HMW-like prolamins are much more closely related

to non-prolamin proteins (alpha-globulin) than to other prolamins, suggesting that HMW-like prolamins could be the oldest prolamins and the precursor of all other prolamins.

Progenitor of prolamins

If prolamins fall into three groups, are there any protein sequence motifs that they share? Previous studies suggested that regions A, B, and C in prolamins and some 2S storage proteins are also present in other seed proteins like trypsin and alpha-amylase inhibitors (Shewry and Tatham 1990). Based on these biochemical studies it was then hypothesized that storage proteins perhaps originated from an ancestor of trypsin inhibitor. While this still could be the case, the ABC model could be further refined. An interim step could be injected if the alpha globulin and the HMW prolamin genes arose from a tandem gene duplication of a storage protein gene followed by the divergence of each progeny copy. Therefore, we were intrigued with finding an HMW-like prolamin gene copy linked to an alpha gobulin gene copy in wheat. The gobulins are also seed storage proteins, but they are water-soluble in contrast to the water-insoluble prolamins. Water-soluble storage proteins are very ubiquitous and have been found in many species. Moreover, this linkage of the two different types of storage protein genes was conserved in the orthologous regions of barley and *Brachypodium* as well. Therefore, this gene pair seems to have arisen before the progenitors of the species belonging to the *Pooideae* split from the other subfamilies of the *Poaceae*. However, if one aligns the orthologous regions from species of the *Ehrhartoideae* (rice) and *Panicoideae* (maize and sorghum) subfamilies with these regions in the *Pooideae* (wheat, barley, and *Brachypodium*), one can find only the alpha globulin gene preserved in the syntenic position across all subfamilies, but the HMW-like prolamin gene copy in the *Pooideae* is absent in the *Ehrhartoideae* (rice) and *Panicoideae* (maize and sorghum) subfamilies (Fig. 3). Moreover, this globulin gene appears to be present in all other cereals as a single copy gene (Nakase et al. 1996; Woo et al. 2001), although it is missing in *Arabidopsis* (data not shown). The simplest explanation would be that the HMW-like prolamin and the alpha globulin gene arose from a progenitor gene by tandem duplication. However, in the *Ehrhartoideae* (rice) and *Panicoideae* (maize and sorghum) subfamilies the HMW-like prolamin gene copy was lost after copies of this gene were made and inserted somewhere else in the genome. The unlinked copies in *Ehrhartoideae* (rice) and *Panicoideae* (maize and sorghum) subfamilies diverged over time, but still cluster in the same group of prolamins (Fig. 4).

If this scenario is correct, then one would expect that phylogenetic analysis of amino acid sequences of the prolamins and the alpha globulin should yield a consistent relationship. Indeed, the phylogenetic tree shows that the alpha globulin is closer to the HMW glutenins/HMW-like prolamins than to inhibitors of trypsin and alpha-amylase (Fig. 5, Suppl. Fig. 3). Based on the genomic and phylogenetic analysis, we then can suggest that alpha globulins and HMW-like prolamins shared a common ancestral gene and that prolamins appear to have undergone a greater change in structure and degree in gene amplification than the globulins.

Prolamin expression and regulation

An interesting aspect of gene amplification is that it includes the regulatory flanking sequences. In case of the storage protein genes it ensures that the copies also express proteins during seed development. As we can now match EST data with individual gene copies, we also can investigate changes and conservation of regulatory relative to coding sequences. It is interesting that in a previous analysis

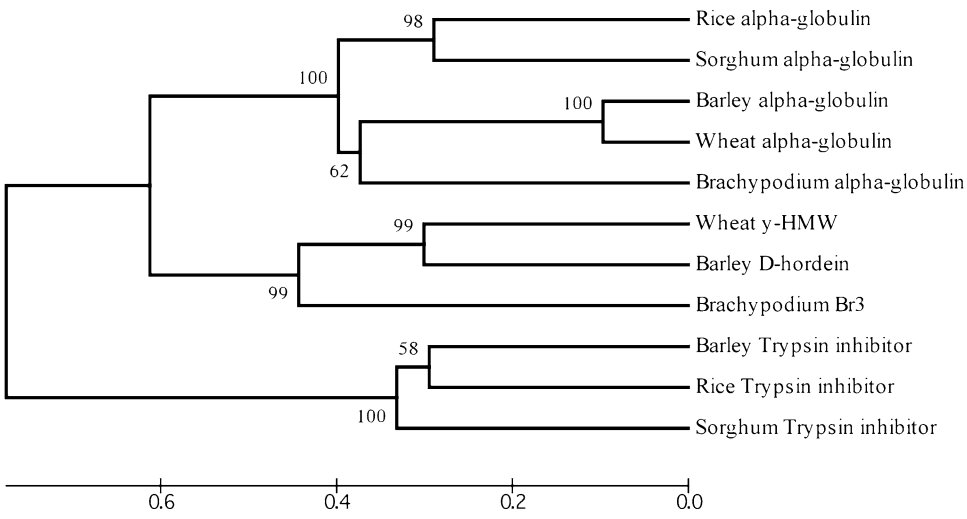

Fig. 5 The phylogenetic tree was produced using MEGA4 program with UPGMA method (Tamura et al. 2007). All HMW-like prolamins and alpha-globulin genes are the same as shown in Fig. 4. Trypsin inhibitors are CAA35188.1 from barley, ABK34470 from rice, and Sb02g006570.1 from sorghum. A tree was drawn using MEGA4 program with UPGMA method with 1,000 replicates (Tamura et al. 2007)

gene copies in maize have shown an uneven contribution of gene copies to the protein composition in the seed (Xu and Messing 2008b). For instance in maize inbred B73, out of 42 alpha zein gene copies, only 25 are transcribed, although at drastic different levels (Miclaus et al. ms in preparation). Moreover, within maize analysis of genomic regions from different inbred lines has shown that cluster sizes can vary through additional amplifications or deletions of amplified copies. Allelic and non-allelic copies in different haplotypes contribute at different levels to the total pool of closely related proteins, although their tissue-specific expression is well conserved (Llaca and Messing 1998; Song and Messing 2003).

