Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species

Jian-Hong Xu and Joachim Messing*

Waksman Institute, Rutgers, The State University of New Jersey, Piscataway, NJ 08854-8020

Communicated by Brian A. Larkins, University of Arizona, Tucson, AZ, July 29, 2008 (received for review April 9, 2008)

***Zea mays*, commonly known as corn, is perhaps the most greatly produced crop in terms of tonnage and a major food, feed, and biofuel resource. Here we analyzed its prolamin gene family, encoding the major seed storage proteins, as a model for gene evolution by syntenic alignments with sorghum and rice, two genomes that have been sequenced recently. Because a high-density gene map has been constructed for maize inbred B73, all prolamin gene copies can be identified in their chromosomal context. Alignment of respective chromosomal regions of these species via conserved genes allow us to identify the pedigree of prolamin gene copies in space and time. Its youngest and largest gene family, the alpha prolamins, arose about 22–26 million years ago (Mya) after the split of the *Panicoideae* (including maize, sorghum, and millet) from the *Pooideae* (including wheat, barley, and oats) and *Oryzoideae* (rice). The first dispersal of alpha prolamin gene copies occurred before the split of the progenitors of maize and sorghum about 11.9 Mya. One of the two progenitors of maize gained a new alpha zein locus, absent in the other lineage, to form a nonduplicated locus in maize after allotetraplodization about 4.8 Mya. But dispersed copies gave rise to tandem duplications through uneven expansion and gene silencing of this gene family in maize and sorghum, possibly because of maize's greater recombination and mutation rates resulting from its diploidization process. Interestingly, new gene loci in maize represent junctions of ancestral chromosome fragments and sites of new centromeres in sorghum and rice.**

allotetraploidy | centromere formation | chromosome structure | comparative genomics

The major seed storage proteins in maize and sorghum are the prolamins, hydrophobic proteins rich in proline and glutamine. The function of the prolamin gene family in maize and sorghum is of great environmental and economic importance because these crops represent a major biological nitrogen storage source and a renewable source of essential amino acids for the human and animal diet (1). Although humans prefer animal protein, most animal protein is derived from the grains of cereal and, to a lesser extent, vegetable species. Because mature seeds contain most of their amino acids stored in proteins rather than as free amino acids, a relative accumulation of protein from different gene copies can shift the balance of essential amino acids proportionally to their amino acid composition. Therefore, both gene copy number and the regulation of individual gene copies have a direct impact on the nutritional value of our most important food sources. From a breeding standpoint, it is also interesting to note that a major domestication locus of modern maize compared with its wild relative, teosinte, is the prolamin-binding factor, which controls the transcription of the entire prolamin gene family (2). Complicating breeding efforts is the multigenic nature of the genes that encode storage proteins. On the other hand, multigene families are quite common in plants (3). Considering only tandem gene duplications and not dispersed gene families, size estimates have been between 25% and 33% of the gene content in *Arabidopsis* and rice, the most complete plant genome sequences (4, 5).

Two critical questions can be asked: When did these genes arise in space and time, and which copy is expressed? The answer to the first question is also fundamental for our understanding of the whole genome duplication (WGD) event that formed the maize genome and its role in the divergence from its close relatives. Because maize and sorghum split just 11.9 million years ago (Mya) from a common progenitor (6) and about 50 Mya from the progenitor of rice (7), gene duplication events before and after the splits can trace the origin of chromosome lineages. To resolve the ancestry of gene copies, we used collinearity of contiguous chromosomal segments of diploid rice, diploid sorghum, and the two homoeologous regions of allotetraploid maize to correlate the position of a duplication event with individual nucleotide substitution rates, or Ks values (8).

Alignments of orthologous regions also provide information on general changes in chromosome structure. We need to account for differences in sixfold size ranges from 0.38 to 2.3 gigabases (Gb; billion bases), changes in chromosome numbers, and polyploidy. We know already that despite the fact that rice and sorghum have different chromosome numbers (12 and 10, respectively), they appear to be largely collinear (9). Maize, like sorghum, has 10 chromosomes but originated from 2 progenitors by allotetraploidization as recently as 4.8 Mya (6, 10). Alignment of rice and maize chromosomes indicates that maize temporarily consisted of 20 chromosomes. Through perhaps as many as 62 chromosomal breakages and fusions, modern maize shed 10 centromeres and restructured 10 larger chromosomes (11). Nonetheless, the polyploid origin of maize can be illustrated by the 2:1 relationship of most chromosomal regions of maize relative to sorghum and rice. Earlier alignments of chromosomal regions between maize and rice also demonstrated that genes missing in collinear regions of the rice genome not only could be found in other parts of the rice genome, but also could be duplicated in their new positions (12).

Consequently, we aligned all regions of the maize genome containing prolamin gene copies with orthologous regions that arose from the WGD and those of rice and sorghum. If gene copies were collinear in two chromosomes but absent in the others, then these alignments facilitated the discovery of sites that lost genes. We not only used the different chromosomal regions to define the ancestry of gene copies, but also discovered that genes were copied and inserted in dispersed sites of the genome, which were prone to the formation of new centromeres and chromosome fusions.

The second question relates to the fact that tandem gene arrays that arose recently often consist of highly conserved gene copies, so that transcripts may be mistakenly considered to be derived from a

Author contributions: J.-H.X. and J.M. designed research; J.-H.X. performed research; J.-H.X. and J.M. analyzed data; and J.-H.X. and J.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

*To whom correspondence should be addressed. E-mail: messing@waksman.rutgers.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0807026105/DCSupplemental.

