Teff, an Orphan Cereal in the *Chloridoideae*, Provides Insights into the Evolution of Storage Proteins in Grasses

Wei Zhang,[1] Jianhong Xu,[2] Jeffrey L. Bennetzen,[3] and Joachim Messing*,[1]

[1]Waksman Institute of Microbiology, Rutgers University

[2]Zhejiang Key Laboratory of Crop Germplasm, Institute of Crop Science, Zhejiang University, Hangzhou China

[3]Department of Genetics, University of Georgia, Athens

Accepted: May 6, 2016

Data deposition: This project has been deposited at NCBI under the accession KX265112 - KX265188.

Abstract

Seed storage proteins (SSP) in cereals provide essential nutrition for humans and animals. Genes encoding these proteins have undergone rapid evolution in different grass species. To better understand the degree of divergence, we analyzed this gene family in the subfamily *Chloridoideae*, where the genome of teff (*Eragrostis tef*) has been sequenced. We find gene duplications, deletions, and rapid mutations in protein-coding sequences. The main SSPs in teff, like other grasses, are prolamins, here called eragrostins. Teff has γ- and δ-prolamins, but has no β-prolamins. One δ-type prolamin (δ1) in teff has higher methionine (33%) levels than in maize (23–25%). The other δ-type prolamin (δ2) has reduced methionine residues (<10%) and is phylogenetically closer to α prolamins. Prolamin δ2 in teff represents an intermediate between δ and α types that appears to have been lost in maize and other *Panicoideae*, and was replaced by the expansion of α-prolamins. Teff also has considerably larger numbers of α-prolamin genes, which we further divide into five sub-groups, where α2 and α5 represent the most abundant α-prolamins both in number and in expression. In addition, indolines that determine kernel softness are present in teff and the panicoid cereal called foxtail millet (*Setaria italica*) but not in sorghum or maize, indicating that these genes were only recently lost in some members of the *Panicoideae*. Moreover, this study provides not only information on the evolution of SSPs in the grass family but also the importance of α-globulins in protein aggregation and germplasm divergence.

Key words: grass genomes, seed protein genes, gene copy number variation.

Introduction

Seed storage proteins (SSPs), one of the major components in cereal kernels besides starch and oil, have been extensively studied in wheat, rice, and maize (Shewry and Halford 2002). This is largely due to the importance of these crops in agriculture. Wheat belongs to the subfamily *Pooideae*, rice to *Ehrhartoideae*, and pearl millet, sorghum, and maize to *Panicoideae*. Subfamily *Chlorodoideae*, which is more closely related to *Panicoidea*e than to either the *Pooideae* or the *Ehrhartoideae* (fig. 1), consists mainly of weedy and forage grasses (Kellogg 2001) and therefore lacks broad research investigation compared with major crops. Recently, the genome of teff, a Chloridoid grass, has been sequenced by Illumina HiSeq and 454 platforms, providing the first draft genome in this subfamily (Cannarozzi et al. 2014). Teff has been cultivated for human consumption in Ethiopia for centuries. Over the past decade, the recognition that teff grain has few toxic epitopes against celiac disease patients if at all, high levels of essential amino acids like lysine and methionine, and high levels of minerals (especially calcium and iron) has attracted global interest (Baye 2014).

SSPs are classified based on solubility into albumins (soluble in water), globulins (soluble in saline), prolamins (soluble in 60–70% alcohol), and glutelins (soluble in alkali) (Osborne 1908; Shewry and Casey 1999). Prolamins are the major SSPs in most common cereals including teff (Adebowale et al. 2011). Two exceptions are rice and oats, which accumulate more globulins and glutelins (Shewry and Halford 2002). It has been suggested that the ancestral prolamin gene arose from a tandem duplication of an α-globulin gene (Xu and Messing 2009). During the evolution of the grasses, prolamin genes were copied and inserted tandemly

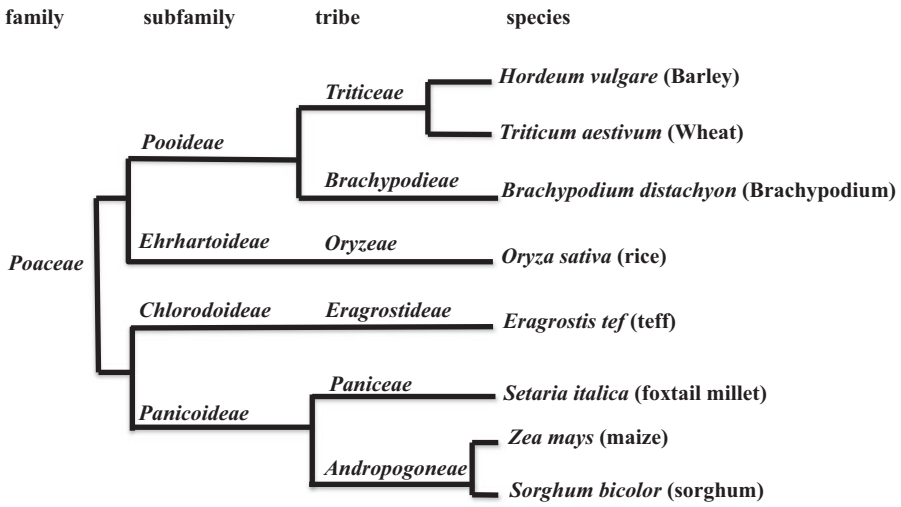

Fig. 1.—Phylogenetic relationships among several common cereals. Names of family, subfamily, tribe, and species are shown. This is a generalized tree of relationships adapted from previous studies (Kellogg 2001; Vincentini et al. 2008; Xu and Messing 2009) with no computational support and phylogenetic distances were not drawn to scale.

or dispersed over different chromosomal regions, which in turn gave rise to further amplifications. Donor copies could either be maintained or lost by unequal crossover events (Xu et al. 2012). Genome-wide dispersal and subsequent divergence gave rise to new groups of prolamins, like the δ- and α-prolamins. The major prolamins in wheat, high-molecular-weight (HMW)-glutenins, belong to the group III-HMW-type of prolamins, low-molecular-weight (LMW)-glutenins and gliadins belong to the group II-γ-type of prolamins. Rice has group II-γ-type and group I-δ-type prolamins and maize group I-α- and δ-type, and group II-γ-type (Xu and Messing 2009). All investigated species in the *Panicoideae* have α-prolamins as the major storage proteins, associated with recent gene amplifications (Song et al. 2001; Song and Messing 2002; Xu and Messing 2008). Detailed analysis of the types of prolamins in teff is needed to fill the gap in understanding the evolution of SSPs in the grasses.