In respect to the lower number of active gene copies, there is a certain degree of similarity among zeins and oryzeins. Out of 34 oryzeins, 21 are transcribed. It is interesting that in maize six of the transcribed genes would not translate to full-size proteins due to premature stop codons. In rice, two oryzeins that have premature stop codons and one that has no start codon are transcribed (Table 1), indicating that promoter regions are quite conserved, even if transcribed mRNAs are turned over more rapidly, as mRNAs of these defective genes in maize and rice accumulate at much reduced levels. On the other hand, three intact oryzein gene copies appear not to be transcribed, indicating that their promoters might be silenced (Table 1). Analysis of EST collections from wheat showed that as with zeins and oryzeins not all copies of gliadins and glutenins are expressed (Kawaura et al. 2005). For instance, 1–3 copies of HMW prolamin genes are silenced in wheat (Forde et al. 1985; Payne et al. 1981; Payne and Lancaster 1983).

As some prolamin gene copies with premature stop codons are still transcribed while others are not, they appear to be controlled at the transcriptional and/or post-transcriptional level. As all prolamin genes are derived from a common ancestor, function in seeds, and show similar patterns of expression, one would also expect that their regulation would be based on conserved protein–nucleic acids interactions. Comparison of the promoter regions of cereal prolamin genes have yielded a common conserved promoter element that is called the prolamin box (P-box), located about 300 bp upstream of the translation start codon (Boronat et al. 1986; Kreis et al. 1985; Ueda et al. 1994), which consists of two conserved sequence motifs: the endosperm (TGTAAAG) and *GCN4*-like motif (GLM) (A/G)TGAGTCAT. All alpha zeins and kafirins promoters have a conserved endosperm motif, but not the *GCN4*-like motif (Table 3). Gamma and beta zein/kafirin gene promoters have endosperm and *GCN4*-like motifs. However, *GCN4*-like motifs are not in close proximity to the endosperm motif (de Freitas et al. 1994). Surprisingly, the 18 kDa delta zein and kafirin promoters have neither endosperm nor *GCN4*-like motifs, although the 10-kDa delta zein does have an endosperm motif. In wheat, S-rich and S-poor prolamins have common endosperm and *GCN4*-like motifs. In rice, *ory13a* has the entire P-box as do S-rich and S-poor prolamins with both motifs in close proximity. Like gamma and beta zeins and kafirins, *ory16* has not only an endosperm motif, but also the *GCN4*-like motif, however, not in close proximity to the endosperm motif. Interestingly, *ory13b* has two conserved *GCN4*-like motifs, but no endosperm motif (Table 3). HMW glutenins and *Ory10* promoter regions have the conserved sequence TGCAAAG that is similar to endosperm motif (TGTAAAG). The three intact oryzein gene copies that are not transcribed have common promoter regions with the transcribed ones, indicating that absence of transcripts is either due to mRNA turnover or epigenetic regulation. Therefore, prolamin genes seem to be composed of regulatory modules that would require different transcriptional activators that either can act alone or in combination as it has been shown for the prolamin-box-binding factor (PBF) and the opaque-2 DNA-binding protein (O2) (Wang and Messing 1998).

Discussion

Alpha-globulin as progenitor of prolamins

Previous studies suggested that cereal prolamin genes have different origins among subfamilies of the grass family, *Ehrhartoideae* (rice), *Panicoideae* (maize and sorghum), and *Pooideae* (wheat and barley) (Fig. 1) based on sequence similarity and antibody cross-reactivity (Okita et al. 1988; Shyur et al. 1994). Furthermore, it was hypothesized that prolamins have a common ancestor with inhibitors of trypsin and alpha-amylase because of the presence of short amino acid sequence motifs, called A, B, and C (Kreis et al. 1985; Shewry and Tatham 1990). Using chromosomal linkage information based on multiple pairwise alignments of chromosomal regions either within subfamilies or across subfamilies of the grass family of species, we discovered that an alpha-globulin storage protein was a more recent progenitor to prolamins than inhibitors of trypsin and alpha-amylase (Fig. 5, Suppl Fig. 3). The alpha-globulin gene was not found in *Arabidopsis*, but it remained a single gene in grass genomes analyzed so far, while the donor copy for prolamins amplified and diverged in parallel to the evolution of different subfamilies of the *Poaceae*. These findings are consistent with the previous ABC model because amplification and divergence of prolamins retained the amino acid sequence motifs A, B, and C described earlier.

