Dynamic Gene Copy Number Variation in Collinear Regions of Grass Genomes

Jian-Hong Xu,† Jeffrey L. Bennetzen,2 and Joachim Messing1,*

1Waksman Institute of Microbiology, Rutgers, The State University of New Jersey
2Department of Genetics, University of Georgia
†Present address: Institute of Crop Science, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang, China
*Corresponding author: E-mail: messing@waksman.rutgers.edu.
Associate editor: Neelima Sinha

Abstract

A salient feature of genomes of higher organisms is the birth and death of gene copies. An example is the alpha prolamin genes, which encode seed storage proteins in grasses (Poaceae) and represent a medium-size gene family. To better understand the mechanism, extent, and pace of gene amplification, we compared prolamin gene copies in the genomes of two different tribes in the Panicoideae, the Paniceae and the Andropogoneae. We identified alpha prolamin (setarin) gene copies in the diploid foxtail millet (Paniceae) genome (490 Mb) and compared them with orthologous regions in diploid sorghum (730 Mb) and ancient allotetraploid maize (2,300 Mb) (Andropogoneae). Because sequenced genomes of other subfamilies of Poaceae like rice (389 Mb) (Ehrhartoideae) and Brachypodium (272 Mb) (Pooideae) do not have alpha prolamin genes, their collinear regions can serve as “empty” reference sites. A pattern emerged, where genes were copied and inserted into other chromosomal locations followed by additional tandem duplications (clusters). We observed both recent (species-specific) insertion events and older ones that are shared by these tribes. Many older copies were deleted by unequal crossing over of flanking sequences or damaged by truncations. However, some remain intact with active and inactive alleles. These results indicate that genomes reflect only a snapshot of the gene content of a species and are far less static than conventional genetics has suggested. Nucleotide substitution rates for active alpha prolamin genes were twice as high as for low copy number beta, gamma, and delta prolamin genes, suggesting that gene amplification accelerates the pace of divergence.

Key words: copy number variation, gene duplication/loss, collinear regions, grass genomes, prolamin gene family.

Introduction

Prolamins are the major seed storage proteins in most grass species. Although each grass species has its own name for the prolamin gene family and subfamily (fig. 1), all prolamins can be divided into three groups (I, II, and III) based on amino acid sequence conservation (Xu and Messing 2009). Group I includes alpha, delta, and Ory10 prolamins (fig. 1, in black). Group II is the largest group and comprises beta, gamma, Ory13, Ory16, and S-rich prolamins (fig. 1, in red). Group III consists of HMW-like prolamins that are distinct from all the others (fig. 1, in blue). This classification also reflects the taxonomic division of the grasses into different subfamilies (fig. 1). All of them share gamma prolamins, but Pooidae (wheat, barley, and Brachypodium) also have HMW-glutenins, important for the baking quality of bread wheat. In the Panicoideae, groups I and II are present, with group I only detected so far in this subfamily that utilizes C4 photosynthesis (foxtail millet, maize, sorghum, and switchgrass). Prolamins accumulate in the endoplasmic reticulum (ER) and are deposited into protein bodies (PBs) of the seed endosperm. In maize (Zea mays), alpha and delta prolamins accumulate in the inner core of PBs, which are surrounded by gamma and beta prolamins (Lending and Larkins 1989). In a transgenic system using tobacco, alpha and delta prolamins require gamma and beta prolamins to stably accumulate, confirming the subfunctionalization that divergent proteins have assumed in seed architecture (Coleman et al. 1996; Bagga et al. 1997).

Although alpha prolamins in maize and sorghum (Sorghum bicolor), called zeins and kafirins respectively, are composed of multiple gene clusters, beta, gamma, and delta prolamins exist as single or low copy number loci. In maize and sorghum, there are a respective 41 and 23 alpha prolamin genes in the reference genomes B73 and Bxx623 that can be distinguished as 19 kDa (z1A, z1B, z1D, and k1A) and 22 kDa (z1C and k1C) subgroups, located in five and three chromosomal loci, respectively (Song and Messing 2002, 2003; Xu and Messing 2008b). However, the higher number of gene copies in maize is not due to the whole genome duplication (WGD) about 48 Ma (Swigonova et al. 2004) but rather to recent gene amplification. Furthermore, the z1B locus on chromosome 7S arose about 9 Ma only in one of the two progenitors of maize. After hybridization of the two progenitors, tandem duplications occurred in only one of the two homoeologous regions.