Similar to prolamins in structure, 2S albumins also have A, B, and C domains. In addition, 2S albumins have 10 conserved cysteines to form disulphide bridges (Shewry et al. 1995; Shewry and Pandya 1999). Indolines are basic, cysteine-rich proteins with a unique tryptophan-rich domain that forms indole rings of tryptophan and binds to lipid granules. The hardness (Ha) locus in wheat comprises puroindolines (puros is the Greek name of wheat) genes, *Pina*, *Pinb*, and *Gsp*-1, present in the wheat D genome (Chantret et al. 2005). These genes also have 10 conserved cysteine residues and are believed to have originated from 2S albumins. Homologous genes are also present in other *Triticeae* species, where soft endosperm is a dominant trait (Gautier et al. 1994). Related grain softness genes can be found in brachypodium and rice but not in sorghum, although the flanking genes are conserved in all three species (Charles et al. 2009).

In this study, we looked in detail at the prolamins in two different cultivars of teff: Tsedey, the sequenced genome, and Dabbi (PI 524434, www.ars-grin.gov). We found 42 prolamin and Ha-like genes in Tsedey and PCR amplified a similar number of prolamin genes from Dabbi. These prolamins genes were used to analyze the evolution of the prolamins in the grass family.

Results

Prolamins in Teff

Amino acid composition was determined for seeds from teff cultivar Dabbi. We found that this teff grain contains 8.5% protein. The content of the essential amino acids methionine and lysine is comparable to that in rice, but higher than that in other cereal crops (supplementary table S1, Supplementary Material online).

Following the standard nomenclature for prolamins, we named the prolamins in teff eragrostins. Eragrostins were separated into ~50, ~27, ~22, ~19, ~15, and ~10 kDa components by SDS-PAGE gel electrophoresis (fig. 2), which is very similar to the results previously seen for species of the subfamily *Panicoideae*, like maize and sorghum.

Determination of the type of prolamin and sequences of 50 and 27 kDa eragrostins was achieved through sequence analysis and LC-MS (liquid chromatography-mass spectrometry). Because maize 50 and 27 kDa zeins are γ-prolamins, we suspected that these two protein bands might also encode γ-prolamins in teff. First, screening the sequenced teff genome (http://www.tef-research.org/genome.html) for homologous genes to γ-prolamin genes in the *Panicoideae* subfamily found three sequences that could encode teff γ-type prolamins. Indeed, a previous study has identified a DNA

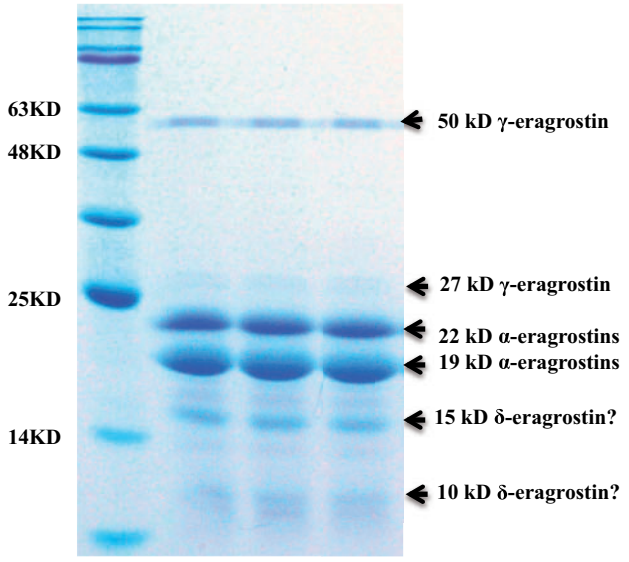

Fig. 2.—Alcohol soluble proteins from mature tefs seed resolved in a 15% SDS-PAGE gel. Lane 1, protein markers. Lane 2–4, extractable prolamin components by 70% ethanol. The types of 15 and 10 kDa eragrostins were predicted with no experimental validations.

sequence (C8175559) as a γ-prolamin gene in its phylogenetic analysis (Cannarozzi et al. 2014). However, we could identify two additional prolamin gene sequences in the same dataset: scaffold5275:48155–46417 and scaffold512:118947–115807. Although predicted prolamin sequences in both scaffolds are incomplete, the protein sequences at the start and end positions could be derived. Aligning these amino acid sequences indicates that there are at least two versions of γ-eragrostins in teff: one ends with MAGAAAI and the other one ends with MAGAGVI (translated from the sequences in scaffold5275 and scaffold512 in the Supplementary File online). We excised the two protein bands of 50 and 27 kDa from SDS-PAGE and analyzed these bands by trypsin–LC-MS. Proteins of 50 kDa size contain a peptide fragment of EFLKQQCSPSAMPFLQSRVSPPTRCQVLRRKCCQQLKQVEPLY RQQAIFEMVQSIIQQQPQQQEEQAAGG, whereas proteins at the 27 kD position contain the peptide fragment QQCSPSAMPFLQSRVSPPTRCQVLRRKCCQQLKQVEPLYR, the same as predicted from the sequences of Et_C8175559 and Et_scaffold5275. Alignment of these sequences with maize zeins, allowed us to classify the three sequences in the teff genome as 50 and 27 kDa γ-eragrostins, respectively.

We could locate a total of 40 copies of eragrostins in the published Tsedey genome including the above three sequences that correspond to the γ-type (see Supplementary File online). The other prolamins were named t1α and t1δ based on similarities to alpha and delta prolamins (fig. 3). These prolamins were clustered into sub-groups by similarity of amino acid sequences with the program MEGA. According to their phylogenetic relationships, prolamins in teff can be divided into five α subgroups and two δ subgroups (fig. 3). The major differences between α and δ prolamins are: (1) the higher level of methionine and cysteine in δ compared with 0–2% of these essential amino acids in α prolamins and (2) the higher level of glutamine in α-prolamins (24% to over 40%) than in δ prolamins (around 10%) (table 1). The δ prolamins are divided into two subgroups: δ1 has much higher levels of methionine and cysteine and a lower level of glutamine compared to δ2. This composition places δ2 between α and δ1 prolamins. We found 12 copies of δ1 prolamins and 11 copies of δ2 prolamins in teff, contrary to only two copies of δ-prolamin genes in maize.

The α and δ eragrostins correspond to 22 and 19 kDa protein bands. In a previous report, two major prolamin peaks resolved by SDS-PAGE were identified by HPLC (Tatham et al. 1996). Prolamin peak tef 6 was recently assembled by 454 sequencing (Cannarozzi et al. 2014). However, no full sequences with tef2 were identified. We found that Et_Scaffold1101.126800–127390 has a full-length prolamin gene with a tef2 profile. In addition, we identified seven prolamin genes in this α2 subgroup and five prolamin genes in the α5 subgroup that corresponds to a tef 6 profile. According to figure 3 in Tatham et al. (1996), α5 prolamins should represent the major protein of 22 kDa eragrostins and α2 prolamins the major 19 kDa eragrostins. Possibly, α1, α2, and α4 prolamins were separated into different peaks by HPLC such that their sequence features were not identified in that study (Tatham et al. 1996). However, based on the predicted sizes of these prolamins (table 1), it is possible that α1, α2, and α4 prolamins also contribute to the 19 kDa eragrostin band in SDS-PAGE.

