Genome-wide identification and characterization of seed storage proteins (SSPs) of foxtail millet (*Setaria italica* (L.) P. Beauv.)

Vikram Singh Gaur¹ · Salej Sood² · Sharad Tiwari³ · Anil Kumar⁴

Received: 2 November 2017 / Accepted: 10 September 2018 / Published online: 17 September 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

We report the identification of 47 foxtail millet (*Setaria italica* (L.) P. Beauv.) seed storage proteins (SSPs) consisting of 14 albumins, 12 prolamins, 18 globulins and 3 glutelins using computational approaches and compared their essential amino acid composition with 225 SSPs of rice, barley, sorghum and maize. Comparative analysis revealed several unique foxtail millet SSPs containing high amounts of essential amino acids. These include three 2s-albumin proteins containing 11.9%, 10.9%, 9.82% lysine, one 10-kDa prolamins containing 20% methionine residues and one each 7S-globulin, 10-kDa prolamins, alpha-zein proteins containing 9.2% threonine, 9.35% phenylalanine and 2.5% tryptophan, respectively. High lysine containing albumins and high methionine containing prolamins were also detected in other cereals indicating that these SSPs are widespread in cereals. Phylogenetic studies revealed that the foxtail millet SSPs are closer to sorghum and maize. The lysine-rich albumins and the methionine-rich prolamins formed a separate cluster. Motif analysis of lysine-rich albumins displayed several lysine containing conserved motifs across cereals including foxtail millet. The 10-kDa prolamins protein containing 20% methionine was unique as it lacked the characteristic repeat motifs of methionine found in the high methionine containing zeins and kafirins. The motif “NPAAFWQQQLL” was uniquely repeated in the foxtail millet high tryptophan prolamins protein. The findings of the present study provide new insights in foxtail millet seed storage protein characterization and their nutritional importance in terms of essential amino acid composition.

Keywords Seed storage proteins · Foxtail millet · 2s-Albumin · Prolamin · Globulin · Glutelin

Abbreviations

SSP	Seed storage proteins
kDa	Kilo Dalton
HMM	Hidden Markov models
MEME	Multiple Em for motif elicitation

RGAPD	Rice Genome Annotation Project Database
Myr	Million year

Introduction

Cereals and legumes are the most important crops in the world. They are not only the source of energy, but also provide proteins for the nutrition of humans, animals and livestock. Since proteins derived from animal sources are expensive, people in developing countries virtually depend on seed protein alone for their entire protein requirement (Mandal and Mandal 2000). Seed storage proteins (SSPs) are specifically synthesized at high levels in the developing seed (in cotyledon in case of pulses or in endosperm in case of cereals) with temporal and spatial control of synthesis during endosperm development. Cereal SSPs are deposited mostly in special storage organelles called protein bodies and accumulate in high amounts during mid-maturation stage of seed development. Subsequently, the SSPs are used up during germination. Based on the solubilities in solvents such as

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-018-1431-8>) contains supplementary material, which is available to authorized users.

✉ Vikram Singh Gaur
vikramsinghgaur@gmail.com

- ¹ College of Agriculture, Waraseoni, Balaghat, JNKVV, Jabalpur 481331, India
- ² Division of Crop Improvement, ICAR-Central Potato Research Institute, Shimla, Himachal Pradesh 171001, India
- ³ Department of Plant Breeding and Genetics, JNKVV, Jabalpur 482004, India
- ⁴ Department of Molecular Biology and Biotechnology, GB Pant University of Agriculture and Technology, Pantnagar 263145, India