Chromosomal linkage and phylogenetic analysis indicated that the HMW-like prolamin is the oldest prolamin in cereal genomes and its gene should be present in all cereal

Table 3 Prolamin promoter regions in maize (zeins), sorghum (kafirins), wheat (gliadins and glutenins), and rice (oryzeins)

Prolamin	Endosperm motif		GCN4 motif				TATA box		Start codon
Zeins									
z2δ10	−261	CTTTACA					−93	TATAAATA	CCGCCATG
z2δ18							−74	TATAAATA	TCGCCATG
z2γ16			−187	ATGAGTCAT			−99	TATAAATA	ACACCATG
z2γ27	−352	TGTAAAG	−487	GTGAGTCAT			−104	TATAAATA	ACACCATG
z2γ50	−330	TGTAAAG							AAACCATG
z2β15	−241	TGTAAAG	−318	ATGAGTCAT			−99	TATAAATG	ACAGCATG
Kafirins									
k2δ18							−85	TATAAATA	CCGCCATG
k2γ27	−339	TGTAAAG	−192	ATGAGTCAT			−108	TATAAATA	ACACCATG
k2γ50	−334	TGTAAAG							CAACCATG
k2β15	−300	TGTAAAG	−486	ATGAGTCAT	−439	ATGAGTCAT	−138	TATAAATA	GTAGCATG
Gliadins									
α-gliadin	−315	TGTAAAGT	−300	ATGAGTCAT			−104	TATAAATA	CCACCATG
γ-gliadin	−301	TGTAAAGT	−286	ATGAGTCAT			−88	TATAAATA	CAACCATG
ω-gliadin	−316	TGTAAAGT	−301	ATGAGTCAT			−103	TATAAATA	CAAACATG
Glutenin									
LMW	−299	TGTAAAGT	−284	ATGAGTCAT			−79	TATAAATA	CCAACATG
HMW	−478	TGCAA-GC	−581	GTGAGTCAT			−91	TATAAAAG	TCGAGATG
Oryzeins									
ory10.1	−151	TGCAAAAA			−195	ATGACTCAT	−108	TATAAATA	CAACAATG
ory10.2	−117	TGCAAAGG					−73	TATAAATA	CAACAATG
ory10.3	−142	TGCAAAGG					−98	TATAAATA	CAGCAATG
ory10.4	−154	TGCAAAGG	−457	ATGACTCAT			−104	TATAAATA	CAGCAATG
ory13a1	−347	TGTAAAGT	−334	ATGACTCAT			−101	TATAAATA	TAGCTATA
ory13a2	−344	TGTAAAGT	−331	ATGACTCAT			−100	TATAAATA	TAGCCATG
ory13b1					−185	ATGACTCAC	−92	TATAAATA	CAGCAATG
ory13b2			−431	ATGAGTCAT			−92	TATAAATA	TAGCAATG
ory13b3			−447	ATGAGTCAT			−92	TGTAAATA	TAGCAATG
ory13b4			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b5			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b6			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b7			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b8			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b9			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b10					−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b11			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b12					−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b13			−445	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATG	CAGCAATG
ory13b14			−445	ATGAGTCAT			−92	TATAAATG	CAGCAATG
ory13b15					−185	ATGACTCAC	−92	TATAAATG	TAGCAATG
ory13b16			−449	ATGAGTCAT	−186	ATGACTCAC	−92	TATAAATA	TAGCAATG
ory13b17			−448	ATGAGTCAT	−185	ATGACTCAC	−92	TATAAATA	TAGCAATG
ory13b18			−442	ATGAGTCAT	−179	ATGACTCAC	−86	TATAAATA	TCGCAATG
ory13b19			−306	ATGAGTCAT	−187	ATGACTCAC	−94	TCTAAATG	CAACAATG
ory13b20			−306	ATGAGTCAT	−187	ATGACTCAC	−94	TATAAATG	CAACAATG
ory13b21			−306	ATGAGTCAT	−187	ATGACTCAC	−94	TATAAATG	CAACAATG

Table 3 continued

Prolamin	Endosperm motif		GCN4 motif				TATA box		Start codon
ory13b22			−306	ATGAGTCAT	−187	ATGACTCAC	−94	TATAAATG	CAACAATG
ory16.1	−543	TGTAAAGT	−278	ATGAGTCAT	−160	ATGACTCAC	−90	TATAAATA	CAACAATG
ory16.2	−537	TGCAAAGT	−277	ATGAGTCAT	−159	ATGACTCAC	−89	TATAAATA	CAAAAATG
ory16.3					−182	ATGACTCAC	−90	TATAAATA	CCACAATG
ory16.4	−397	TGTAAAGT			−182	ATGACTCAC	−90	TATAAATA	CCACAATG
ory16.5					−182	ATGACTCAC	−90	TATAAATA	CCACAACG
ory16.6	−397	TGTAAAGT			−182	ATGACTCAC	−90	TATAAATA	CCACAATG

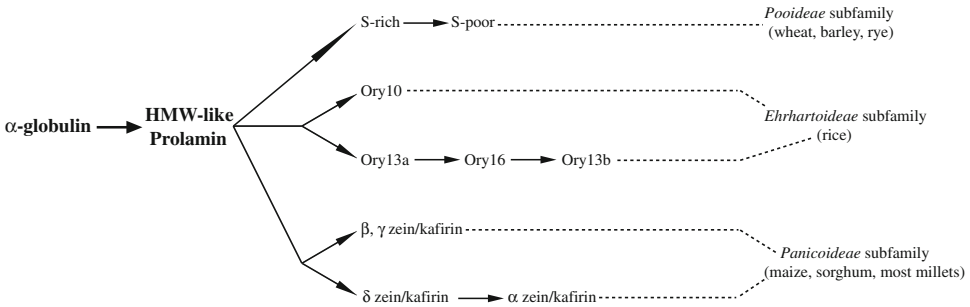

Fig. 6 Hypothetical prolamin evolution pathways in cereals. Because of the conservation of protein structure between alpha-globulin and HMW-like prolamin, their locations in an ancestral chromosome, and previous studies (alpha-globulin has prolamin function with C-terminal change in rice, ancient duplicated transcriptional factor *O2*-like and *OHP*-like genes regulate prolamins and alpha-globulin, respectively), it is hypothesized that HMW-like prolamin arose from alpha-globulin. This specific order of change derives from the fact that alpha-globulin is conserved in a syntenous position of diverged subfamilies of the grasses, while HMW-like prolamin is not. Only the progenitor of the *Pooideae* subfamily kept the original HMW-like prolamin gene in the orthologous position, while progenitors for the other subfamilies lost this gene after it was copied and inserted into another location, resulting in lineage-specific divergence of prolamin copies. *Horizontal arrows* indicate these lineages as time progresses