Table 1. Gene copy number of prolamin genes in orthologous regions of rice, sorghum, and maize

Locus	Maize 1*	No.	Maize 2*	No.	Sorghum†	No.	Rice‡	No.	Origin
Alpha									
A1	Chromosome 4.2§	9	No homoeolog	0	Chromosome 5	2	Chromosome 11	0	Orthologous
A2	Chromosome 4.4	3	Chromosome 10.3¶	0	Chromosome 7	0	Chromosome 8	0	Paralogous
B	Chromosome 7.1	9	Chromosome 2.6	0	Chromosome 2	0	Chromosome 7	0	Paralogous
C1	Chromosome 4.2	14	No homoeolog	0	Chromosome 5	20	Chromosome 11	0	Orthologous
C2	Chromosome 4.3	1	Chromosome 2.9	0	Chromosome 5	0	Chromosome 11	0	Paralogous
D	Chromosome 1.6	5	Chromosome 3.4	0	Chromosome 8	1	Chromosome 12	0	Orthologous
Beta	Chromosome 6.1	1	Chromosome 8.5	0	Chromosome 9	1	Chromosome 5	0	Orthologous
Gamma	Chromosome 2.7	1	Chromosome 7.5	2	Chromosome 2	2	Chromosome 9	0	Orthologous
Delta	Chromosome 6.4	1	Chromosome 9.3	1	Chromosome 10	1	Chromosome 6	0	Orthologous

*Maize: inbred B73.
†Sorghum: Sorghum bicolor, cv Btx623.
‡Rice: *Oryza japonica*, cv Nipponbare.
§4.2: second bin on Chromosomeomosome 4.
¶Distal to junction.

single gene when they in fact are derived from several gene copies. Knowledge of sequence divergence between tandem gene copies then can be applied to cDNA libraries from different inbred lines to detect allelic transcripts. Indeed, it has been shown that the EST contig TUC02–07-16440.1 from ZmDB contains a mix of transcripts derived from five different gene copies (13). When we aligned gene members occurring in a tandem array and compared them with full-length cDNAs from another inbred line, we found that allelic gene copies were less divergent than different gene copies, illustrating the importance of identifying genes by their chromosomal position as well. We show here that systematic and genome-wide analysis of a single gene family provides new insights into polyploidy, the pattern and dynamics of gene duplications, and their affect on gene product accumulation.

Results and Discussion

Divergence of the Progenitor of Rice and Maize. In sorghum and maize, the prolamin genes fall into two groups (Table 1), one group encoded by one to three gene copies and the other group encoded by large sets of gene copies (14, 15). The latter group comprises the alpha prolamins (zein 1 [*z1*], kafirin 1 [*k1*]), and the former group comprises the beta, gamma, and delta prolamins (zein 2 [*z2*], kafirin 2 [*k2*]). All of these genes fall into distinct clusters (Fig. 1). Although the rice prolamin gene family (16, 17) has homologous genes to the beta, gamma, and delta prolamins, but not to the alpha prolamins, they differ from the sorghum and maize prolamin gene copies with respect to sizes and chromosomal positions. This becomes clear when orthologous regions are aligned via conserved genes. We first identified all maize BAC contigs or FPCs (11) containing zein genes. We then used linked non-zein genes to identify duplicated regions that arose from the WGD event in maize. Based on the conserved gene order, orthologous regions from both rice and sorghum were aligned with the two maize regions. A total of 31 chromosomal regions, varying in size from a few 100 kilobases (Kb; 1000 bases) to several megabases (Mb; million bases), were manually annotated and aligned via orthologous gene copies [Figs. 2 and 3*B* and supporting information (SI) Figs. S1–S6]. In these alignments of chromosomal regions, the alpha, beta, gamma, and delta zein genes have a counterpart in sorghum but not in rice, consistent with the results of phylogenetic analysis (Fig. 1). The beta and gamma zein genes appear to be the oldest prolamin genes, followed by the delta prolamins, which arose before the split of the *Panicoideae*.

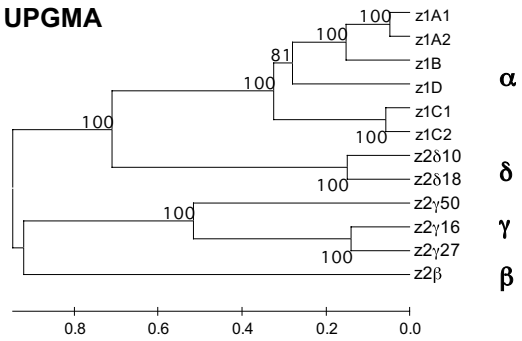

Fig. 1. Phylogenetic analysis of the prolamin genes. Phylogenetic analysis was performed as described in *Materials and Methods*. The graphic representation shows the results by the unweighted pair-group with arithmetic mean (UPGMA) method. The same analysis also was performed with the neighbor-joining method, which confirmed the results of the UPGMA method (data not shown).