The seeds of grasses include a sizeable endosperm, containing the bulk of the nutrients. In maize, the endosperm is ~85% starch and ~10% protein. Of the total protein, nearly 70% is provided by prolamins, and alpha prolamins account for about 70% of total prolamins. Because of alpha prolamin amino acid composition and abundance, some
essential amino acids like lysine are deficient in the endosperm, necessitating supplementation in corn-based diets. In contrast, mutants that decrease the expression of α-zein genes increase lysine levels in maize seed (Gibbon and Larkins 2005). The same effect can be achieved with RNA interference (RNAi) (Segal et al. 2003; Huang et al. 2004, 2006; Kawakatsu et al. 2010).

To better understand the process of subfunctionalization, value can be added by studying different time frames of genome divergence. Genomes of several grass taxa are now becoming available. The rice (Oryza sativa) genome sequence in 2005 (International Rice Genome Sequencing Project 2005) was followed by the sorghum genome sequence (Paterson et al., 2009), then the maize (Schnable et al., 2009) and Brachypodium (Brachypodium distachyon) (International-Brachypodium-Initiative 2010) sequences. All the genomic regions in the maize genome containing prolamin genes were sequenced for two inbred lines (B73 and BSSS53) before a draft genome was assembled for inbred B73 (Song and Messing 2002, 2003). Sorghum and Brachypodium were sequenced by whole genome shotgun (WGS) sequencing with good assemblies of intact pseudomolecules, which made it possible to identify all prolamin gene copies in these species as well (Xu and Messing 2002b, 2009; Larre et al. 2010).

Foxtail millet (Setaria italica) is a diploid species with nine chromosomes in the Paniceae tribe of the same grass subfamily Panicoideae as sorghum and maize, which are in the Andropogoneae tribe, and has a relatively small genome (1C ≈ 490 Mb) (Doust et al. 2009). It is one of the oldest cultivated foods and was a staple in China before rice. Because it is gluten-free, it is a healthy alternative to food products derived from wheat. Recently, the Joint Genome Institute (JGI) sequenced the cultivar Yugu1 (http://www.phytozome.net/foxtailmillet). Here, we used the 8.3× foxtail millet WGS sequence data to identify all alpha setarin genes and compared their orthologous regions to sorghum, maize, rice and Brachypodium.

**Materials and Methods**

**Plant Materials**

Maize inbred B73, sorghum inbred Btx623, and rice cultivar Nipponbare were from our own stocks. Teosinte line PI384062 and foxtail millet line Ames27267 were obtained from the North Central Regional Plant Introduction Station (NCRPIS), USDA-ARS.

**Prolamin Protein Extraction and Sodium Dodecyl Sulfate–Polyacrylamide Gel Electrophoresis**

Prolamin proteins were extracted from dehusked seeds of all plant materials with a method described previously (Wu and Messing 2010). The zein concentration was measured by ND-1000 Spectrophotometer (Nanodrop), and a total of 0.2 μg of total prolams were separated on a 15% sodium dodecyl sulfate–polyacrylamide gel electrophoresis.

**Databases for Grass Genome Sequences**

Sequence Analysis
Each zein and kafirin (z1A, z1B, z1C, and z1D; k1A and k1C) amino acid sequence was used to do tBlastN against the foxtail millet genome sequence assembly (E value cutoff e-6). All candidate genes were extracted from the assembly with 5 kb flanking sequences and manually annotated. About 100 kb−1 Mb chromosomal regions of alpha setarin gene loci were annotated by Fgenesh software (http://linux1.softberry.com/berry.phtml) for gene models, and repeats sequences were masked by Repeatmasker (http://www.repeatmasker.org/). The gene models were further supported by comparison with orthologous genes from rice, sorghum, maize, and Brachypodium.

Phylogenetic Analysis
Nucleotide and protein sequences were aligned using Clustalw at default settings with manual modification. The phylogenetic analyses were conducted using the Maximum Likelihood (ML) method based on the Kimura 2-parameter model in the MEGA4 program (Kimura 1980; Tamura et al. 2007). The tree with the highest log likelihood is shown. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. Phylogeny was tested by the bootstrap method with 1000 replications. Synonymous substitution (Ks) was estimated using the MEGA4 program (Tamura et al. 2007), and substitution rates (r) for x2 and x3 prolamin genes were calculated based on the sorghum and maize split ~11.9 Ma (Swigonova et al. 2004). The divergence time (r) was calculated with the same method as described previously (Xu and Messing 2006).