species. However, this HMW-like prolamin gene seems to be present only in species of the *Pooideae* subfamily (Fig. 3). Still, the alpha-globulin gene is preserved in an orthologous position across subfamilies, both in rice and sorghum. Therefore, the most parsimonious explanation is that the HMW-like prolamin gene was copied and a paralogous copy inserted somewhere else. Only the progenitor of the *Pooideae* subfamily kept the original HMW-like prolamin gene in the orthologous position, while progenitors for the other subfamilies lost this gene. Indeed, it seems to be quite common that donor gene copies can be lost when new gene copies (younger genes) take over (Xu and Messing 2008b). Paralogous genes had greater freedom to evolve when the donor copy is not available for copy correction, resulting in the divergence of prolamin genes in the other subfamilies of the *Poaceae* (Fig. 6). Although such copy correction is probably infrequent, we recently showed that such an event took place for one gene between two segmental duplications in rice, indicating non-homologous pairing of chromosomes (Xu and Messing 2008a). Although such events are prevented in meiosis, they could occur as mitotic events, which could be transmitted because the germline in plants is contiguous with reproductive tissue.

Diversification of prolamins by compartmentalization and digestibility

Given this new perspective of ancestry of prolamins from a globulin, it is useful to consider the compartmentalization of storage proteins in the seed. Prolamins in maize, sorghum, and rice are translated at the rough endoplasmatic reticulum (ER) in the inner endosperm and are deposited into protein bodies (PB) as an extension of the ER while globulins and glutelins are further transported to vacuole-like organelles (Li et al. 1993; Woo et al. 2001). Based on these two locations, protein bodies are classified as PB-I and PB-II, respectively. Interestingly, in the early stage of rice seed development, alpha-globulin is present in PB-I, but then is transported to PB-II in the immature endosperm and sequestered in the matrix that surrounds the crystalloids (Kawagoe et al. 2005). However, when the *C*-terminal sequence of alpha-globulin is deleted, the truncated version accumulates at high levels in PB-I (Kawagoe et al. 2005), suggesting it recovers the prolamin targeting properties. Therefore, one can envision that prolamins originated from alpha-globulin by a *C*-terminal truncation. From a mechanistic point of view, such diversification could have

occurred by unequal crossing over of tandem gene copies. Indeed, we can find an alpha-globulin gene linked to an HMW-like prolamin gene in *Brachypodium*, wheat, and barley (Fig. 3). Changes in protein size by unequal crossing over of tandem copies have also been found among alpha prolamins of maize and sorghum with 22-kDa and 19-kDa alpha zeins/kafirins (Xu and Messing 2008b).

Parameters for selection in protein structure could also be based on the interaction with proteases because the function of storage proteins is to provide amino acids during germination of the seed. Furthermore, humans would select species based on the digestibility of proteins, again a question of the hydrolysis of seed proteins. In addition to alpha globulin, most of the rice prolamins are group II (gamma 2) prolamins. However, in contrast to maize prolamins, which are mainly alpha prolamins, rice prolamins are not easily digestible by monogastric animals, reducing the nutritional quality of the rice seed. Although prolamin proteins in wheat and barley belong to the group III (HMW) and group II (gamma 2) prolamins, they are easily digested in feed and food. However, in contrast to rice, they are synthesized in ER but not retained within the lumen of the rough ER and transported into protein storage vacuoles through the Golgi complex or by autophagy (Herman and Larkins 1999; Levanony et al. 1992). Therefore, it appears that diversification of prolamin proteins have led to structures that accumulate in different subcellular structures and complexes and in such configurations acquired different nutritional qualities as well.

Another case of the loss of a progenitor gene are the group I prolamins, which are present only in the species of *Ehrhartoideae* and *Panicoideae* subfamilies, and absent in the species of the *Pooideae* subfamily (Fig. 4), suggesting group I prolamins are only copied in the species of *Ehrhartoideae* and *Panicoideae* subfamilies, or lost in the species of the *Pooideae* subfamily. The prolamins in species of *Panicoideae* subfamily further diverged into two additional subgroups, the delta and alpha prolamins, while prolamins in the *Ehrhartoideae* subfamily did not, which might explain that no further amplification of prolamin genes occurred in rice and as a consequence prolamins remained a minor component in rice compared to sorghum and maize.