Gene Loss After Allotetraploidization. Although the beta zein gene should be present on two maize chromosomes, it actually is present only on maize chromosome 6 (maize 6). It is present in the orthologous position on sorghum chromosome 9 (sorghum 9) but absent on the homoeologous region of maize 8 (Fig. S1). Therefore, it appears that it was present before allotetraplodization but then was deleted afterward. Indeed, it has been shown that losses of genes duplicated by allotetraploidization are quite common in the maize genome (18). In contrast to the beta zein gene, the gamma zein gene has retained copies on both homoeologous regions of the maize genome, maize 2 and 7, collinear with sorghum 2 (Fig. S2). Interestingly, sorghum 2 and maize 7 share tandemly duplicated copies with different sizes, encoding 50- and 27-kDa prolamins, whereas maize 2 has a single 16-kDa copy. Haplotypes exist in which the copy number of 27-kDa gamma zein genes on maize 7 can vary as well (19). Therefore, the 16-kDa gamma zein gene is likely the result of unequal crossing-over with loss of the 50- and 27-kDa zein genes on maize 2 after allotetraplodization. The third-oldest prolamin gene is the delta prolamin gene, which also has retained both zein copies after allotetraploidization and is present on maize 6 and 9, collinear with sorghum 10 (Fig. 2). These copies also differ in size, encoding a 10-kDa protein on maize 9 and an 18-kDa protein on maize 6. Because sorghum has a 10-kDa gene, we propose that the 18-kDa protein on maize 6 also arose by unequal crossing-over with loss of the duplicated 10-kDa zein gene after allotetraplodization.

Therefore, for all three types of prolamin genes, older gene copies were lost after new ones were produced either in tandem or by the WGD, complicating the reconstruction of the gene family chronology by phylogenetic analysis alone. All gene copies in B73 are intact, and cDNAs from multiple inbreds correspond to the B73 coding regions, indicating that these genes and their alleles are expressed (20). Gel electrophoresis of proteins has confirmed this, although the amount of proteins has been found to vary in different inbred lines (21).

Sugarcane Prolamins as a Reference for Maize Duplications. In the case of the gamma and delta zein genes, the corresponding kafirin

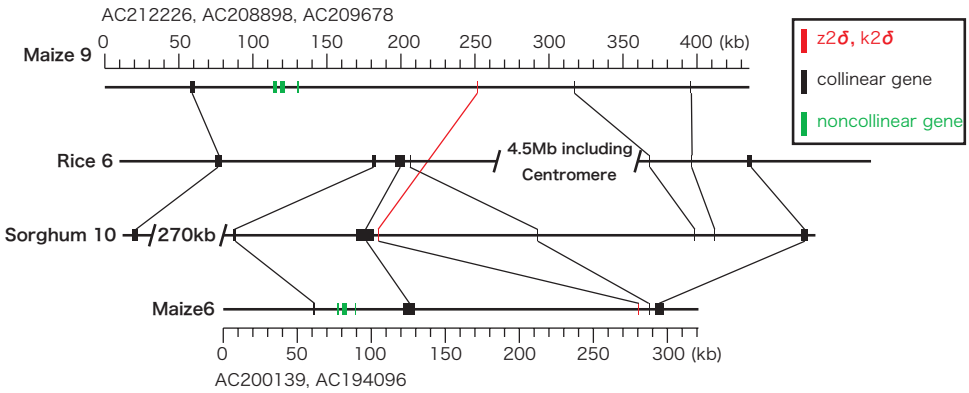

Fig. 2. New centromere formation. The duplicated regions from maize are aligned at the top and the bottom. Vertical bars connect conserved genes with insert providing a key. Clone numbers represent GenBank entries of overlapping maize BAC clones. In the middle, the sorghum and the rice orthologous regions are aligned. Because of the scale, large insertions are indicated in kb or Mb. The rice region contains the largest insertion, which is the centromere of rice chromosome 6. The key to the color-coding is shown in the *Inset*.

genes encoded the same size gene products as one of the maize homeologs, perhaps suggesting that out of the two progenitors that formed maize, one was closer to sorghum than the other. A similar hypothesis has been advanced with another set of duplicated loci in maize (10). But if one progenitor of maize were closer to sorghum, then three progenitors could not have arisen at the same time, as has been proposed in a more recent study (6). The difference between these two studies is that the earlier study had no orthologous sorghum and only a few orthologous rice sequences to compare with the duplicated maize genes. To seek an additional reference to these alignments, we compared the prolamin genes from sugarcane with those from sorghum and maize. Sugarcane split from the progenitor of sorghum 8–9 Mya after the split from the maize progenitors, and thus it is more distant to maize than