Puroindoline Genes in Teff

The hardness locus (*Ha*) in wheat has three functional genes *Pina*, *Pinb*, and *Gsp*-1, plus a PseudoPinb and a Pinb-relic (Chantret et al. 2005). Genes encoding BGGP, a β-1-3-galactosyl-*O*-glycosyl-glycoprotein and HIPL, a Hedgehog-interacting-like protein, flank the *Ha* locus and are conserved in wheat, brachypodium, and rice (Charles et al. 2009). The Tsedey genome has two *Ha*-like genes in Scaffold4919:528-94 and Scaffold1023:17200–17628, respectively. In addition, the two genes flanking the *Ha* locus, HIPL and BGGP, are also in the same scaffold as the *Ha*-like gene in Scaffold1023 (fig. 4*A*). However, in Scaffold4919, only HIPL is present. It is not known whether BGGP is absent from scaffold4919 due to its limited length. It is possible that Scaffold4919 and Scaffold1023 represent two ancestral *Ha* loci because teff is a tetraploid. Moreover, foxtail millet (*Setaria italica*) that belongs to the *Paniceae* tribe in the subfamily *Panicoideae*, also has a *Ha*-like gene. In XM_004963284, the *Ha*-like gene is located between HIPL and BGGP (fig. 4*A*). Similarly, the sorghum genome has HIPL (sb08g023170) and BGGP (Sb08g023160) genes next to each other but no *Ha*-like

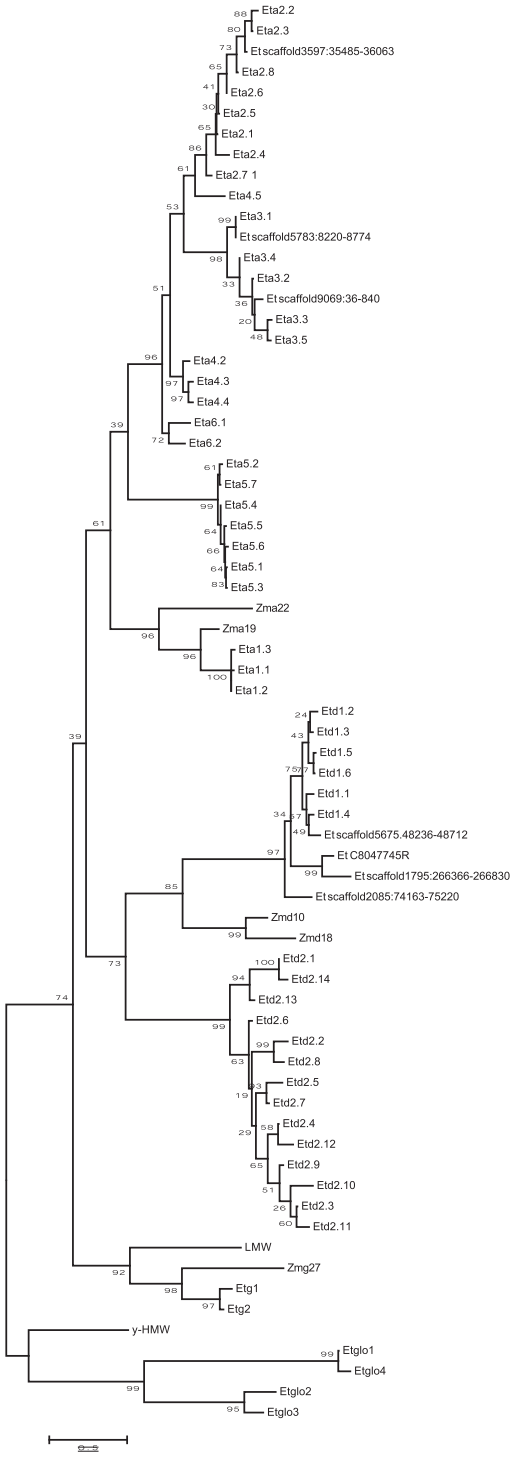

Fig. 3.—Phylogenetic analysis of teff eragrostins. Genomic sequences of all eragrostins from Dabbi were obtained with PCR and traditional sequencing (Methods). A few eragrostins from Tsedey that are absent from Dabbi and a few representative maize genes were also used for phylogenetic analysis. The sequences of the eragrostins used can be found in supplementary table S2, Supplementary Material online. A phylogenetic tree was drawn using the MEGA5 program with the Neighbor-Joining method.

genes between HIPL and BGGP (fig. 4*A*). However, the available maize genome of inbred B73 does not have *Ha*-like genes, but has many copies of BGGP in different chromosomal locations based on our BLAST results. Therefore, it is likely that BGGP was copied and reinserted, whereas the original *Ha* locus was deleted in maize..

The difference between the Pina and Pinb-group is the number of tryptophans in the tryptophan-rich domain. Aligning the tryptophan-rich domains of *Ha*-like genes suggests that *Ha*-like genes in teff are closely related to the Pinb-group with only three tryptophans (fig. 4*B* and *C*). This group of *Ha*-like genes has previously only been seen in species in the *Pooideae*: *Triticum aestivum* (ta), *Hordeum vulgare* (hv), *Secale cereale* (sc), and *Brachypodium sylvaticum* (bs), with the exception of only one from *Panicoideae*: *S. italica* (si) and now one from the *Chloridoideae*, *Eragrostis tef* (et). Hence, these genes must have been present in an ancestral grass species, and then were lost specifically in the Panicoid lineages.

α-Globulins in Teff

There are four α-globulins in the Tsedey draft genome assembly: contig5581:3032–3682, contig5581:5157–5933, contig5582:9918–9268, and contig5582:8208–7336. This result indicates that teff had amplification of α-globulins in both of its progenitor genomes, in contrast to low copies of α-globulin genes in other grasses (Belanger and Kriz 1989; Wallace and Kriz 1991; Shorrosh et al. 1992; Nakase et al. 1996; Woo et al. 2001; Loit et al. 2009).

Comparison of Teff Sequences among Cultivars

To validate the *in silico*-predicted SSPs in teff, primers were designed to amplify these genes in cultivar Dabbi based on gene sequences from Tsedey. We found all groups of prolamin genes in Dabbi, including two genes, Etg1 and Etg2, for the γ-eragrostins, 30 genes for α-eragrostins, and 36 genes for δ-eragrostins (supplementary file 1 and table S2, Supplementary Material online). Specifically, 18 of the α-and δ-eragrostins are 100% identical to those found in Tsedey (supplementary table S2, Supplementary Material online). Four *Ha*-like genes were also identified in Dabbi.