It is interesting that divergence also seems to occur through tandem rather than dispersed amplification of gene copies. It appears that the degree of tandem amplification has increased with the formation of the most recent paralogous gene copies. For instance, all species have group II prolamins, which are younger than the HMW-like prolamins and includes large tandem clusters in rice as described above. The *Panicoideae* subfamily added two new prolamin subfamilies, alpha and delta prolamins, although *Ehrhartoideae* subfamily has few copies of delta prolamins. Again, the younger protein subfamily, the alpha prolamins, is the larger gene family with large tandem gene clusters. In addition to gene family size, the younger gene subfamily also generated protein variants to seed development that differentiated from a compartmentalization point of view. It has been shown in maize that beta and gamma zein genes are expressed earlier in development than delta and alpha zein genes. While the beta and gamma zein accumulate at the periphery of the protein bodies, the delta and alpha accumulate in the inner core (Esen and Stetler 1992; Lending and Larkins 1989). Furthermore, in transgenic experiments it could be shown that the early expression pattern of beta and gamma zein genes in maize endosperm presumably provide an initializing role for compartmentalization so that delta and alpha zeins can accumulate in a stable fashion, indicating a higher structure of protein bodies. In a heterologous transgenic system of tobacco, alpha prolamins needed gamma prolamins to stably accumulate and were co-localized with gamma prolamin in ER-derived protein bodies (Coleman et al. 1996). Although delta prolamin was fairly stable and formed protein bodies in transgenic tobacco, it became even more stable and accumulated at much higher levels when co-expressed with beta prolamin; it also was co-localized in the beta-containing protein bodies (Bagga et al. 1997). Therefore, beta and gamma prolamins as the older prolamins appear to play an important structural and functional role for the stability of protein bodies.

Evolution of promoters

The expression of prolamins is restricted to endosperm cells during seed development. Studies of promoter regions of prolamin genes have shown multiple *cis*-acting elements and transcriptional activators. A major element is the P-box conserved in most of seed prolamin storage protein gene promoters, which contains the endosperm motif (TGTAAAG) and *GCN4*-like motif [(A/G)TGAGTCAT] (Table 3). It is believed that they are key elements in controlling the endosperm-specific expression. The corresponding transcriptional activators in maize are called prolamin-box binding factor (PBF) and O2. PBF specifically interacts with a 5′-AAAG-3′ or its complementary 5′-CTTT-3′ sequence. It binds the endosperm motif and controls the expression of prolamins and interacts with the O2 protein as well because of the short spacing between the two motifs (Ueda et al. 1994; Vicente-Carbajosa et al. 1997; Wang and Messing 1998; Wang et al. 1998).

The O2 protein recognizes a specific target site (TCCACGTAGA) in the promoters of 22 kDa zeins (Muth et al. 1996; Schmidt et al. 1992; Ueda et al. 1992). It can bind to a different target site (GACATGTC) in the promoters of alpha kafirin and alpha coixin genes (Yunes et al. 1994). O2 has also been shown to bind *GCN4*-like motif in the

b-32 gene (Lohmer et al. 1991), LMW-glutenin (Holdsworth et al. 1995), gamma-zein (Marzabal et al. 1998), and rice glutelin (Wu et al. 1998). The O2-like genes were identified in other cereals, such as *SPA* in wheat (Albani et al. 1997), *BLZ2* in barley (Onate et al. 1999), *RISBZ1* in rice (Onodera et al. 2001), and they also can bind the *GCN4*-like motif and control the expression of prolamins. Interestingly, wheat *SPA* and barley *BLZ2* can also interact with PBF (Conlan et al. 1999; Mena et al. 1998) and, indeed, *GCN4*-like motif and endosperm motif are both present in promoters of prolamins of maize, wheat, and barley (Table 3). The major rice prolamin *ory13b* has two *GCN4*-like motifs (Table 3), which could explain why in rice the *RISBZ1* protein only binds to the *GCN4*-like motif to activate transcription without interacting with PBF (Onodera et al. 2001).

Interestingly, a bZIP protein (REB), which is similar to the O2 heterodimerizing protein (OHP) of maize, can bind to the promoter sequence (GCCACGTCAG) of the alpha-globulin gene in rice, resembling the O2 motif in the promoter of 22 kDa alpha-zeins (Nakase et al. 1997). Unexpectedly, we found that the REB protein is the ortholog of OHP1 and OHP2, which arose by an ancient segmental duplication of the ancient O2 locus (≈56 mya) before the Poaceae subfamilies split 50 mya (Xu and Messing 2008a). Therefore, it appears that both the regulatory and the target gene underwent duplication in an ancestral species of the *Poaceae*, which would simplify for each target gene to diverge from each other. However, unlike the amplification and dispersal of target gene copies in parallel of multiple subfamilies of the *Poaceae*, the pair of regulatory genes remained fixed in their chromosomal locations. Furthermore, the conservation of the regulation of alpha-globulins by *OHP*-like genes and prolamins by *O2*-like genes lends additional credence to our hypothesis that prolamins arose from alpha-globulin by an ancient duplication before the cereals split 50 mya.

Comparison to amplification of disease resistance genes

Gene amplification has been analyzed in several plant genomes, but mostly in respect to disease resistance genes (Leister 2004). These studies have investigated potential mechanisms of gene amplification. Similar to storage protein genes, disease resistance genes are thought to have amplified by tandem duplication and ectopic duplication (Meyers et al. 2003; Richly et al. 2002), here called unlinked copies. It is interesting that older copies of disease resistance genes also frequently become inactivated, which has led to a model of birth and death of genes (Michelmore and Meyers 1998). This model has been used to explain diversifying selection and would apply to the differential trafficking and digestibility of storage proteins described above. However, in contrast to the ectopic duplication of gene copies, it has been proposed that most duplication of disease resistance genes might have occurred via segmental duplication (Baumgarten et al. 2003). In this study, we can contrast both mechanisms. In case of the regulatory gene, O2-like genes, the model of segmental duplication appears to be correct, whereas the syntenic alignments of storage protein genes argue for ectopic duplication events of gene copies (Xu and Messing 2008a).

Acknowledgments The research described in this manuscript was supported by the Selman A. Waksman Chair in Molecular Genetics and a grant from the DOE (# DE-FG05-95ER20194) to JM. The *Brachypodium* sequence data was produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/.