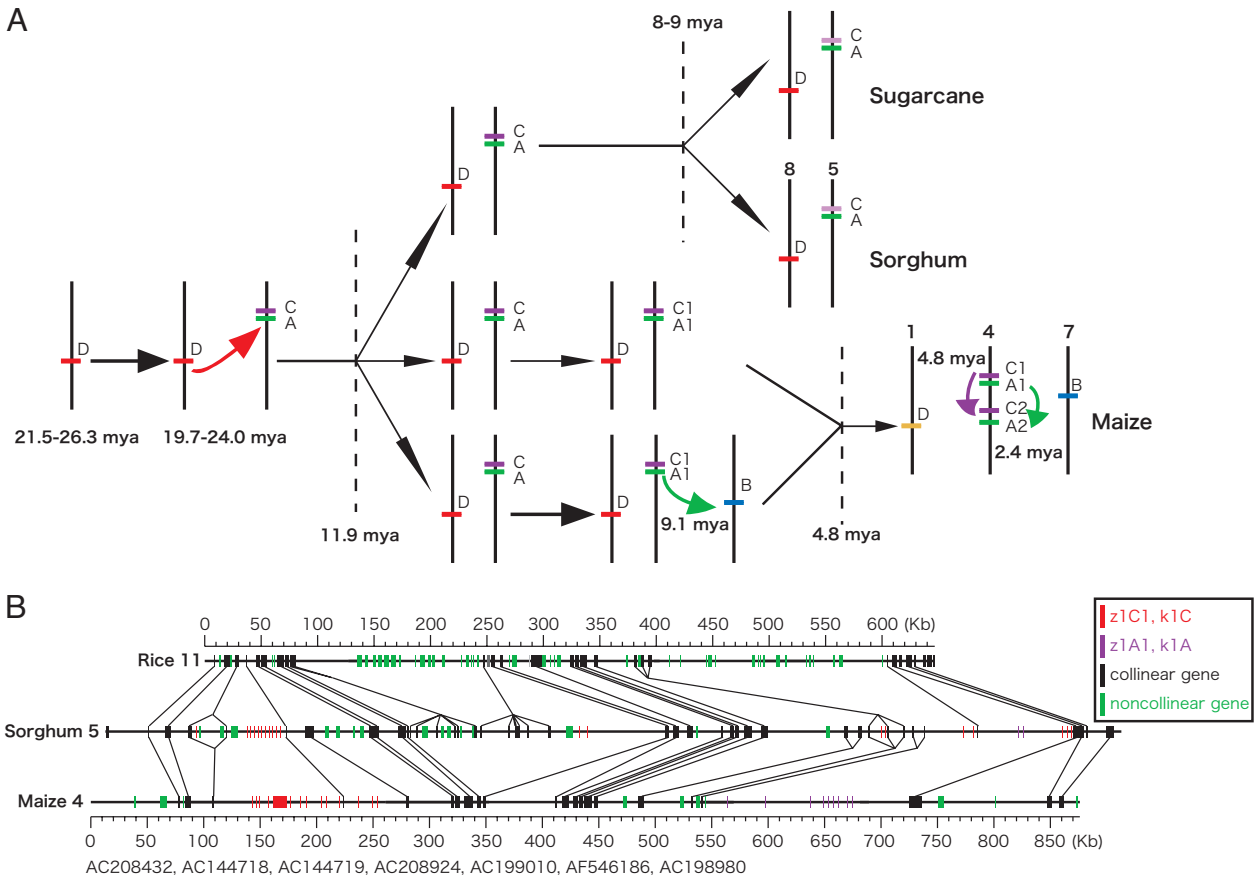

Fig. 3. (*A*) Gene pedigree of alpha prolamins in maize, sorghum, and sugarcane. Alpha prolamins fall into subfamilies, four in maize (A, B, C, and D) and three in sorghum (A, C, and D), based on distance analysis (see *Materials and Methods*). Increases in copy number and speciation are illustrated in progression from left to right. Vertical lines indicate chromosomal segments containing alpha-prolamin genes. Dotted vertical lines indicate the split of progenitor genomes. The progenitor of maize, sorghum, and sugarcane split about 11.9 Mya. One of the progeny genome then split into the progenitors of sorghum and sugarcane, whereas the two other ones hybridized to form maize by allotetraploidization about 4.8 Mya. The approximate timing and direction of gene copying events is indicated in Mya and by arrows. Genes also are color-coded to indicate orthologous and paralogous gene copying, as well as gene- silencing events. Note that we do not have a map of contiguous polamin genes in sugarcane, but only of EST resources. Their relationships are based solely on their sequence homology to kafirin genes. (*B*) Orthologous regions of the alpha prolamin *C1* and *A1* genes. To better follow the first duplication event in (*A*), the orthologous regions of the prolamin *C1* and *A1* genes of rice 11, sorghum 5, and maize 4 are shown below the diagram. For an explanation, see the legend of Fig. 2.

sorghum is to sugarcane (22). Because gamma and delta prolamins are found in homoeologous regions of maize, sorghum, and sugarcane, we can question whether the distance of sorghum genes falls between those of one maize homoeolog and sugarcane. That is not the case, however—the sorghum and sugarcane genes are always the closest pair (Fig. S7). Therefore, the two progenitors of maize split from the progenitor of sorghum and sugarcane before 8–9 Mya and must have diverged before that time. The divergence of the two parental lineages indicates that maize formed by allotetraploidization of two species that evolved independently of the sorghum lineage.

Collinearity Interrupted by a New Centromere. Alignment of the three older prolamin gene loci in maize and sorghum with rice has yielded an unexpected and interesting chromosomal feature. The chromosomal landscape of the delta prolamin genes, including rice 6, is rather gene-poor and has expanded intergenic regions. The distance of two genes on maize 9, about 200 kb apart, is twice the distance on sorghum 10. In rice, this collinear region is split by about 4.5 Mb, containing the centromere of the chromosome. The loss or gain of centromeres appears to be lineage-independent, because the relative distance to the centromere is different in all four chromosomes (the two from maize, the one from sorghum, and the one from rice). The new centromere on rice 6 arose by the insertion and tandem amplification of retrotransposons and centromere-specific repeats (23) after the split of the progenitor of rice, sorghum, and maize. Indeed, all species that deviate from a basic chromosome number of an ancestral genome had to form new centromeres. Interestingly, the centromeric region of rice 6 contains noncollinear genes as well (not shown), indicating that this region likely was hypersensitive to illegitimate recombination.