Organization of Teff Storage Proteins in the Endosperm

The nature of the organization of SSPs in the mature endosperm affects kernel hardness and the ultimate usage of the seed flour. For example, wheat storage proteins are packed in protein bodies (PBs) and merged into large storage vacuoles upon seed maturation, whereas maize storage proteins are packed in individual PBs that do not fuse (Arcalis et al. 2014). It has long been thought that wheat HMW glutenins and certain LMW glutenins mainly contribute to merging and collapsing the PB complexes (Rubin et al. 1992). Teff seeds are rich in α- and δ- prolamins, similar to maize seeds that have α-prolamins as the major prolamins. However, teff has large

Table 1
Protein Composition of Teff Storage Proteins

Gene predicted MW	Gamma-eragrostins		Indoline	Delta-eragrostins-1		Delta-eragrostin-2	Alpha-eragrostins				
	γ1 42 kDa (%)	γ2 42 kDa (%)	pin1 14 kDa (%)	α1.3 14 kDa (%)	α1.4 14 kDa (%)	α2.1 19 kDa (%)	α3.1 20 kDa (%)	α2.3 20 kDa (%)	α4.2 20 kDa (%)	α5.1 26 kDa (%)	α1.1 22 kDa (%)
Ala (A)	5.5	4.1	6.5	3.9	7.0	13.2	7.4	4.7	5.8	12.2	5.4
Arg (R)	1.3	1.8	6.5	0.0	0.4	0.6	2.5	0.6	1.2	0.9	1.1
Asn (N)	0.5	0.8	2.4	0.8	0.4	6.0	1.8	2.9	2.9	3.5	3.8
Asp (D)	0.5	1.0	2.4	0.8	0.9	0.0	1.8	0.0	0.0	0.0	0.5
Cys (C)	3.7	3.6	9.7	8.7	5.3	3.6	1.2	1.2	0.6	2.2	0.5
Gln (Q)	25.8	24.3	8.1	11.8	8.8	14.4	39.9	41.5	40.7	24.3	39.1
Glu (E)	2.4	2.3	8.1	0.0	0.0	0.0	1.8	0.0	1.2	0.0	0.5
Gly (G)	13.2	12.9	3.2	3.9	16.7	1.2	1.8	1.8	0.6	1.3	1.1
His (H)	5.8	6.5	2.4	0.8	0.4	1.2	1.8	1.2	1.2	1.3	1.1
Ile (I)	1.1	0.8	4.8	1.6	1.8	5.4	4.3	4.1	5.8	6.1	5.4
Leu (L)	2.6	3.1	4.8	2.4	2.2	4.2	7.4	8.8	8.7	11.7	10.9
Lys (K)	1.1	0.8	5.6	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Met (M)	3.9	4.1	7.3	29.9	18.9	9.0	0.0	2.3	0.6	1.7	0.5
Phe (F)	3.4	3.1	0.0	2.4	5.3	7.2	7.4	4.7	6.4	5.2	4.9
Pro (P)	13.2	15.0	5.6	10.2	9.2	9.6	4.3	3.5	5.8	7.8	4.9
Ser (S)	7.4	7.2	5.6	12.6	13.2	10.2	2.5	6.4	4.1	5.7	4.3
Thr (T)	2.4	2.6	8.1	6.3	4.8	5.4	4.3	4.1	2.9	4.8	3.3
Trp (W)	1.3	1.3	2.4	0.8	0.0	1.8	1.8	0.6	0.6	0.9	1.1
Tyr (Y)	3.7	3.1	4.0	2.4	4.4	3.6	4.3	4.1	2.9	4.8	4.9
Val (V)	1.3	1.6	2.4	0.0	0.4	3.6	3.7	7.6	8.1	5.7	6.5

fused PBs in its endosperm (fig. 5). Under transmission electron microscope, one endosperm cell only has three to four big protein aggregates. The biggest PB in teff is ~15 μm in diameter, whereas PBs in maize are only ~1 μm in diameter (fig. 5). Another characteristic of teff PBs are the smaller electron-dense PBs. These PBs usually locate on the membrane surface of the big electron-light PBs or aggregate into bigger, electron–dense protein complexes.

The SSPs responsible for the aggregation of PBs in teff should localize on the surface of PBs. In the first endosperm layer, PBs vary in size, and are smaller than those in outer endosperm layers (fig. 6), whereas electron dense PBs are hardly visible in this layer. In the fourth endosperm layer, both the electron–light and electron-dense PBs are larger, but protein aggregation is not obvious. In the fifth endosperm layer, electron-dense PBs emerge; protein budding and protein aggregation become obvious.

Discussion

Tandem Duplication Is a Common Feature of Seed Storage Proteins in Grasses

Different species of the grass family accumulate different types of SSPs associated with the expansion of different SSP gene numbers in these species. For example, wheat accumulates HMW- and γ-prolamins as its major SSPs (Payne et al. 1984), whereas rice mainly accumulates globulins and glutelins in its seeds, although in contrast to wheat, it also accumulates prolamins (Krishnan and White 1995). Species of *Panicoideae* on the other hand accumulate the young α-prolamins as their major SSPs (Thompson and Larkins 1994; Xu et al. 2012). Table 2 summarizes the types of prolamins expressed in representative species in different grass subfamilies.

Even within the same subfamily, species differentially amplify prolamin genes. An example is the α-prolamins in foxtail millet, sorghum, and maize (Xu and Messing 2008; Xu et al. 2012). The oldest α-prolamin genes, α1, are found in foxtail millet, sorghum and maize but the youngest (α3) exhibit no copies in foxtail millet, three copies in sorghum and 20 copies in maize (Xu et al. 2012).

With this in mind, α-prolamins are believed to have originated from δ-prolamins. (Xu and Messing 2009). Maize has only two copies of δ-prolamin genes, rice has four copies and wheat has no δ-prolamin genes. Therefore, the massive expansion of δ-prolamins is a unique feature of the teff genome, although it will be interesting to see if this phenomenon is shared with other Chloridoid grasses. In this study, 23 copies of δ-prolamins were found in the Tsedey draft genome, whereas 36 were found in Dabbi. Because the Tsedey sequence was not complete, and because we did not have access to Tsedey seed or DNA, we do not know if the variation in gene number of δ-prolamins observed between Dabbi and