Results
Identification and Analysis of Alpha Prolamin Genes in Foxtail Millet
Alpha zein and kafirin subgroup (z1A, z1B and z1C, z1D; k1A and k1C; loci A, B, C, and D refer to loci in maize and sorghum, fig. 1) amino acid sequences were used as query in a tBlastN analysis against the 8.3× foxtail millet genome sequence assembly. Fifteen copies of alpha prolamin genes, called alpha setarin genes following standard nomenclature rules, were identified in five chromosomal locations (table 1). Based on their sequence similarity to zein and kafirin genes, alpha setarin genes were named s1D1-6 and s1C1-9. Phylogenetic analysis confirmed that they were clustered into two groups (fig. 2), corresponding to the C and D subgroups in maize and sorghum. Nine copies of s1C were located at two loci, one with three copies, and the other with six copies, whereas six copies of the s1D subgroup were represented as one, two, and three copies at three different sites.

A consensus sequence was generated from the nine s1C gene copies and a second one was generated from the six s1D gene copies (supplementary fig. 1, Supplementary Material online). The size of s1C genes varies between 600 and 781 bp. The s1C1-3 genes share 120 bp tandem duplications, and s1C3 has an additional three deletions (a 72 bp deletion and two 6 bp deletions of the sequence CCCCCCG). All these deletions maintain the protein reading frame. The s1C2 gene has 58 bp repeats, resulting in a frame shift and a premature stop codon. The s1C7 gene has one 3 bp insertion (TAG), which generates a premature stop codon, whereas s1C9 has mutations from TTG to TAA to produce a premature stop codon. The s1C4-6 gene has one 6 bp insertion (CCCCCG) and two small deletions (CAA and CTTATG), whereas s1C9 has a 15 bp deletion (fig. 3). The nucleotide sequences of s1C4 and s1C5 are identical, suggesting that the s1C gene was recently duplicated. The longest gene of the s1D subgroup is 798 bp (s1D2) and has internal tandem duplication events of 135 bp. Two 3 bp deletions (CAA, CTT) were observed in s1D1 and s1D2, whereas s1D3 and s1D4 share an additional 30 bp deletion (fig. 3). The data indicate that internal rearrangements of the C subgroup setarin genes have occurred more frequently than for the D subgroup setarin genes.

Because three s1C gene copies have premature stop codons (s1C2, s1C7, and s1C9), only the remaining six copies of s1C and s1D genes were analyzed for their amino acid composition (supplementary tables 1 and 2, Supplementary Material online). The predicted proteins encoded by s1C and s1D genes are very similar, with four amino acids (cysteine, lysine, glutamine, and aspartic acid) at very low levels (<1%). The essential amino acid lysine was only 0.48% of the encoded protein. Therefore, foxtail millet seed is predicted to be lysine deficient, similar to maize and sorghum (Naren and Virupaksha 1990). The molecular weight for both subgroups (C and D) of alpha setarins was observed to be very similar, migrating with the size of 16 kDa, while alpha zeins and kafirins have the sizes of 19 and 22 kDa (fig. 4) (supplementary tables 1 and 2, Supplementary Material online).

Phylogenetic Analysis of Alpha Prolamin Genes in the Panicoideae
Excellent evolutionary conservation of gene order across the grasses (Gale and Devos 1998) allows the discovery of orthologous and paralogous gene copies. Based on