water, dilute saline, alcohol–water mixtures, and dilute alkali and acid (Shewry and Halford 2002), the storage proteins have been classified into four groups namely—albumins, globulins, prolamins, and glutelins. Albumins and globulins comprise the major storage proteins of dicots (e.g. pulses), whereas except oats (*Avena sativa*) and rice (*Oryza sativa*), the major endosperm storage proteins of all cereal grains including maize (*Zea mays*), sorghum (*Sorghum bicolor*), coix (*Coix lacryma-jobi*) and millets are the water–alcohol mixture soluble prolamins, which accounts for more than 50% of the total protein. The 2S albumins are widely distributed in both mono- and di-cotyledonous plants (Youle and Huang 1981), and have been reported to play important roles in plant defence against fungal attack (Agizzio et al. 2006). Because of the presence of high content of sulphur-containing amino acids, the 2S albumins are of increasing interest in nutritional and clinical studies. However, in recent years, some members of this protein family have been described as major food allergens. The globulin SSPs are broadly categorized into two types (1) the 7S globulins—present in the embryo and outer aleurone layer of the endosperm (Wallace and Kriz 1991) and (2) the 11–12S globulins located in the starchy endosperm. The 7S globulins have limited sequence similarity with the 7S vicilins of legumes and other dicotyledonous plants, and have similar structures and properties. Related proteins have been found in embryos and/or aleurone layers of wheat, barley and oats (Shewry and Halford 2002). Since, aleurone and embryo is usually removed by milling (wheat), polishing (rice), pearling (barley) or decortication (sorghum), before human consumption, the globulins in these tissues have limited impact on the end use properties of the grain. In oats and rice, the 11–12S storage globulins form the major endosperm storage protein fraction and are related to the widely distributed ‘legumin’ type globulins, which occur in most dicotyledonous species (Casey 1999). The rice proteins classically defined as glutelins belong to the 11–12S globulin family. Based on the primary sequence comparisons, the glutelins have been classified into A and B types (Takaiwa et al. 1991). The rice B-type glutelin has more lysine content, and therefore has been suggested as a good genetic resource to improve rice protein quality. The prolamins are the dominant class of seed storage protein in many cereals and are characterized as having a high glutamine and proline content. Prolamins are named differently in different cereals such as in wheat (gliadin), barley (hordein), rye (secalin), corn (zein), sorghum (kafirin) and as avenin in oats. The rice, barley, maize and sorghum prolamins accounts for about 35%, 50%, 60%, 80% of the total endosperm protein, respectively. Although prolamins were initially distinguished as a group of proteins soluble in 70% ethanol (Osborne 1897), the differences in aqueous solubility, ability to form disulfide interactions, molecular weight, gene sequence were subsequently used to classify prolamins

sub-families. For example, the zeins and the kafirins are grouped into α , β , γ and δ types (Holding 2014). Nutritionally, the 13-kDa rice prolamin polypeptide has a higher content of leucine, glutamic acid and aspartic acid and a low content of lysine and sulphur-containing amino acids, whereas the 10-kDa and 16-kDa polypeptides have a higher content of sulphur-containing amino acids (Mitsukawa et al. 1999). The S-rich (sulphur rich) B-hordeins of barley have very high glutamine–proline content and relatively higher cysteine content, while the S-poor (sulphur poor) C-hordeins lack cysteine and have very low methionine and lysine content (Shewry et al. 1981). Similarly, the wheat α , β and γ -gliadins, oat α , β and γ -avenins and the maize α , β , γ and δ zeins are poor in lysine, arginine, histidine and tryptophan content (Matta et al. 2009). Overall, the low content of lysine and tryptophan in the prolamin fractions makes the prolamin proteins inferior in nutritional quality.

The identification of nutritionally superior SSPs (such as methionine rich) has opened up possibilities to engineer crops to provide higher essential amino acid content in human and animal diets (Chakraborty et al. 2000; Molvig et al. 1997; Tabe et al. 1995). Nucleotide sequences of several genes encoding seed storage proteins from important cereals are now available in the public databases and the list is increasing with the sequencing of cereal genomes.

Recently, the foxtail millet (*Setaria italica* (L.) P. Beauv) genome has been sequenced (Zhang et al. 2012) and its genome sequence is available. Since, millets are known for their nutraceutical value and quality proteins (Kumar et al. 2015, 2016b; Gupta et al. 2017), the availability of the genome sequence provides a unique opportunity to identify and characterize its seed storage proteins. In this study, we report the identification SSPs of foxtail millet, access their nutritional potential and show their phylogenetic relationship with the seed storage proteins of other cereals.