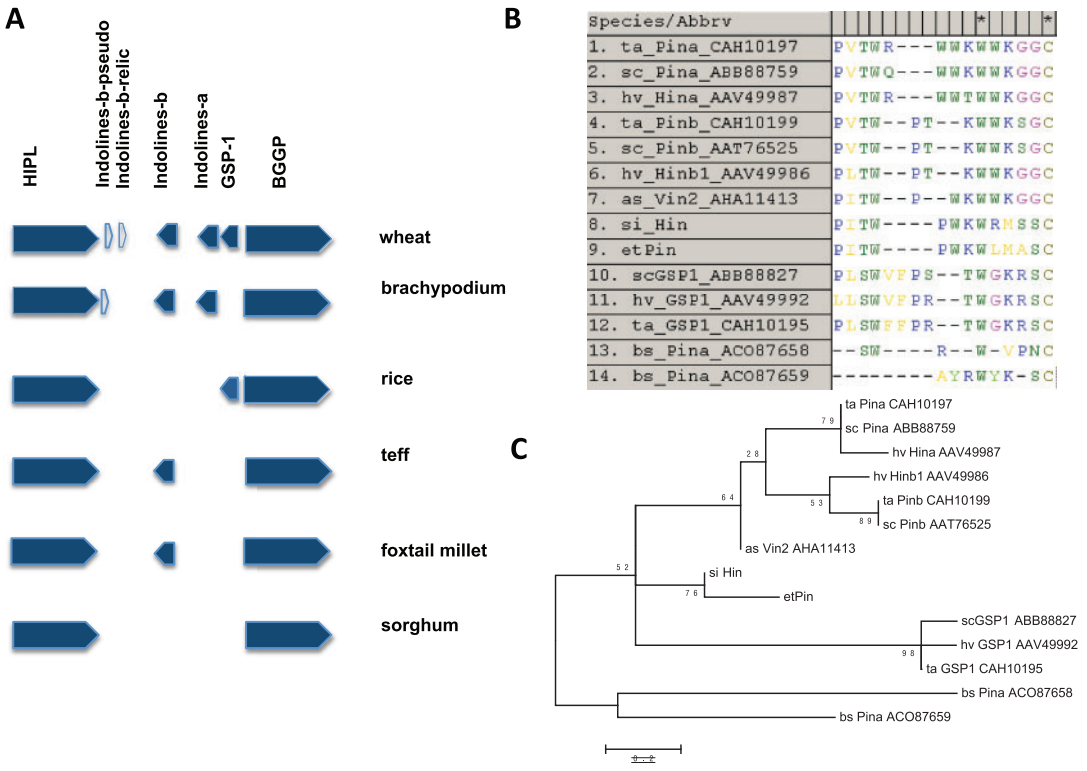

Fig. 4.—Comparison of the *Ha* locus and *Ha*-like genes in wheat (*T. aestivum* [ta]), brachypodium (*B. sylvaticum* [bs]), rice (*Oryza sativa* [os]), teff (*E. tef* [et]), foxtail millet (*S. italica* [si]) and sorghum (*Sorghum bicolor* [bc]). (*A*) Comparison of loci orthologous to *Ha*. *Ha* and *Ha*-like loci of wheat, *B. sylvaticum*, rice and sorghum were from Charles et al. (2009). The teff *Ha*-like locus was from Scaffold1023:14107-23158 (http://www.tef-research.org/genome.html) and the foxtail millet *Ha*-like locus was downloaded from NCBI as XM_004963284. Gene sizes are not drawn to scale. Solid blue represents expressed genes while unfilled genes indicate pseudo-genes or partial genes. (*B*) Comparison of tryptophan-rich domains of *Ha*-like genes. (*C*) Phylogenetic analysis of the relatedness of *Ha*-like genes. The phylogenetic tree shown was drawn with MEGA5 using the Maximum Likelihood method.

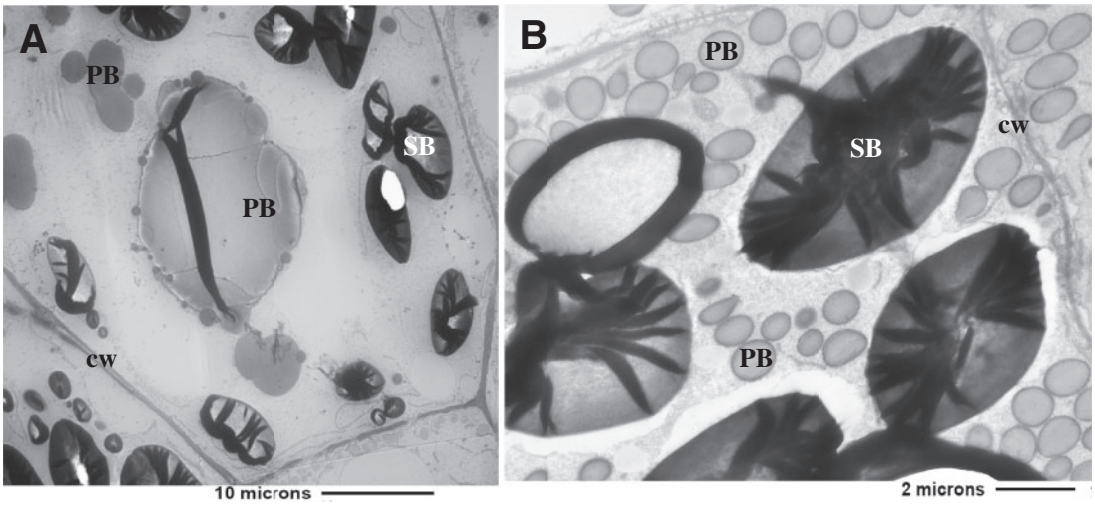

Fig. 5.—Comparison of organization of storage proteins in (*A*) teff and (*B*) maize. CW, cell wall; PB, protein body; SB,: starch body.

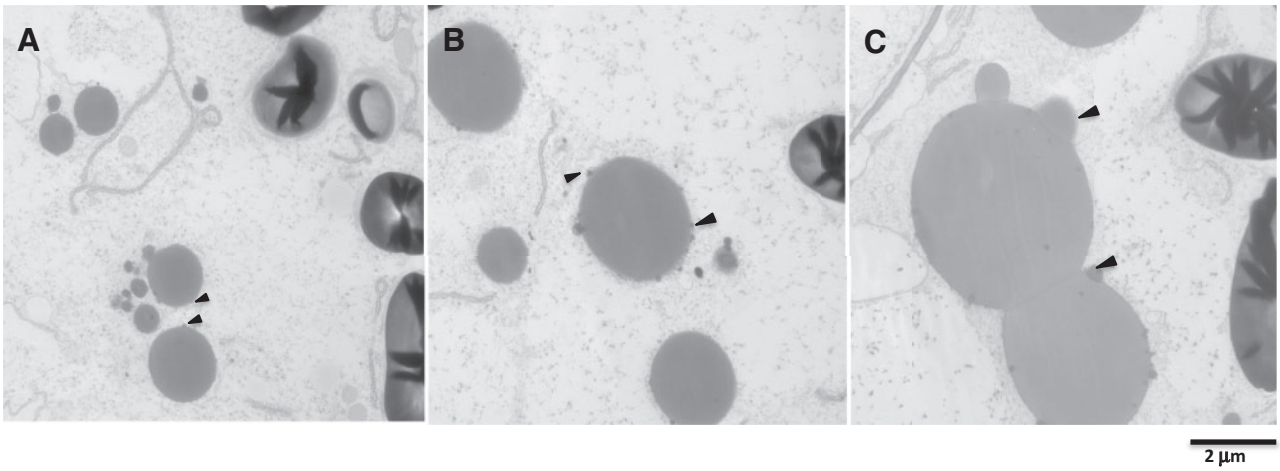

Fig. 6.—Development of PBs in teff endosperm. (*A*) First layer of endosperm. (*B*) Fourth layer of endosperm. (*C*) Fifth layer of endosperm. Arrowheads point to electron-dense PBs.