<table>
<thead>
<tr>
<th>Setarin</th>
<th>Supercontig</th>
<th>Position</th>
<th>Size</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1C1</td>
<td>PROV_5</td>
<td>33756447–33755715</td>
<td>723</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C2</td>
<td>PROV_5</td>
<td>33760626–33759846</td>
<td>781</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C3</td>
<td>PROV_5</td>
<td>33764651–33764028</td>
<td>624</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C4</td>
<td>PROV_5</td>
<td>33788235–33787621</td>
<td>615</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C5</td>
<td>PROV_5</td>
<td>33790897–33790283</td>
<td>615</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C6</td>
<td>PROV_5</td>
<td>33793561–33792947</td>
<td>615</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C7</td>
<td>PROV_5</td>
<td>14243003–14242386</td>
<td>618</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C8</td>
<td>PROV_5</td>
<td>14245682–14245068</td>
<td>615</td>
<td>Intact</td>
</tr>
<tr>
<td>S1C9</td>
<td>PROV_5</td>
<td>14268773–14268174</td>
<td>600</td>
<td>Intact</td>
</tr>
<tr>
<td>S1D1</td>
<td>PROV_1</td>
<td>31465344–31466006</td>
<td>663</td>
<td>Intact</td>
</tr>
<tr>
<td>S1D2</td>
<td>PROV_1</td>
<td>31469451–31470248</td>
<td>798</td>
<td>Intact</td>
</tr>
<tr>
<td>S1D3</td>
<td>PROV_1</td>
<td>43414426–43413788</td>
<td>639</td>
<td>Intact</td>
</tr>
<tr>
<td>S1D4</td>
<td>PROV_1</td>
<td>43435265–43434627</td>
<td>639</td>
<td>Intact</td>
</tr>
<tr>
<td>S1D5</td>
<td>PROV_1</td>
<td>43446137–43445464</td>
<td>669</td>
<td>Intact</td>
</tr>
<tr>
<td>S1D6</td>
<td>PROV_6</td>
<td>12310665–12311225</td>
<td>561</td>
<td>Intact</td>
</tr>
</tbody>
</table>

* In base pair for the predicted protein-encoding portion of the gene.
evolutionary distance, one can then reconstruct the time of gene duplication, including those non-collinear events associated with insertion of paralogs at non-tandem sites. We now can use the genomes of foxtail millet inbred Yugu1, sorghum inbred Btx623, and maize inbred B73 for this purpose. They contain 15, 23, and 41 copies of alpha prolamin genes, respectively. Eliminating four truncated genes leaves 75 (including those with premature stop codons) as input for the creation of a phylogenetic tree. This tree was generated by the ML method (fig. 2). All z1C, k1C, and s1C genes clustered together, indicating that the C subgroup of alpha prolamin genes originated from the same donor. Gene copies of z1A, z1B, and k1A subgroup clustered into one clade, consistent with previous results (Xu and Messing 2008b). All paralogous copies of z1D and s1D genes were clustered with the founder of this family, k1D, in species-specific subclades at the root of the tree. The k1D gene has previously determined as the oldest alpha prolamin gene in a phylogenetic tree rooted with the further distant globulin storage protein genes (Xu and Messing 2009).

The original naming of alpha prolamins was based on sequence homologies of cDNAs rather than phylogenetic criteria, although they are located in collinear regions (Heidecker and Messing 1986). Because of transposition events, some phylogenetic relationships are not predicted by current chromosomal positions. To conform to phylogenetic data across species, a new nomenclature is therefore proposed for alpha prolamins. Based on these considerations, alpha prolamin genes can be divided into three new subgroups, a1, a2, and a3 prolamins (fig. 2). The a1 prolamins include the D subgroup, the oldest alpha prolamins. The a2 prolamins consists of the C subgroup, whereas the a3 prolamins comprises both A and B subgroups (fig. 2).

Synonymous substitutions (Ks) were calculated for all alpha prolamin gene copies by the MEGA4 program (Tamura et al. 2007). Because the sorghum and maize genomes have evolved independently for ∼11.9 my (Swigonova et al. 2004), the substitution rate (r) for the a2 and a3 genes was calculated to be 1.3 × 10⁻⁸ and 1.24 × 10⁻⁸, respectively. Based on the a3 gene substitution rate of 1.24 × 10⁻⁸, the first z1B gene copy arose ∼9 Ma, about the same time that the progenitors of sorghum and sugarcane diverged (Jannoo et al. 2007). The first z1A2 gene amplification occurred ∼3.0 Ma, after the allotetraploidization of maize at ∼4.8 Ma (Swigonova et al. 2004). Because the z2 prolamins were conserved within the Panicoideae subfamily, the substitution rate of the z2 genes was applied to calculating the z2 gene copying events. This analysis predicted that the first z2 gene arose ∼21.3 Ma. Because the foxtail millet and sorghum copy formed the oldest subclades that also are in collinear positions, it suggested that this event occurred before the tribal split. An additional translocation of an z2 gene copy occurred ∼14.4 Ma, forming the C locus in the progenitors of the Andropogoneae. An independent insertion event ∼13.3 Ma at a different chromosomal location was added in foxtail millet, at almost the same time that the progenitors of foxtail millet and switchgrass diverged (Vicentini et al. 2008).