Materials and methods

Retrieving seed storage protein sequences of rice, maize, barley and sorghum

Protein sequences of seed storage proteins of rice were downloaded from “Rice Genome Annotation Project Database” (RGAPD) (<http://rice.plantbiology.msu.edu/>) and the SSPs of maize, barley and sorghum were downloaded from NCBI. For more comprehensive information about the SSPs of maize, barley and sorghum, the genome-wide protein sequence file was downloaded from their respective genome databases, viz., for maize—ftp://ftp.ensemblgenomes.org/pub/plants/release-35/fasta/zea_mays, for barley—ftp://ftp.ensemblgenomes.org/pub/plants/release-35/fasta/hordeum_vulgare, and for sorghum—ftp://ftp.ensemblgenomes.org/pub/plants/release-35/fasta/sorghum_bicolor.

org/pub/plants/release-35/fasta/sorghum_bicolor. Using the tool HMMER3.1 and the protein sequences obtained from NCBI and RGAPD, individual HMM profiles of each SSPs protein family were generated using “hmmbuild” program. Using the “hmmsearch” program hmmprofiles of individual protein family used as query against the maize, sorghum and barley protein databases and protein hits were recovered. Other hmm programs such as the “phmmer” and “jackhammer” were also used to detect homologous target SSP protein sequences in the respective target databases. To filter out the nonspecific SSPs, the protein sequence hits were further analysed using the BLAST2GO software and the protein sequences showing homologies with known SSPs were identified and stored in separate files. Finally duplicate sequences were removed by running a simple perl script. The final dataset of SSP is shown in Table 1.

Retrieving foxtail millet genomic information and identification of foxtail millet SSP sequences

Foxtail millet (*Setaria italica* strain Yugu1) individual chromosome sequence information, CDS (coding DNA sequence) and predicted protein sequences were downloaded from NCBI-FTP resource (ftp://ftp.ncbi.nlm.nih.gov/genomes/Setaria_italica/). As described before, using the HMMER3.1—hmmbuild program the non-redundant dataset generated was again used to generate SSP hmmprofiles of each SSP protein family, viz., prolamin, globulin (7S, 11S and 12S), 2S-albumin, vicilin, con-vicilin, legumin, zein, glutelin, glutenin, hordeins and was subsequently used to identify homologous foxtail millet SSPs using the

“hmmsearch” program. The BLAST-like program “phmmer”, which searches a single protein sequence against a protein sequence database (foxtail millet protein sequence database), and the PSI BLAST-like program “jackhammer”, which iteratively searches a protein sequence against a protein sequence database were also used to detect homologues target SSP protein sequences. The hits were filtered using the BLAST2GO software and the annotated protein sequences showing homologies with known SSPs were identified and saved.

Determination of the percentage of each amino acid in each SSP sequence

The retrieved fasta formatted seed storage protein sequences were saved in a single txt file (.txt) and batch counting of individual amino acids in a protein sequence was carried out using the ‘Biostring’ package of Rstudio ver3.2.4 (revised). The output file was managed using MS-Excel and percentage of individual amino acids in each protein sequence was calculated using a simple formula (number of amino acids in a protein sequence*100)/total number of amino acids).

Statistical analysis

The classification of seed storage proteins dataset based on their percent individual essential amino acid content was carried out by *k*-mean clustering method. The R-packages “psych”, “stats”, “mclust” and “rattle” were used for the analysis. The Kaiser–Meyer–Olkin (KMO) test [overall measures for sampling adequacy (MSA)] for the set of

Table 1 Total number of seed storage proteins of different cereals included in the present investigation

	Albumin	Globulin	Prolamin	Glutelin
Rice	6 (TIGR)	3 (7S globulin-2, globulin2-1) (TIGR)	29 (TIGR)	15 (TIGR)
Sorghum	16 (identified in this study)	7 (legumin-3, globulin-4) (NCBI)	16 (NCBI)	0
Maize	34 (identified in this study)	14 (7S-3, globulin-1s allele-5, legumin-5, vicilin-1) (NCBI)	16 (NCBI)	0
Barley	21 (NCBI-2, identified in this study-19)	16 (7S-7, 11S-3, legumin-2, vicilin-2, globulin-1s-allele-1, embryo-globulin-1) (NCBI)	28 (NCBI-prolamin-2, NCBI-hordein-22, identified in this study-avenin-3, farinin-1)	2 (NCBI-1, identified in this study-1)
Fox tail millet (identified in the present study)	14	18 (7s-11, 11S-2, globulin-1S-allele-4, 19-kDa-globulin-1)	12 (prolamin-4, zein-8)	3