Alpha Prolamins Arose at the Same Time as C4 Photosynthesis. Although a limited degree of tandem gene duplications and duplications by polyploidization occurred with the older prolamin genes, the most extensive gene duplications occurred with the youngest prolamin gene subfamily, the alpha prolamins. Alpha prolamins fall into two size classes, 19- and 22-kDa proteins, that can be subdivided into four subfamilies, *z1A*, *z1B*, *z1C*, and *z1D*, based on sequence homology and copy number estimates by nucleic acid hybridization experiments in descending order (24). Because of polymorphisms, two-dimensional gel electrophoresis has facilitated the mapping of alpha prolamins to different chromosomal locations (21). We know that newly created dispersed gene copies also seeded extensive tandem gene duplications (Fig. S8), as was previously observed (albeit to a lesser degree) in noncollinear genes in rice (12). Based on Ks values, we selected the oldest copy in each alpha zein subfamily to conduct a phylogenetic analysis of dispersed gene copies in maize, sorghum, and sugarcane and reconstructed the dispersals of gene copies at different stages of speciation (Fig. 3*A*).

The oldest alpha prolamin gene is the prolamin *D* gene, which arose 21–26 Mya (Fig. 3*A*) after the *Panicoideae* subfamily of grasses split about 28 Mya into the tribes of the *Paniceae* (including the millets) and *Andropogoneae* (including maize, sorghum, and sugarcane). This divergence also coincided with the appearance of C4 photosynthesis, an improved system of food production (7). Before the progenitor of sorghum and maize split, a dispersed prolamin gene copy *C1* was created about 20–24 Mya, which in turn produced a tandem prolamin gene copy *A1*. In addition, both *A* and *C* alpha prolamin genes were duplicated tandemly after the split of maize and sorghum. Interestingly, one *z1A1* gene encodes a 22-kDa zein, and one *z1C1* gene encodes a 19-kDa zein, whereas the rest of the tandem array of *z1A* gene copies encode 19-kDa zeins and the *z1C* zein gene copies encode 22-kDa zeins, indicating a change in protein size during gene duplication either by deletion or insertion events in the coding regions.

Loss of Tandemly Duplicated Genes. Similar to the delta and gamma prolamin genes, tandem gene duplications appear to have resulted in the deletion of older copies. Here we can infer such deletions also from synteny. The *z1C1* (*azs22*;*13*) and the *k1C18* gene appear to be orthologous based on phylogenetic analysis, but not on synteny (Fig. 3*B*). Therefore, we suggest that the older *z1C1* gene was deleted after tandem duplication and insertion of sequences between both copies. Interestingly, we found a different maize haplotype (inbred BSSS53) in which *z1C1* (*azs22.18*) arose at the same time and in the same place as *z1C1* (*azs22.13*), but the segment containing the *z1C1* (*azs22.18*) copy was deleted in the B73 haplotype (25). This finding would not have been possible by mere clustering methods, but it necessitates the alignment of orthologous chromosomal segments.

Divergence of the Two Maize Progenitors. Although all three progenitors of maize and sorghum carried the alpha prolamin *D*, *A1*, and *C1* loci, one of them split into two additional progenitors preceding sorghum and sugarcane about 8–9 Mya (Fig. 3*A*). During the same time frame, one of the maize progenitors underwent a new copying event that created the alpha prolamin *B* locus. After polyploidization, that locus is present only on one of the two homoeologous regions of maize, maize 7 (Fig. S3), illustrating the divergence of the two progenitors of maize. After allotetraploidization, eight additional tandem copies were formed at the alpha prolamin *B* locus (Table 1). This locus is absent in sorghum, confirming the unique copying event in only one of the three progenitors of maize and sorghum. Interestingly, similar to the beta zein gene, one of the maize homoeologous regions lost the orthologous copy of the alpha prolamin *D* gene after allotetraploidization, as can be seen from the alignment of maize 1 and 3 (Fig. S4). In the case of the prolamin *C1* and *A1* genes, the entire region on maize 4 is one of the very few, if not the only, region that lacks a duplicated counterpart in the maize genome (11).

Another New Centromere and a Junction of Ancient Chromosome Fusion. There are two types of paralogous alpha zein gene copies in maize, some inserted into different chromosomal locations and others in tandem. A copy of the maize *z1C1* gene, the *z1C2* (*azs22*;*16*) gene, was inserted 20 cM closer to the centromere of maize 4 (Fig. S6). Whereas all other alpha kafirins and zeins were tandemly amplified, this alpha zein gene remained a single copy and constitutes the *floury-2* locus (26). An interesting feature is that the *floury-2* region resembles the delta prolamin region by having expanded intergenic regions, except that new centromere formation occurred on sorghum 5. There is also a very large insertion of 2 Mb in rice. Furthermore, this region appears to have segments of duplicated regions on maize 2, as if one duplicated homoeolog of ancient rice 11 was fragmented into much smaller pieces than the other homoeologs. As a result, the smaller fragments have landed in different places of the genome, thereby disrupting the extended syntenic alignments of maize and rice chromosomes. One additional 19-kDa zein gene locus originated from the *z1A1* locus (Fig. 3*A*). Its paralogous gene copies can be confirmed by their Ks values and their absence in the syntenous region of sorghum (Fig. S5). The *z1A2* gene was placed about 22 cM closer to the centromere of maize 4 and was subsequently duplicated in tandem. Interestingly, the *A2* genes appear to have translocated to the junction that fused ancient portions of rice 11 and rice 8 (Figs. S5 and S6). This junction lies between the *z1C2* and *z1A2* loci; it appears that only the portion distal to the *z1A2* genes was duplicated on maize 10 (Fig. S5). Obviously, we can question whether the chromosome break occurred in a chromosome structure that is inherent for illegitimate recombination.