**FIG. 2.** Phylogenetic analysis of alpha prolamin genes in maize, sorghum, and foxtail millet. In total, 75 copies of alpha prolamin genes, except four truncated genes, were aligned with ClustalW. The phylogenetic tree was drawn using the ML method from the MEGA4 program, and bootstrap values were calculated with 1000 replications. All alpha prolamin genes can be divided into three subgroups, z1, z2, and z3. The z1 prolamins includes the D subgroup, the z2 prolamins consists of the C subgroup, and the z3 prolamins comprises both A and B subgroups.
The first translocated alpha gene copy underwent an increase in molecular weight from 19 kDa ($\alpha_1$) to 22 kDa ($\alpha_2$) through tandem duplication of the internal glutamine repeat block of 20 amino acids (Geraghty et al. 1981). Because only paralogous $\alpha_1$ prolamin copies were available for analysis, the substitution rates of $\alpha_2$ and $\alpha_3$ genes were used for calculating the age of the oldest prolamin gene copy. This analysis suggested a date about 30–32 Ma as the age of the first alpha prolamin, about 10–15 my after the Panicoideae subfamily diverged from other subfamilies of the Poaceae. Based on the maize $\alpha_1$ prolamin (z1D), one might suggest that $\alpha_3$ prolamin could have arisen directly from $\alpha_1$ prolamin; 19–21 Ma, but the maize $\alpha_1$ prolamin paralogs are much younger than the original sorghum $\alpha_1$ prolamin copy (k1D). Considering the loss of the original z1D and z1C copies of maize but their conservation in sorghum, more weight has been given to the distances of sorghum alpha prolamins in respect to the chronology and divergence of alpha prolamins. Furthermore, considering that k1C and k1A are physically linked in sorghum, the most parsimonious conclusion would be that the $\alpha_3$ copy arose from the $\alpha_2$ copy by unequal crossing over.

Chromosomal Organization of Alpha Prolamins in the Panicoideae Subfamily

Positional information can also be used to infer insertions and deletions. Gene copies of $\alpha_1$ at locus D included the oldest alpha prolamin gene in the Andropogoneae tribe (fig. 2 and Xu and Messing 2008a). However, when flanking orthologous genes were aligned in foxtail millet, sorghum, maize, rice, and Brachypodium, the alpha prolamin gene was found to be missing not only from rice and Brachypodium (which do not have alpha prolamins) but also from foxtail millet (fig. 5). Furthermore, the position of alpha prolamin gene copies in maize was shifted relative to collinear genes, suggesting that they were paralogous copies of the founding ancestral gene, which was subsequently deleted. Interestingly, a pair of cytochrome P450 genes was found to be flanking the $\alpha_1$ kafirin gene in sorghum, whereas foxtail millet and maize had only one copy (fig. 5). Phylogenetic analysis showed that one of the two cytochrome P450 gene copies from sorghum, Sb08g019430, was closely related to the maize GRMZM2G132450 gene and the other, Sb08g019470, to the foxtail millet SiPROV007183m gene (fig. 6, supplementary fig. 2, Supplementary Material online), suggesting that Sb08g019430 and Sb08g019470 were duplicated before the progenitor of foxtail millet, maize, and sorghum split, but after the Panicoideae formed as a subfamily, and the original alpha prolamin gene was deleted by homologous recombination in foxtail millet and maize (fig. 7).

The $\alpha_2$ genes at locus A were subjected to subsequent tandem duplications in foxtail millet and sorghum that each gave rise to six $\alpha_2$ (s1C) and five $\alpha_2$ (k1C) gene copies at this location. No $\alpha_2$ (z1C) gene copies were found in the collinear position of the maize genome, suggesting that the first $\alpha_2$ gene copy in maize was deleted (fig. 8). However, this model predicts that, before this deletion occurred, an $\alpha_2$ gene copy in both maize and sorghum was inserted into a distal region only a few 100 kb from the donor position. This insertion event could be simply explained by occurrence after the tribal split because foxtail millet was missing the gene copy in this collinear position. Alternatively, the insertion could have been at the base of the Panicoids, but later lost by
deletion in *Setaria*. Additional taxa need to be further investigated in order to differentiate between these two models.