References

Albani D, Hammond-Kosack MC, Smith C, Conlan S, Colot V, Holdsworth M, Bevan MW (1997) The wheat transcriptional activator SPA: a seed-specific bZIP protein that recognizes the GCN4-like motif in the bifactorial endosperm box of prolamin genes. Plant Cell 9:171–184

Bagga S, Adams HP, Rodriguez FD, Kemp JD, Sengupta-Gopalan C (1997) Coexpression of the maize delta-zein and beta-zein genes results in stable accumulation of delta-zein in endoplasmic reticulum-derived protein bodies formed by beta-zein. Plant Cell 9:1683–1696

Baumgarten A, Cannon S, Spangler R, May G (2003) Genome-level evolution of resistance genes in *Arabidopsis thaliana*. Genetics 165:309–319

Boronat A, Martìnez MC, Reina M, Puigdomënech P, Palau J (1986) Isolation and sequencing of a 28 kD glutelin-2 gene from maize. Common elements in the 5′ flanking regions among zein and glutelin genes. Plant Science 47:95

Coleman CE, Herman EM, Takasaki K, Larkins BA (1996) The maize gamma-zein sequesters alpha-zein and stabilizes its accumulation in protein bodies of transgenic tobacco endosperm. Plant Cell 8:2335–2345

Conlan RS, Hammond-Kosack M, Bevan M (1999) Transcription activation mediated by the bZIP factor SPA on the endosperm box is modulated by ESBF-1 in vitro. Plant J 19:173–181

de Freitas FA, Yunes JA, da Silva MJ, Arruda P, Leite A (1994) Structural characterization and promoter activity analysis of the gamma-kafirin gene from sorghum. Mol Gen Genet 245:177–186

Esen A, Stetler DA (1992) Immunocytochemical localization of delta-zein in the protein bodies of maize endosperm cells. Am J Bot 79:243–248

Forde J, Malpica JM, Halford NG, Shewry PR, Anderson OD, Greene FC, Miflin BJ (1985) The nucleotide sequence of a HMW glutenin subunit gene located on chromosome 1A of wheat (*Triticum aestivum* L.). Nucleic Acids Res 13:6817–6832

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

Gao S, Gu YQ, Wu J, Coleman-Derr D, Huo N, Crossman C, Jia J, Zuo Q, Ren Z, Anderson OD, Kong X (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

Gibbon BC, Larkins BA (2005) Molecular genetic approaches to developing quality protein maize. Trends Genet 21:227–233

Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G (2006) Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. Nature 439:749–752

Gu YQ, Salse J, Coleman-Derr D, Dupin A, Crossman C, Lazo GR, Huo N, Belcram H, Ravel C, Charmet G, Charles M, Anderson OD, Chalhoub B (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. Genetics 174:1493–1504

Herman EM, Larkins BA (1999) Protein storage bodies and vacuoles. Plant Cell 11:601–614

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res 27:297–300

Holdsworth MJ, Munoz-Blanco J, Hammond-Kosack M, Colot V, Schuch W, Bevan MW (1995) The maize transcription factor Opaque-2 activates a wheat glutenin promoter in plant and yeast cells. Plant Mol Biol 29:711–720

Huo N, Lazo GR, Vogel JP, You FM, Ma Y, Hayden DM, Coleman-Derr D, Hill TA, Dvorak J, Anderson OD, Luo MC, Gu YQ (2007) The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences. Functional and Integrated Genomics 8:135–147

Huo N, Vogel JP, Lazo GR, You FM, Ma Y, McMahon S, Dvorak J, Anderson OD, Luo MC, Gu YQ (2009) Structural characterization of Brachypodium genome and its syntenic relationship with rice and wheat. Plant Mol Biol 70:47–61

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–298

Kawagoe Y, Suzuki K, Tasaki M, Yasuda H, Akagi K, Katoh E, Nishizawa NK, Ogawa M, Takaiwa F (2005) The critical role of disulfide bond formation in protein sorting in the endosperm of rice. Plant Cell 17:1141–1153

Kawaura K, Mochida K, Ogihara Y (2005) Expression profile of two storage-protein gene families in hexaploid wheat revealed by large-scale analysis of expressed sequence tags. Plant Physiol 139:1870–1880

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

Leister D (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. Trends Genet 20:116–122

Lending CR, Larkins BA (1989) Changes in the zein composition of protein bodies during maize endosperm development. Plant Cell 1:1011–1023

Levanony H, Rubin R, Altschuler Y, Galili G (1992) Evidence for a novel route of wheat storage proteins to vacuoles. J Cell Biol 119:1117–1128

Li X, Franceschi VR, Okita TW (1993) Segregation of storage protein mRNAs on the rough endoplasmic reticulum membranes of rice endosperm cells. Cell 72:869–879

Llaca V, Messing J (1998) Amplicons of maize zein genes are conserved within genic but expanded and constricted in intergenic regions. Plant J 15:211–220

Lohmer S, Maddaloni M, Motto M, Di Fonzo N, Hartings H, Salamini F, Thompson RD (1991) The maize regulatory locus Opaque-2 encodes a DNA-binding protein which activates the transcription of the b-32 gene. EMBO J 10:617–624

Marzabal P, Busk PK, Ludevid MD, Torrent M (1998) The bifactorial endosperm box of gamma-zein gene: characterisation and function of the Pb3 and GZM cis-acting elements. Plant J 16:41–52