Uneven Expansion in Homoeologous Regions of Maize. The alignment of orthologous regions by linked non-zein genes supports previous observations from smaller studies. A considerable number of

Table 2. Alpha-zein genes and their expression

Gene copy	Comment	mRNA	Length, bp	Expression	Gene copy	Comment	mRNA	Length, bp	Expression
z1A1–1	Prestop	87.2%	600	No	z1C1–1	Prestop	96.5%	801	No
z1A1–2	Intact	100%	804	Yes	z1C1–2	Insertion	2 SNPs*	801	No
z1A1–3	Intact	100%	705	Yes	z1C1–3	Iruncated		79	No
z1A1–4	Intact	100%	705	Yes	z1C1–4	Intact	1 SNP	801	Yes
z1A1–5	Intact	100%	705	Yes	z1C1–5	Prestop	1 SNP	801	Yes
z1A1–6	Intact	100%	705	Yes	z1C1–6	Prestop	97.0%	807	No
z1A1–7	Intact	100%	705	Yes	z1C1–7	Intact	3 SNPs	804	Yes
z1A1–8	No start codon	90.1%	706	No	z1C1–8	Intact	1 SNP	801	Yes
z1A1–9	Truncated		332	No	z1C1–9	Intact	1 SNP	801	Yes
z1A2–1	Intact	100%	702	Yes	z1C1–11	Prestop	93.9%	801	No
z1A2–2	Intact	100%	702	Yes	z1C1–12	Intact	100%	801	Yes
z1A2–3	Prestop	1 SNP	702	Yes	z1C1–13	Prestop		807	No
z1B1	Prestop	98.9%	726	Yes	z1C1–19	Intact	1 SNP	801	Yes
z1B2	Prestop	2 SNPs	726	Yes	z1C1–20	Prestop	1 SNP	798	Yes
z1B3	Prestop	97.7%	725	Yes	z1C2	Intact	2 SNPs	792	Yes
z1B4	Intact	1 SNP	723	Yes	z1D1	Truncated		201	No
z1B5	Prestop	94.9%	719	No	z1D2	Intact	100%	723	Yes
z1B6	Intact	100%	723	Yes	z1D3	Insertion	93.1%*	723	No
z1B7	Prestop	92.8%	719	No	z1D4	Intact	100%	726	Yes
z1B8	Prestop	93.3%	726	No	z1D5	Insertion	92.5%[a]	714	No
z1B9	Prestop	95.9%	723	No					

SNP, single nucleotide polymorphism.
*After removal of insertion.

non-zein genes are not collinear with rice and sorghum and likely are the result of gene insertions (green bars in Figs. 2 and 3*B*, and S1–S6). Actually, the noncollinear genes appear equally frequently in rice and sorghum. Moreover, genes that are not tandemly duplicated in maize might be tandemly duplicated in rice and sorghum (Figs. 2, 3*B*, and S1–S6). Therefore, the events described here for the prolamin gene family likely represent a general pattern of gene families in plants. Furthermore, expansion of chromosomal regions in maize relative to the smaller genomes of rice and sorghum do not occur evenly in the genome. The homoeologous regions frequently differ in length substantially (Figs. 2, 3*B*, and S1–S6), validating an earlier observation for a single region (27). Expansion of the maize genome after allotetraploidization was due mainly to retrotransposition; therefore, retrotransposition contributes to uneven expansion in the intergenic regions of the two homoeologous chromosomal segments, which could have played an important role in the diploidization of maize. It also can be envisioned that unequal crossing-over between two closely linked copies of the same retrotransposon yielded deletions, including genes, which may have been tolerated because of the copy present in the other homoeologous region. Moreover, during the time of tandem duplication of prolamin genes, the maize genome doubled in size by retrotransposition (8). Whether these two types of chromosomal expansion have any cause in common except the requirement for chromosome breaks is less clear, however.

Gene Expression and Silencing. What is the function of gene duplications? Part of the answer lies in the expression of duplicated genes. But gene expression analysis of gene families is hampered by the fact that recently duplicated genes are so homologous that nucleic acid hybridization and sequence alignments often are insufficient to infer which gene member is actually expressed from mRNA samples. As described earlier, this is true for some of the tandem clusters of the alpha prolamin genes as well. To address this problem, we constructed polymorphism grids of coding regions of all alpha zein gene copies in maize inbred B73 (Table S1). As can be seen, sequence conservation within each subfamily can vary between 75% and 99%; therefore, the assignment of an mRNA species to a single gene member is more stringent for recent gene duplications (see *Materials and Methods*). For instance, we found allelic mRNAs in three inbreds for *z1A1–5* and *z1A1–6*, although these two gene copies were 99% homologous. Full-length cDNAs from different inbreds were 100% homologous to *z1A1–5* or to *z1A1–6* (Table 2). These findings illustrate the need for a polymorphism grid of a gene family for gene expression analysis.