The new \( \alpha_2 \) prolamin gene copy insertion at the distal position in maize and sorghum was followed either by removal of the donor sequence or subsequent tandem amplification of the inserted copy. The last insertion events also produced an expansion in the copy number of genes in the Andropogoneae tribe. In foxtail millet, \( \alpha_2 \) (s1C) gene copies were found at both a closely linked site and one other chromosomal location. In all cases, these transposed setarin genes underwent subsequent tandem duplications (fig. 2; supplementary fig. 3, Supplementary Material online). Nine \( \alpha_3 \) (z1A) and two \( \alpha_3 \) (k1A) genes were in colinear positions (fig. 8), but no \( \alpha_3 \) (s1A) gene was present in the orthologous region of foxtail millet, suggesting that the \( \alpha_3 \) subgroup of prolamin was generated in the Andropogoneae or lost from the *Setaria* lineage. Moreover, the \( \alpha_3 \) subgroup of prolamins was amplified in the genus *Zea*, as described recently (Xu and Messing 2008b). It appears that maize had yet another independent insertion event, of \( \alpha_3 \) (z1A), after allotetraploidization and followed by tandem duplication on chromosome 4S. Interestingly, the collinear prolamin gene regions in the Panicoideae appear to have been chromosomal hot spots for tandem gene amplification because other clusters of patatin-like, disease resistance, and DNA-binding protein genes of various sizes were observed. Gene duplication and gene loss events are summarized in fig. 9. The modes of amplification appear to be mostly unequal recombination or non-homologous events like transposition, although ploidy variation explains some events.

**Discussion**

**Differential Amplification of Alpha Prolamin Genes in the Panicoideae Subfamily**

Here, we used the 8.3× foxtail millet sequence data set assembled into chromosomal pseudomolecules to identify 15 alpha setarin genes at five different chromosomal loci, indicating a higher dispersal rate than in sorghum, but a lower rate of tandem duplication than in sorghum (23 genes in Btx623) or maize (41 genes in B73) (table 1) (Song and Messing 2002, 2003; Xu and Messing 2008b). Given that foxtail millet has approximately two-third the size of the sorghum genome and approximately one-fifth the size of the maize genome, the dispersal rate suggests a dynamic foxtail millet genome in the absence of genome expansion by retrotransposable elements. Phylogenetic analysis showed that the younger \( \alpha_3 \) prolamins are only present in the Andropogoneae tribe. Furthermore, sorghum has only two copies of \( \alpha_3 \) prolamins instead of 20 in maize, demonstrating the rapid expansion of alpha prolamin gene clusters in maize mostly after allotetraploidization. Whether this is a general theme in maize is difficult to determine because of the lack of contiguous sequences across entire BAC clones in the maize genome sequence, where some duplications might be missed with incomplete assemblies, as exemplified by the P1 locus (Goettel and Messing 2009). It seems that the species-specific copy number variation of \( \alpha_3 \) prolamin genes had the biggest impact on the total number of alpha prolamin genes. However, between maize inbreds, copy number variation was found to be the most extensive for \( \alpha_2 \) prolamins (Song and Messing 2003).