Matsumoto T, Wu JZ, Kanamori H, Katayose Y, Fujisawa M, Namiki N, Mizuno H, Yamamoto K, Antonio BA, Baba T, Sakata K, Nagamura Y, Aoki H, Arikawa K, Arita K, Bito T, Chiden Y, Fujitsuka N, Fukunaka R, Hamada M, Harada C, Hayashi A, Hijishita S, Honda M, Hosokawa S, Ichikawa Y, Idonuma A, Iijima M, Ikeda M, Ikeno M, Ito K, Ito S, Ito T, Ito Y, Iwabuchi A, Kamiya K, Karasawa W, Kurita K, Katagiri S, Kikuta A, Kobayashi H, Kobayashi N, Machita K, Maehara T, Masukawa M, Mizubayashi T, Mukai Y, Nagasaki H, Nagata Y, Naito S, Nakashima M, Nakama Y, Nakamichi Y, Nakamura M, Meguro A, Negishi M, Ohta I, Ohta T, Okamoto M, Ono N, Saji S, Sakaguchi M, Sakai K, Shibata M, Shimokawa T, Song JY, Takazaki Y, Terasawa K, Tsugane M, Tsuji K, Ueda S, Waki K, Yamagata H, Yamamoto M, Yamamoto S, Yamane H, Yoshiki S, Yoshihara R, Yukawa K, Zhong HS, Yano M, Sasaki T, Yuan QP, Shu OT, Liu J, Jones KM, Gansberger K, Moffat K, Hill J, Bera J, Fadrosh D, Jin SH, Johri S, Kim M, Overton L, Reardon M, Tsitrin T, Vuong H, Weaver B, Ciecko A, Tallon L, Jackson J, Pai G, Van Aken S, Utterback T, Reidmuller S, Feldblyum T, Hsiao J, Zismann V, Iobst S, de Vazeille AR, Buell CR, Ying K, Li Y, Lu TT, Huang YC, Zhao Q, Feng Q, Zhang L, Zhu JJ, Weng QJ, Mu J, Lu YQ, Fan DL, Liu YL, Guan JP, Zhang YJ, Yu SL, Liu XH, Zhang Y, Hong GF, Han B, Choisne N, Demange N, Orjeda G, Samain S, Cattolico L, Pelletier E, Couloux A, Segurens B, Wincker P, D'Hont A, Scarpelli C, Weissenbach J, Salanoubat M, Quetier F, Yu Y, Kim HR, Rambo T, Currie J, Collura K, Luo MZ, Yang TJ, Ammiraju JSS, Engler F, Soderlund C, Wing RA, Palmer LE, de la Bastide M, Spiegel L, Nascimento L, Zutavern T, O'Shaughnessy A, Dike S, Dedhia N, Preston R, Balija V, McCombie WR, Chow TY, Chen HH, Chung MC, Chen CS, Shaw JF, Wu HP, Hsiao KJ, Chao YT, Chu MK, Cheng CH, Hour AL, Lee PF, Lin SJ, Lin YC, Liou JY, Liu SM, Hsing YI, Raghuvanshi S, Mohanty A, Bharti AK, Gaur A, Gupta V, Kumar D, Ravi V, Vij S, Kapur A, Khurana P, Khurana JP, Tyagi AK, Gaikwad K, Singh A, Dalal V, Srivastava S, Dixit A, Pal AK, Ghazi IA, Yadav M, Pandit A, Bhargava A, Sureshbabu K, Batra K, Sharma TR, Mohapatra T, Singh NK, Messing J, Nelson AB, Fuks G, Kavchok S, Keizer G, Llaca ELV, Song RT, Tanyolac B, Young S, Il KH, Hahn JH, Sangsakoo G, Vanavichit A, de Mattos LAT, Zimmer PD, Malone G, Dellagostin O, de Oliveira AC, Bevan M, Bancroft I, Minx P, Cordum H, Wilson R, Cheng ZK, Jin WW, Jiang JM, Leong SA, Iwama H, Gojobori T, Itoh T, Niimura Y, Fujii Y, Habara T, Sakai H, Sato Y, Wilson G, Kumar K, McCouch S, Juretic N, Hoen D, Wright S, Bruskiewich R, Bureau T, Miyao A, Hirochika H, Nishikawa T, Kadowaki K, Sugiura M (2005) The map-based sequence of the rice genome. Nature 436:793–800

Mena M, Vicente-Carbajosa J, Schmidt RJ, Carbonero P (1998) An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. Plant J 16:53–62

Messing J (2009) Synergy of two reference genomes for the grass family. Plant Physiol 149:117–124

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant cell 15:809–834

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 8:1113–1130

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

Muench DG, Ogawa M, Okita TW (1999) The prolamins of rice. In: Shewry PR, Casey R (eds) Seed proteins. Kluwer Academic Publishers, Dordrecht, pp 93–108

Muth JR, Muller M, Lohmer S, Salamini F, Thompson RD (1996) The role of multiple binding sites in the activation of zein gene expression by Opaque-2. Mol Gen Genet 252:723–732

Nakase M, Hotta H, Adachi T, Aoki N, Nakamura R, Masumura T, Tanaka K, Matsuda T (1996) Cloning of the rice seed alpha-

globulin-encoding gene: sequence similarity of the 5′-flanking region to those of the genes encoding wheat high-molecular-weight glutenin and barley D hordein. Gene 170:223–226

Nakase M, Aoki N, Matsuda T, Adachi T (1997) Characterization of a novel rice bZIP protein which binds to the alpha-globulin promoter. Plant Mol Biol 33:513–522

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Okita TW, Krishnan HB, Kim WT (1988) Immunological relationships among the major seed proteins of cereals. Plant Sci 57:103–111

Onate L, Vicente-Carbajosa J, Lara P, Diaz I, Carbonero P (1999) Barley BLZ2, a seed-specific bZIP protein that interacts with BLZ1 in vivo and activates transcription from the GCN4-like motif of B-hordein promoters in barley endosperm. J Biol Chem 274:9175–9182