Interestingly, six mRNAs contained short reading frames and, if translated, would make shorter peptides. However, mRNAs for those clearly accumulated at low concentrations, consistent with an earlier study of an mRNA with a premature stop codon (28). It has been suggested that mRNAs with premature stop codons decrease mRNA stability, which would explain their relative lower concentration (29). Given their shorter reading frame, it is not surprising that those genes exhibited greater allelic diversity than genes with full-length reading frames (Table 2). Surprisingly, promoters of these genes were kept intact. Either these genes serve as a reserve or they indeed have a new function, perhaps as processed regulatory small RNAs. There is evidence for the former possibility, because one of the paralogs of the 22-kDa zein gene, *azs22;8*, has an allele with a premature stop codon in inbred W22 and one that is intact in inbreds BSSS53 and B73 (25). Nonetheless, out of 41 alpha zein genes, only 25 appeared to be expressed, whereas out of 23 kafirins, 19 were expressed at the transcript level (Table S2). This suggests that protein accumulation during seed development is less a function of gene dosage than the relative expression of individual gene copies. Strikingly, maize underwent more extensive changes in gene expression compared with sorghum.

There is clearly a trend toward gene expression shifting from the older copies to the newer copies. Mutant analysis has shown that 19- and 22-kDa zein genes are under the regulation of different transacting factors (1), illustrating that copying events result in the differential regulation of the same gene product as described for other duplicated genes (30). Interestingly, even within the 22-kDa zein gene cluster, one haplotype with the youngest gene copies are also under the regulation of different *trans*-acting factors (25). The beta, gamma, and delta zein genes are expressed in most inbreds, although the amounts of protein accumulated in mature seeds vary widely (1). However, except for the variation in the copy number of the 27-kDa gamma zein gene, quantitative differences in expression appear to be due largely to alleles of *trans*-acting factors. Interestingly, an allele of a *trans*-acting factor specific for the delta zeins, *dzr1*+Mo17, is imprinted, providing a parental effect on protein accumulation in seeds of reciprocal hybrid crosses (31).

Possible Mechanisms. Copying events of the prolamin gene family appear to have occurred in two phases. Early copying events are separated in a lineage-dependent manner by speciation, validating the use of syntenic data in phylogenetic analysis. Furthermore, older copies are frequently silenced or lost, particularly after tandem duplication, indicating rapid changes in the control of gene expression. Such a rapid mode of action likely reflects plants' need to cope with environmental changes because of their immobility compared with animals. Therefore, the main purpose of gene copying is likely a change

in the regulation of the same gene product. From a mechanistic standpoint, the simplest explanation is that far-distance placement occurs through translocation of an extrachromosomal copy and tandem copies by unequal crossover. Because these genes are expressed in endosperm tissue, we are less inclined to assume RNA intermediates in gene copy dispersal, because somatic events in endosperm are not transmissible through the germline. There are also no indications that dispersed copies arose as retroposons, as in animal systems (32). Prolamin genes lack introns, which usually provide the distinguishing feature between retroposons and donor genes, but we cannot detect the polyA stretch that would be expected at the 3′ end. On the other hand, it has been shown that recombination at the gamma zein locus occurred in seedlings, producing a different gene copy that transmitted into the next generation (19). Several pathways, including early somatic events, can lead to rapid expansion of gene families and their adaptation to differential control of gene expression. Moreover, gene families provide an invaluable molecular link between ancient and modern chromosomes and for the origin of species.

Materials and Methods

Databases. The rice genome has been completely sequenced (5). The sorghum genome also has been sequenced and assembled in contiguous sequence information (www.phytozome.net/cgi-bin/gbrowse/sorghum/). Although the maize genome is still being sequenced by a BAC-by-BAC strategy, the BAC sequences available are anchored to the maize genetic map in form of fingerprinted contigs (11).

Sequence Annotation and DNA Sequencing. Sequences of maize and rice BAC clones were downloaded from GenBank, and sorghum sequences were obtained from the Joint Genome Institute website (http://www.phytozome.net/cgi-bin/gbrowse/sorghum/). Repetitive DNA was masked with RepeatMasker before gene prediction methods were used, as described previously (27). Predicted genes were manually annotated using EST and protein sequence resources. Standard features of transposable elements were used to determine their position and endpoints. Many maize BAC clones containing zein genes were finished sequences from our own laboratory. Clones from the maize-sequencing project consisted of unordered contigs that were linked with unspecified nucleotides by inserting Ns in gaps of unknown length. We could experimentally verify that their order often was incorrect, but that collinearity with sorghum and rice could be used to reorder these contigs. In one case, AC208898 had a gap that contained the 10-kDa delta zein gene. This clone was retrieved from our library, and with two primer extensions, this gap was closed and indeed contained the 10-kDa delta zein gene. Collinear regions were aligned in a pairwise fashion using dot plot graphics from Lasergene.

Phylogenetic Analysis. Phylogenetic analyses were performed by multiple alignments of nucleotide or amino acid sequences of the entire coding regions using the Clustal_X program (33). Distance analysis and Ks values were calculated with the MEGA3 program using either the neighbor-joining or UPGMA method to draw phylograms (34).