**Expansion and Contraction of Alpha Prolamin Genes**

Alpha prolamins represent the youngest prolamins within the grass family and have been found only in the Panicoideae subfamily so far. Based on phylogenetic data, the \( \alpha_1 \) prolamin gene copy could have originated from a copy of a delta
prolamin gene (Xu and Messing 2008b, 2009). In contrast to sorghum, there are two delta prolamin gene copies in two homoeologous regions of maize, formed by allotetraploidization \( \sim 4.8 \) Ma. In addition, the two copies differ by an internal duplication, giving the second copy a coding product of 18 kDa rather than 10 kDa (Swarup et al. 1995). When alpha prolamins of different sizes were compared, it was noted that the internal structure of their amino acid sequences varied between 7 and 8 copies of a 20 amino acid repeat (Messing et al. 1983). This repeat block is characteristic of a glutamine track that is common to all prolamins. Therefore, one can envision the same mechanism that gave rise to tandem gene amplification by unequal crossing over also giving rise to size variation for the internal repeat structure. Alternatively, slipped strand replication could have led to this tandem duplication, although this process is more pronounced with very short sequences (1–6 bp) (Toth et al. 2000; Goldfless et al. 2006) as proposed for Huntington disease (Kovtun and McMurray 2001) but has also been found in *Escherichia coli* to be as large as 624–787 bp (Lovett et al. 1993; Bierne et al. 1997). Consistent with the expansion and contraction via the internal repeat structure, the amino acid composition of different sizes of alpha prolamins of the same subgroups are very similar (supplementary tables 1 and 2, Supplementary Material online).

**Alpha Prolamin Gene Deletion by Homologous Recombination**

The phylogenetic relationship of two cytochrome P450 genes in sorghum, which are flanking the \( a_{1} \) prolamin gene (k1D), and the ones in foxtail millet and maize suggest a model that the tribal progenitor of foxtail millet underwent unequal homologous recombination that was sited near the 5’ end of the cytochrome P450 genes, resulting in the deletion of the prolamin D gene in foxtail millet and maize.

**Fig. 7.** Model for deletion of the alpha prolamin founder gene, D, in the Panicoideae subfamily. The prolamin D gene is the short arrow and the two flanking cytochrome P450 gene copies are the long arrows in gray and black. Homologous recombination presumably occurred between two cytochrome P450 genes to delete the prolamin D gene in foxtail millet and maize.

**Fig. 8.** Sequence alignment of orthologous regions of C and A types of prolamin genes in rice, foxtail millet, sorghum, and maize. The \( a_{2} \) prolamin genes are shown in red, the \( a_{3} \) genes are in purple, and all other syntenic genes are in black.
in the loss of the alpha prolamin gene and one cytochrome P450 gene (fig. 7). The model further predicts an independent unequal recombination site near the 3′ ends of the two cytochrome P450 genes after sorghum and maize progenitors diverged, leading to the loss of the maize orthologous α1 prolamin gene (α1D) (fig. 7). Therefore, only sorghum is predicted to retain the oldest alpha prolamin gene copy (k1D), which is closely related to the HMW-like prolamin genes present only in the subfamily of the Pooideae (Xu and Messing 2009). In sorghum, this subgroup was not further duplicated, whereas additional copies were produced and retained in maize close to the donor copy. In contrast to maize, copies were inserted into dispersed sites of the foxtail millet genome (supplementary fig. 4, Supplementary Material online). After the two modes of gene duplication in maize and foxtail millet, the donor gene was deleted as described above.

**Accelerated Evolution of Alpha Prolamin Genes**

Based on the predicted substitution rates for alpha prolamin genes, we could reconstruct the dispersal and amplification of gene copies in foxtail millet, sorghum, and maize (fig. 10A). The α1 prolamin gene in sorghum appears to be the oldest copy of this gene family, originating around 30–32 Ma, well after the Panicoideae subfamily split from other subfamilies of the grasses. This divergence time is

---

**Fig. 9.** Modes of gene duplication and gene loss. At least four processes can duplicate a gene: (a) WGD (polyploidy); (b) segmental duplication; (c) tandem duplication; (d) ectopic duplication (translocation) or transposition duplication. Genes can be deleted by homologous recombination, often between flanking genes or transposable element. (e) Genes also can be deleted frequently after WGD (polyploidy), (f) by unequal recombination or deletion associated with chromosome break repair (Puchta 2005) or (g) ectopic duplication events.