Table 2

Summary of Kernel Softness, SSP Content, and Storage Structures of SSPs in Several Grass Species

	Kernel Type	Ha-like genes	Globulin gene no.	Storage organelle	Types of prolamin proteins	Percentage in prolamins (%)	Names of prolamin proteins
Wheat	hard or soft	*Pina, Pinb, GSP-1*	3	PBs and PSVs	HMW	6–12[a]	HMW-glutenins
					γ	25–38[a]	LMW-glutenins
						38–50[a]	α-gliadins
							γ-gliadins
							ω-gliadins
Rice	hard	*GSP-1*	1	PBs and PSVs	γ	40[b]	13 kDa Ory13
						60[b]	13 kDa Ory16
							16 kDa Ory13
					δ		10 kDa Ory10
Teff	hard	*Pinb*-like	4	PBs and PSVs	γ	15[c]	50 kDa γ-eragrostin
							27 kDa γ-eragrostin
					δ	25[c]	15 kDa δ-eragrostins
							10 kDa δ-eragrostins
					α	60[c]	22 kDa α-eragrostins
							19 kDa α-eragrostins
Maize	hard	*GSP-1* (PCR)	2	PBs	γ	20–25[d]	50 kDa γ-zein
							27 kDa γ-zein
							16 kDa γ-zein
							15 kDa β-zein
					δ	<5[d]	18 kDa δ-zein
							10 kDa δ-zein
					α	60–70[d]	22 kDa α-zein
							19 kDa α-zein

[a]The percentage of different types of prolamins in total wheat prolamins is based on a previous report (Payne et al. 1984), with ~80% of wheat proteins being glutenins and gliadins.

[b]The percentages of different prolamin species in total prolamins were from a previous study (Ogawa et al. 1987).

[c]The percentage of different types of teff prolamins in total prolamins was calculated by protein band densitometry from Image J, following a previous protocol (Garcia et al. 2015).

[d]Percentages of different zeins in total maize prolamins were from previous reports (Thompson and Larkins 1994; Wu et al. 2009).

Tsedey is a technical artifact or an indication of a real difference between these two cultivars. It is known, however, that accessions within the same species often differ in the copy number of specific genes, especially when those genes are in large gene families. For instance, the number of SSP genes for z1C1 differs between BSSS53 and B73 in Maize (14 copies in B73 and 23 copies in BSSS53) (Miclaus et al. 2011). The differences of the numbers of eragrostins found

in the two cultivars could also be due to different methods used to find these sequences: prolamins in Tsedey were found through blasting to the incomplete genome whereas those in Dabbi were found by PCR purification and sequencing. In Tsedey, three different copies derived from Scaffold 2,167 have the same exact sequence (supplementary table S2, Supplementary Material online), but if it is the case in Dabby, only one sequence will be derived. Although PCR amplification could introduce errors that lead to false positive sequences of the genes predicted. However, as noted in supplementary table S2, Supplementary Material online, most of the genes were obtained from sequences of two or more clones, found both in Tsedey and Dabbi, but the sequences obtained from one clone could still provide useful information of prolamins in teff.

Among the Tsedey δ-prolamins, many of them are duplicated in the same scaffold. For example, Scaffold2167 has four copies of δ1 genes, Scaffold2085 has two copies of δ1 genes, whereas Scaffold5655 and Scaffold7847 each have two copies of δ2 genes. Although teff seems to have many δ-prolamin gene copies, δ-prolamins are not the major SSPs in teff (fig. 2, table 2). Rather, α-prolamins constitute the major prolamins in teff.

Evolution of Grain Hardness in the Grasses

Two genes in the *Ha* locus contribute to grain softness, namely *Pina* and *Pinb* (Giroux and Morris 1998; Giroux et al. 2003), but the other gene (*Gsp*-1) does not seem to affect grain hardness (Tranquilli et al. 2002; Elmorjani et al. 2013). *Gsp*-1 has less tryptophan compared with *Pina* or *Pinb* (fig. 4*B*). A recent study amplifying *Gsp*-1 from different species suggested that crops in the subfamily *Panicoideae* also have *Gsp*-1 (Wilkinson et al. 2013). Aligning the *Ha*-like genes from brachypodium and teff found significant divergence at the start and end of the *Ha*-like genes compared with those in the *Triticeae*. Teff has two *Ha*-like genes represented in two different contigs of Tsedey. These genes are similar in sequence to the *Pinb* group of *Ha*-like genes but lack one of the tryptophan residues present in *Pinb* (fig. 4*B*). Teff *Ha*-like genes are closely related to the *Ha*-like gene in foxtail millet but have diverged greatly from *Pinb* genes in *Triticeae* (fig. 4). Therefore, our results indicate that *Pinb*-like genes are highly variable among grasses and have mostly been lost in panicoid cereals. Considering that the kernel of normal foxtail millet is hard in texture, it appears that the expression of the *Ha*-like genes does not create kernel softness in foxtail millet, perhaps due to multiple amino acid substitutions in the tryptophan domain. *Ha*-like genes in teff likewise might not determine kernel softness.