Onodera Y, Suzuki A, Wu CY, Washida H, Takaiwa F (2001) A rice functional transcriptional activator, RISBZ1, responsible for endosperm-specific expression of storage protein genes through GCN4 motif. J Biol Chem 276:14139–14152

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboobur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556

Payne PI, Lancaster GJ (1983) Catalogue of alleles for the complex gene loci Glu-A1, Glu-B1 and Glu-D1, which code for the high-molecular-weight subunits of glutenin in hexaploid wheat. Cereal Res Commun 11:29–35

Payne PI, Corfield KG, Blackman JA (1981) Correlation between the inheritance of certain high-molecular-weight subunits of glutenin and bread-making quality in progenies of six crosses of bread wheat. J Sci Food Agric 32:51–60

Richly E, Kurth J, Leister D (2002) Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. Mol Biol Evol 19:76–84

Schmidt RJ, Ketudat M, Aukerman MJ, Hoschek G (1992) Opaque-2 is a transcriptional activator that recognizes a specific target site in 22-kD zein genes. Plant Cell 4:689–700

Shewry PR, Tatham AS (1990) The prolamin storage proteins of cereal seeds: structure and evolution. Biochem J 267:1–12

Shewry PR, Tatham AS (1999) The characteristics, structure and evolutionary relationships of prolamins. In: Shewry PR, Casey R (eds) Seed proteins. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 11–33

Shewry PR, Miflin BJ, Kasarda DD (1984) The structural and evolutionary relationships of the prolamin storage proteins of barley, rye and wheat. Philos Trans R Soc Lond 304:297–308

Shewry PR, Tatham AS, Halford NG (1999) The prolamins of the Triticeae. In: Shewry PR, Casey R (eds) Seed Proteins. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 35–78

Shyur LF, Wen TN, Chen CS (1994) Purification and characterization of rice prolamins. Bot Bull Acad Sinica 35:65–71

Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. Proc Natl Acad Sci USA 100:9055–9060

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

Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) Close split of sorghum and maize genome progenitors. Genome Res 14:1916–1923

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, Aono R, Fujii Y, Habara T, Harada E, Kanno M, Kawahara Y, Kawashima H, Kubooka H, Matsuya A, Nakaoka H, Saichi N, Sanbonmatsu R, Sato Y, Shinso Y, Suzuki M, Takeda J, Tanino M, Todokoro F, Yamaguchi K, Yamamoto N, Yamasaki C, Imanishi T, Okido T, Tada M, Ikeo K, Tateno Y, Gojobori T, Lin YC, Wei FJ, Hsing YI, Zhao Q, Han B, Kramer MR, McCombie RW, Lonsdale D, O'Donovan CC, Whitfield EJ, Apweiler R, Koyanagi KO, Khurana JP, Raghuvanshi S, Singh NK, Tyagi AK, Haberer G, Fujisawa M, Hosokawa S, Ito Y, Ikawa H, Shibata M, Yamamoto M, Bruskiewich RM, Hoen DR, Bureau TE, Namiki N, Ohyanagi H, Sakai Y, Nobushima S, Sakata K, Barrero RA, Sato Y, Souvorov A, Smith-White B, Tatusova T, An S, An G, Oota S, Fuks G, Messing J, Christie KR, Lieberherr D, Kim H, Zuccolo A, Wing RA, Nobuta K, Green PJ, Lu C, Meyers BC, Chaparro C, Piegu B, Panaud O, Echeverria M (2008) The rice annotation project database (RAP-DB): 2008 update. Nucleic Acids Res 36:D1028–D1033

Ueda T, Waverczak W, Ward K, Sher N, Ketudat M, Schmidt RJ, Messing J (1992) Mutations of the 22- and 27-kD zein promoters affect transactivation by the Opaque-2 protein. Plant cell 4:701–709

Ueda T, Wang Z, Pham N, Messing J (1994) Identification of a transcriptional activator-binding element in the 27-kilodalton zein promoter, the -300 element. Mol Cell Biol 14:4350–4359

Vicente-Carbajosa J, Moose SP, Parsons RL, Schmidt RJ (1997) A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2. Proc Natl Acad Sci USA 94:7685–7690

Wang Z, Messing J (1998) Modulation of gene expression by DNA-protein and protein-protein interactions in the promoter region of the zein multigene family. Gene 223:333–345

Wang Z, Ueda T, Messing J (1998) Characterization of the maize prolamin box-binding factor-1 (PBF-1) and its role in the developmental regulation of the zein multigene family. Gene 223:321–332

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 Am genomes of wheat. Plant Cell 15:1186–1197

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

Wu CY, Suzuki A, Washida H, Takaiwa F (1998) The GCN4 motif in a rice glutelin gene is essential for endosperm-specific gene expression and is activated by Opaque-2 in transgenic rice plants. Plant J 14:673–683

Xu JH, Messing J (2006) Maize haplotype with a helitron-amplified cytidine deaminase gene copy. BMC Genet 7:52

Xu JH, Messing J (2008a) Diverged copies of the seed regulatory Opaque-2 gene by a segmental duplication in the progenitor genome of rice, sorghum, and maize. Mol Plant 1:760–769

Xu JH, Messing J (2008b) Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species. Proc Natl Acad Sci USA 105:14330–14335

Yunes JA, Cord Neto G, da Silva MJ, Leite A, Ottoboni LM, Arruda P (1994) The transcriptional activator Opaque2 recognizes two different target sequences in the 22-kD-like alpha-prolamin genes. Plant Cell 6:237–249