**Fig. 10.** (A) Gene pedigree of alpha prolamins in foxtail millet, sorghum, and maize. Alpha prolamins have diverged into three subgroups in maize and sorghum (α1, α2, and α3) and two in foxtail millet (α1 and α2). α1, α2, and α3 prolamins are shown in ovals with black, white, and red colors respectively; and A, B, C, and D refer to loci used in maize and sorghum. Prolamin gene copy number and speciation are illustrated in progression from left to right. Vertical lines indicate chromosomal segments containing alpha prolamin genes. The approximate time of gene duplication events is indicated in Ma based on the substitution rates of α2 and α3 prolamin genes. The progenitors of sorghum and maize last shared a common ancestor about 11.9 Ma, and allotetraploidization was at least 4.8 Ma in maize, are shown by the dotted vertical lines. Prolamin gene deletion and duplication models for (B) α1 prolamins and (C) α2 and α3 prolamins.
consistent with the earliest origins of C4 species ~32 Ma in the Oligocene (Christin et al. 2008; Vicentini et al. 2008). Interestingly, the orthologous gene copy was apparently lost after new copies were duplicated in foxtail millet and maize (fig. 7 and 108). The original α2 prolamin gene copy in maize (Z1C) was also deleted, as shown by comparison with sorghum and foxtail millet (fig. 10C), reinforcing the observation that older gene copies could eventually and preferentially be lost when they are members of a large gene family.

The synonymous substitution rate of adh genes (6.5 × 10⁻⁹) has been widely applied to date the divergence times of many genes, although substitution rates can differ by more than an order of magnitude for different genes even within the same species (Gaut et al. 1996; Lai et al. 2004; Swigonova et al. 2004; Xu and Messing 2008a). The average substitution rate of beta, delta, and gamma prolamins would be 6.2 × 10⁻⁹ (Xu and Messing 2009). However, the nucleotide substitution rate for alpha prolamin is 1.3 × 10⁻⁸, twice as high as other prolamin genes, coinciding with the same rate used for the calculation of LTR retrotransposition events (Ma and Bennetzen 2004). The reason for this might be that other prolamin genes have single or low copy numbers. On the other hand, the higher rate of nucleotide substitutions of a gene would also favor neo/subfunctionalization of gene functions, as has been suggested for disease resistance and DNA-binding protein genes (Leister 2004; Gramzow et al. 2010). However, there is no apparent evidence for neofunctionalization or subfunctionalization of the alpha prolamin genes. Many copies are not expressed and they are all expressed in the same tissue (endosperm), although expression levels of some copies are significantly different (Woo et al. 2001; Song and Messing 2003; Zhou et al. 2010; Miclaus et al. 2011). Thus, many gene copies are pseudogenes, which are caused by having premature stop codons, insertions, frame-shift mutations, and truncations. Copies with intact coding regions could also be epigenetically silenced. So, why would damaged copies not further deteriorate? In this respect, disease-resistance genes could serve as an example. At the Rp1 gene cluster in maize, unequal recombination could create recombinant genes to activate non-transcribed genes, which results in non-parental resistance specificity (Smith and Hulbert 2005). One could envision that recombination could rearrange or reactivate copies, suggesting that inactive gene copies resemble a savings account for drastic environmental changes. Therefore, subfunctionalization probably occurred mostly for different prolamins rather than within individual prolamin groups, as shown by RNA interference (Wu and Messing 2010).

Gene Duplication and Gene Loss After Polyploidy

As already described above, the delta zein gene copies arose from a WGD event (polyploidy), which represents another mode of gene duplication. Polyploidy has been quite common in angiosperm lineages, resulting in homoeologous gene pairs. Because meiosis provides a strong selection against intact homoeologous chromosomes, diploidization can result in drastic losses of gene pairs that were duplicated by polyploidy (Seoighe and Wolfe 1998; Vision et al. 2000; Yu et al. 2005). In maize, more than 50% of the homoeologous genes generated by allotetraploidization ~4.8 Ma have been deleted (Lai et al. 2004). Indeed, the half-life of duplicated genes was estimated to be only 3–7 my (Lynch and Conery 2000). Only delta and gamma prolamins retained homoeologous gene pairs in maize (Xu and Messing 2008b). Moreover, genes that are retained after duplication correspond mostly to transcriptional regulators and other genes involved in multi-protein interactions (Murat et al. 2010). In rice, 80% of the retained homoeologous transcription factors were subfunctionalized or neofunctionalized (Throude et al. 2009), and the same is true for O2-like genes that control alpha prolamin transcription in maize (Xu and Messing 2008a).

Supplementary Material

Supplementary tables 1 and 2 and figures 1–4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Selman Waksman Chair in Molecular Genetics to J.M. and the foxtail millet sequence data were produced by the Joint Genome Institute as part of the foxtail millet genome project (Principal Investigators: Bennetzen JL, Devos KM, Doust AN, Kellogg EA, Ware D, and Zale J).

References