Expression Analysis. Z2 genomic DNAs were matched with cDNAs at the National Center for Biotechnology Information website. Because of the large number of *z1* genes and their extensive sequence conservation, a separate collection of genomic sequences was formed. To determine sequence conservation, a two-dimensional grid was composed, matching each coding region in a pairwise fashion and expressing sequence homology in the percentage of nucleotides over the entire length of the cDNA. Because all zeins have relatively short mRNAs and genomic DNA with no introns, each genomic coding region can be directly aligned with cDNAs and subjected to basic local alignment search tool (BLAST) analysis of various cDNA collections (35). We drew on 364,383 ESTs from B73 mixed-tissue cDNA libraries (silks, husks, ears, pollen, shoot tips, leaf, root tips, whole seed, embryo); 6732 ESTs from B73 endosperm cDNA libraries harvested at 11, 13, 15, 18, 21, 24, 27, 29, 30, 35, and 40 days after pollination (DAP) (20); 30,531 ESTs from F352 endosperm cDNA libraries harvested at 10, 15, and 20 DAP (36); and 5326 ESTs from W22 endosperm cDNA libraries harvested at 4–6 DAP and 7–23 DAP (37). Alignments of genomic DNA and cDNA were than used to calculate sequence conservation in percentages of nucleotide homologies. A similar analysis was done recently for the 22-kDa kafirin cluster (38). All kafirin genes are so diverged that a BLAST analysis could easily match genomic sequences with cDNA sequences.

ACKNOWLEDGMENTS. We thank Dr. Hugo Dooner for his critical review. This research was supported by Grant DE-FG05–95ER20194 from the U.S. Department of Energy (to J.M.).

1. Gibbon BC, Larkins BA (2005) Molecular genetic approaches to developing quality protein maize. *Trends Genet* 21:227–233.
2. Jaenicke-Despres V, *et al.* (2003) Early allelic selection in maize as revealed by ancient DNA. *Science* 302:1206–1208.
3. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16:1667–1678.
4. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.
5. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800.
6. Swigonova Z, *et al.* (2004) Close split of sorghum and maize genome progenitors. *Genome Res* 14:1916–1923.
7. Kellogg EA (2001) Evolutionary history of the grasses. *Plant Physiol* 125:1198–1205.
8. Messing J, Bennetzen J (2008) Grass genome structure and evolution. *Genome Dynamics* 4:41–56.
9. Bowers JE, *et al.* (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. *Proc Natl Acad Sci USA* 102:13206–13211.
10. Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc Natl Acad Sci USA* 94:6809–6814.
11. Wei F, *et al.* (2007) Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* 3:e123.
12. Lai J, *et al.* (2004) Gene loss and movement in the maize genome. *Genome Res* 14:1924–1931.
13. Song R, Messing J (2002) Contiguous genomic DNA sequence comprising the 19-kD zein gene family from maize. *Plant Physiol* 130:1626–1635.
14. DeRose R (1989) Characterization of the kafirin gene family from sorghum reveals extensive homology with zein from maize. *Plant Mol Biol* 12:245–256.
15. Song R, Llaca V, Messing J (2002) Mosaic organization of orthologous sequences in grass genomes. *Genome Res* 12:1549–1555.
16. Yamakawa H, Hirose T, Kuroda M, Yamaguchi T (2007) Comprehensive expression profiling of rice grain filling-related genes under high temperature using DNA microarray. *Plant Physiol* 144:258–277.
17. Lin H, *et al.* (2008) Characterization of paralogous protein families in rice. *BMC Plant Biol* 8:18.
18. Messing J, *et al.* (2004) Sequence composition and genome organization of maize. *Proc Natl Acad Sci USA* 101:14349–14354.
19. Das OP, Levi-Minzi S, Koury M, Benner M, Messing J (1990) A somatic gene rearrangement contributing to genetic diversity in maize. *Proc Natl Acad Sci USA* 87:7809–7813.
20. 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.
21. Wilson C (1989) Linkages among zein genes determined by isoelectric focusing. *Theor Appl Genet* 77:217–226.
22. Jannoo N, *et al.* (2007) Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J* 50:574–585.
23. Ma J, Jackson SA (2006) Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res* 16:251–259.
24. Heidecker G, Messing J (1986) Structural analysis of plant genes. *Annu Rev Plant Physiol* 37:439–466.
25. 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.
26. Coleman CE, *et al.* (1997) Expression of a mutant alpha-zein creates the *floury2* phenotype in transgenic maize. *Proc Natl Acad Sci USA* 94:7094–7097.
27. Bruggmann R, *et al.* (2006) Uneven chromosome contraction and expansion in the maize genome. *Genome Res* 16:1241–1251.
28. Liu CN, Rubenstein I (1993) Transcriptional characterization of an alpha-zein gene cluster in maize. *Plant Mol Biol* 22:323–336.
29. van Hoof A, Green PJ (1996) Premature nonsense codons decrease the stability of phytohemagglutinin mRNA in a position-dependent manner. *Plant J* 10:415–424.
30. Cusack BP, Wolfe KH (2007) When gene marriages don't work out: divorce by subfunctionalization. *Trends Genet* 23:270–272.
31. Chaudhuri S, Messing J (1994) Allele-specific parental imprinting of dzr1, a posttranscriptional regulator of zein accumulation. *Proc Natl Acad Sci USA* 91:4867–4871.
32. Vinogradov AE (2002) Growth and decline of introns. *Trends Genet* 18:232–236.
33. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25:4876–4882.
34. Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* 5:150–163.
35. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
36. Verza NC, *et al.* (2005) Endosperm-preferred expression of maize genes as revealed by transcriptome-wide analysis of expressed sequence tags. *Plant Mol Biol* 59:363–374.
37. Lai J, *et al.* (2004) Characterization of the maize endosperm transcriptome and its comparison to the rice genome. *Genome Res* 14:1932–1937.
38. Song R, Segal G, Messing J (2004) Expression of the sorghum 10-member kafirin gene cluster in maize endosperm. *Nucleic Acids Res* 32:e189.