α-Globulins and PB Aggregation in Teff

Considering the similarity in protein sizes of teff with maize prolamins as judged by SDS-PAGE, it is somewhat unexpected to find larger aggregated PBs in teff, contrary to the singular, individual PBs in species of the *Panicoideae* (fig. 5). Under transmission electron microscopy, teff has big electron-light PBs and smaller electron-dense PBs. Electron-dense PBs are usually on the edges of electron-light PBs and appear as budding structures (figs. 5 and 6). Such a budding phenotype was previously reported in maize containing transgenic HMW-glutenin (Zhang et al. 2013). However, teff does not have HMW-glutenin-like genes. Co-existence of electron-dense PBs and electron-light PBs have mostly been studied in rice, where electron-dense PBs are mainly derived from globulins and glutelins, whereas electron-light PBs are mainly composed of prolamins (Krishnan and White 1995; Nagamine et al. 2011). Wheat prolamins are the major SSPs and are deposited into both electron–dense and electron–light protein structures (Rubin et al. 1992; Tosi et al. 2009). Study of prolamin evolution has pointed to the fact that an alpha-globulin gene is the ancestor of prolamin genes and that HMW-glutenin genes are the oldest type of prolamins derived from alpha-globulin genes (Xu and Messing 2009). Consistent with the deposition of globulins into electron-dense PBs in rice (Krishnan and White 1995), HMW-glutenins are also deposited into electron-dense PBs in wheat (Rubin et al. 1992), whereas the relatively younger prolamins, including γ- and δ-prolamins, in various species seem to favor electron-light PBs (Rubin et al. 1992; Krishnan and White 1995; Tosi et al. 2009; Nagamine et al. 2011). Prolamins in maize consist mainly of α-prolamins, with lower amounts of γ- and δ-prolamins (Thompson and Larkins 1994). Maize starchy endosperm only contains electron-light PBs, whereas protein storage vacuoles were only observed in aleurone cells (Reyes et al. 2011). Immunogold labeling of alpha-globulin shows that this protein is found only in the rough endosperm surrounded PBs characterized by empty space between protein accretion and surrounding membrane (Woo et al. 2001). Additionally, alpha-globulin labeled PBs are darker under electron microscopy, irregular in shape in the sub-aleurone, first layer and second layer of the endosperm. In the mature endosperm, the "empty space" in alpha-globulin-labeled protein structures is filled with growing protein accretion and the PBs are sometimes bigger in size than regular PBs (Woo et al. 2001). The above-mentioned PBs, labeled with alpha-globulin, are like the electron-dense PBs in rice, but only constitute a small portion of all PBs in maize because alpha-globulin genes are expressed at very low levels in maize endosperm (Woo et al. 2001). Teff has four copies of alpha-globulin genes in a tandem repeats on two contigs. Higher copy number could result in higher expression, like the expansion of α-prolamin genes in maize. Indeed, teff contains 11% of globulin + albumin, much higher than sorghum does (6%) (Adebowale et al. 2011). We propose that α-globulins in teff play a major role in PB aggregation and ultimately in its dough property. In this regard, it is possible that α-globulins can be used to genetically engineer crop plants for better dough properties.

Materials and Methods

Protein Analysis

Protein content and amino acids analysis of 1 g of teff seeds were provided by the New Jersey Field Lab, Trenton, NJ. Protein extraction of the teff prolamin component was by sequential extraction with borate buffer and 70% ethanol (Zhang et al. 2013). The resulting prolamin components were dissolved in 1% SDS and resolved in a 15% SDS-PAGE gel. Amino acid compositions of other crops were from previous studies (Houston et al. 1969; Lester and Bekele 1981; Morey and Evans 1983; Ejeta et al. 1987; Krishnan et al. 2005; Wu and Messing 2012; Wu et al. 2012; Zhang et al. 2013).

Prolamin Genes, Ha-like Genes, and Globulin Genes Searches in Teff

Teff genomic sequences were downloaded from http://www.tef-research.org/genome.html. Protein sequences of prolamins and globulins in foxtail millet and maize were used to identify teff prolamins using tblastN. All candidate teff prolamin gene sequences were extracted from the teff genome and manually annotated. Primers were then designed from the identified teff prolamin genes as shown in supplementary table S3, Supplementary Material online. *Ha*-like genes were first found in teff and grouped into prolamins by the above tblastN method, and later found to be more close to Hordoindoline-like genes in *S. italica* (XP_004963341). The corresponding REFSEQ for the indoline encoding locus in *S. italica* was located as XM-004963284. Amplified PCR products from Dabbi were purified and ligated into the T-easy vector for sequencing. Sequences with at least 98% similarity were collapsed.

Phylogenetic Analysis

The following *Ha*-like genes were used in our study: *S. italica* (foxtail millet, NCBI accession number XM_004963284), *T. aestivum* (NCBI accession number CAH10197, CAH10199, CAH10195), *H. vulgare* (NCBI accession number AAV49987, AAV49986, AAV49992), *S. cereale* (NCBI accession number ABB88759, AAT76525, ABB88827), *B. sylvaticum* (NCBI accession number ACO87658, ACO87659). Nucleotide and predicted protein sequences were aligned using clustalW at default settings. The phylogenetic analyses were conducted using the Maximum Likelihood method or Neighbor-Joining method with 1,000 bootstraps in the MEGA5 program (Tamura et al. 2011).

Transmission Electron Microscopy

Maize immature kernels were sliced and fixed as described in a previously published method (Wu and Messing 2010) with some modifications. In brief, 18 day-after-pollination kernels were sliced to 1–2 mm and fixed in 2.5% glutaraldehyde in 0.1 M sodium cacodylate buffer, pH 7.4, overnight. In addition, immature teff kernels were harvested at 14–20 days after flowering and the entire kernels were fixed in the same fixation buffer overnight. Kernels or slices of kernels in fixation were then rinsed with 0.1 M sodium cacodylate buffer, postfixed in 1% osmium tetroxide at 4 °C overnight, dehydrated in an increasing concentration series of acetone, and embedded in epon resin. The samples were cut into 90 nm sections with a Leica EM UC6 ultramicrotome and gridded. For teff kernels, the thin sections included aleurones, subaleurones, and endosperm.

Supplementary Material

Supplementary file, tables S1–S3, and figures S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgment

This work was supported by the Waksman Chair in Molecular Genetics to J.M.

Literature Cited

Adebowale AA, Emmambux MN, Beukes M, Taylor JRN. 2011. Fractionation and characterization of teff proteins. J Cereal Sci. 54:380–386.

Arcalis E, Ibl V, Peters J, Melnik S, Stoger E. 2014. The dynamic behavior of storage organelles in developing cereal seeds and its impact on the production of recombinant proteins. Front Plant Sci. 5:439.

Baye K. 2014. Teff: nutrient composition and health benefits. ESSP Working paper 67. Washington, DC and Addis Ababa, Ethiopia: International Food Policy Research Institute (IFPRI) and Ethiopian Development Research Institute (EDRI).

Belanger FC, Kriz AL. 1989. Molecular characterization of the major maize embryo globulin encoded by the glb1 gene. Plant Physiol. 91:636–643.

Cannarozzi G, et al. 2014. Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). BMC Genomics 15:581.

Chantret N, et al. 2005. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). Plant Cell 17:1033–1045.

Charles M, et al. 2009. Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of Pooideae and Ehrhartoideae, after their divergence from Panicoideae. Mol Biol Evol. 26:1651–1661.

Ejeta G, Hassen MM, Mertz ET. 1987. In vitro digestibility and amino acid composition of pearl millet (*Pennisetum typhoides*) and other cereals. Proc Natl Acad Sci U S A. 84:6016–6019.

Elmorjani K, Geneix N, Dalgalarrondo M, Branlard G, Marion D. 2013. Wheat grain softness protein (Gsp1) is a puroindoline-like protein that displays a specific post-translational maturation and does not interact with lipids. J Cereal Sci. 58:117–122.

Garcia N, Zhang W, Wu Y, Messing J. 2015. Evolution of gene expression after gene amplification. Genome Biol Evol. 7:1303–1312.

Gautier MF, Aleman ME, Guirao A, Marion D, Joudrier P. 1994. *Triticum aestivum* puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant Mol Biol. 25:43–57.

Giroux MJ, Morris CF. 1998. Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. Proc Natl Acad Sci U S A. 95:6262–6266.
Giroux MJ, Sripo T, Gerhardt S, Sherwood J. 2003. Puroindolines: their role in grain hardness and plant defence. Biotechnol Genet Eng Rev. 20:277–290.
Houston DF, Allis ME, Kohler GO. 1969. Amino acid composition of rice and rice by-products. Cereal Chem. 46:527–536.
Kellogg EA. 2001. Evolutionary history of the grasses. Plant Physiol. 125:1198–1205.
Krishnan HB, Bennett JO, Kim WS, Krishnan AH, Mawhinney TP. 2005. Nitrogen lowers the sulfur amino acid content of soybean (*Glycine max* [L.] Merr.) by regulating the accumulation of Bowman-Birk protease inhibitor. J Agric Food Chem. 53:6347–6354.
Krishnan HB, White JA. 1995. Morphometric analysis of rice seed protein bodies (implication for a significant contribution of prolamine to the total protein content of rice endosperm). Plant Physiol. 109:1491–1495.
Lester RN, Bekele E. 1981. Amino acid composition of the cereal Tef and related species of Eragrostis (Gramineae). Cereal Chem. 58:113–115.
Loit E, Melnyk CW, MacFarlane AJ, Scott FW, Altosaar I. 2009. Identification of three wheat globulin genes by screening a *Triticum aestivum* BAC genomic library with cDNA from a diabetes-associated globulin. BMC Plant Biol. 9:93.
Miclaus M, Xu J-H, Messing J. 2011. Differential gene expression and epiregulation of alpha zein gene copies in maize haplotypes. PLoS Genet. 7:e1002131.
Morey DD, Evans JJ. 1983. Amino-acid-composition of 6 grains and winter-wheat forage. Cereal Chem. 60:461–464.
Nagamine A, et al. 2011. A role for the cysteine-rich 10 kDa prolamin in protein body I formation in rice. Plant Cell Physiol. 52:1003–1016.
Nakase M, et al. 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.
Ogawa M, et al. 1987. Purification of protein body-i of rice seed and its polypeptide composition. Plant Cell Physiol. 28:1517–1527.
Osborne TB. 1908. Our present knowledge of plant proteins. Science 28:417–427.
Payne PI, Holt LM, Jackson EA, Law CN, Damania AB. 1984. What storage proteins: their genetics and their potential for manipulation by plant breeding [and discussion]. Philos Trans R Soc Lond Series B Biol Sci. 304:359–371.
Reyes FC, et al. 2011. Delivery of prolamins to the protein storage vacuole in maize aleurone cells. Plant Cell 23:769–784.
Rubin R, Levanony H, Galili G. 1992. Evidence for the presence of two different types of protein bodies in wheat endosperm. Plant Physiol. 99:718–724.
Shewry PR, Casey R. 1999. Seed Proteins. In: Shewry PR, editor. Seed proteins. Netherlands: Springer. p. 1–10.
Shewry PR, Halford NG. 2002. Cereal seed storage proteins: structures, properties and role in grain utilization. J Exp Bot. 53:947–958.
Shewry PR, Napier JA, Tatham AS. 1995. Seed storage proteins: structures and biosynthesis. Plant Cell 7:945–956.
Shewry PR, Pandya MJ. 1999. The 2S albumin storage proteins. In: Shwery PR, Casey R, editors. Seed proteins. New York: Springer. p. 563–583.
Shorrosh BS, et al. 1992. A novel cereal storage protein: molecular genetics of the 19 kDa globulin of rice. Plant Mol Biol. 18:151–154.
Song R, Llaca V, Linton E, Messing J. 2001. Sequence, regulation, and evolution of the maize 22-kD alpha zein gene family. Genome Res. 11:1817–1825.
Song R, Messing J. 2002. Contiguous genomic DNA sequence comprising the 19-kD zein gene family from maize. Plant Physiol. 130:1626–1635.
Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.
Tatham AS, et al. 1996. Characterisation of the major prolamins of Tef (*Eragrostis tef*) and Finger Millet (*Eleusine coracana*). J Cereal Sci. 24:65–71.
Thompson GA, Larkins BA. 1994. Characterization of zein genes and their regulation in maize endosperm. In: Freeling M, Walbot V, editors. The maize handbook. New York: Springer. p. 639–647.
Tosi P, et al. 2009. Trafficking of storage proteins in developing grain of wheat. J Exp Bot. 60:979–991.
Tranquilli G, Heaton J, Chicaiza O, Dubcovsky J. 2002. Substitutions and deletions of genes related to grain hardness in wheat and their effect on grain texture. Crop Sci. 42:1812–1817.
Vincentini A, Barber JC, Aliscioni SS, Giussani LM, Kellogg EA. 2008. The age of the grasses and clusters of orignins of C4 photosynthesis. Global Change Biol. 14:2963–2977.
Wallace NH, Kriz AL. 1991. Nucleotide sequence of a cDNA clone corresponding to the maize globulin-2 gene. Plant Physiol. 95:973–975.
Wilkinson MD, Castells-Brooke N, Shewry PR. 2013. Diversity of sequences encoded by the GSP-1 genes in wheat and other grass species. J Cereal Sci. 57:1–9.
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 Y, Goettel W, Messing J. 2009. Non-Mendelian regulation and allelic variation of methionine-rich delta-zein genes in maize. Theor Appl Genet. 119:721–731.
Wu Y, Messing J. 2010. RNA interference-mediated change in protein body morphology and seed opacity through loss of different zein proteins. Plant Physiol. 153:337–347.
Wu Y, Messing J. 2012. RNA interference can rebalance the nitrogen sink of maize seeds without losing hard endosperm. PLoS One 7:e32850.
Wu Y, Wang W, Messing J. 2012. Balancing of sulfur storage in maize seed. BMC Plant Biol. 12:77.
Xu JH, Bennetzen JL, Messing J. 2012. Dynamic gene copy number variation in collinear regions of grass genomes. Mol Biol Evol. 29:861–871.
Xu JH, Messing J. 2008. 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 U S A. 105:14330–14335.
Xu JH, Messing J. 2009. Amplification of prolamin storage protein genes in different subfamilies of the Poaceae. Theor Appl Genet. 119:1397–1412.
Zhang W, Sangtong V, Peterson J, Scott MP, Messing J. 2013. Divergent properties of prolamins in wheat and maize. Planta 237:1465–1473.

Associate editor: Brandon Gaut