All rice SSPs were retrieved from the TIGR database. All globulins of sorghum, maize and barley, and all prolamins of sorghum and maize were retrieved from NCBI database. Two (2) prolamins, 22 hordeins and 1 glutelin protein of barley were also retrieved from NCBI database. Sixteen, 34 and 19 albumin proteins of sorghum, maize and barley, respectively, along with 3-avenin and 1 farinin proteins of barley were identified in the present study. Hmm profiles of the resulting protein data of rice, sorghum, maize and barley was subsequently used to identify the foxtail millet SSPs. The details of each SSP are given in the parenthesis

variables included in the analysis was 0.543, which exceeded the minimum requirement of 0.50 for overall MSA. The probability associated with the Bartlett test was <0.001 , which satisfied the requirement for carrying out principal component analysis (PCA) to determine the number of clusters in the dataset. The PCA revealed 4 eigenvalues greater than 1 explaining cumulative variance of about 64%. Hence, the SSPs were classified by taking the value of $k=4$. The ANOVA test for all the k -mean clustering analysis revealed that the clusters were significantly different ($p < 0.001$) from each other.

Phylogenetic analysis of foxtail millet SSPs with rice SSPs

Phylogenetic relationship of the identified SSPs of foxtail millet was studied with the SSPs of rice, sorghum, maize and barley. To better understand the phylogeny, individual SSP family-specific tree (albumin, globulin, prolamin and glutelin) was studied. The identified foxtail millet seed storage protein sequences, as well as that of rice, sorghum, barley and maize along with high methionine content SSPs earlier reported in rice, maize, sorghum were aligned using the ClustalW multiple sequence alignment tool available online (<http://www.genome.jp/tools/clustalw/>, Kyoto University Bioinformatics Centre). The resultant tree file (with the extension .dnd) was downloaded and a phylogenetic tree was constructed using the neighborhood joining method using the software MEGA Ver. 7.0.20 (Kumar et al. 2016a).

Domain and motif analysis of protein sequences

Since seed storage proteins contain characteristic and unique domains, the foxtail millet protein sequences found to contain high percentage of essential amino acids were analysed using Uniprot (<http://www.uniprot.org/>) and the protein domain finding tool “SMART” (Letunic et al. 2015). To find repeating conserved motif units in the group of proteins sharing high percentage of essential amino acids, the MEME (multiple EM for motif elicitation) version 4.11.3 program (Timothy et al. 2009) was used with the following parameters: minimum motif width was set to 4 while the maximum motif was kept as 12 with any number of motifs repetitions. Maximum number of motifs detection was restricted to 10. Consensus motifs were generated using MAST (Motif Alignment & Search Tool, http://meme.sdsc.edu/meme4_1/cgi-bin/mast.cgi) (Timothy and Michael 1998).

Mapping and visualizing the identified seed storage protein genes onto foxtail genome

To map and visualize the location of all the identified foxtail millet SSPs genes on foxtail millet genome, Genome

Blast search was performed using individual SSPs coding DNA sequences (CDS) as query sequences. Using the RFLP genetic map of *Setaria italica* (annotation release 100), the physical location of the blast hits and the map drawing software “MapChart 2.30”, the foxtail millet chromosomes were constructed and SSP genes were mapped.

Results

Retrieving rice, sorghum, maize, barley SSPs and identification of foxtail millet seed storage protein sequences

Using the public databases, 53 SSPs of rice consisting of 6 albumin, 29 globulin and 15 glutelin, 23 SSPs of sorghum consisting of 7 globulin and 16 prolamin, 30 SSPs of maize consisting of 14 globulin and 16 prolamin proteins, and 41 SSPs of barley consisting of 2 albumin, 16 globulin, 24 prolamin, 1 glutelin proteins were identified. Using the hmm profiles of individual SSP family and individual SSPs as query, 16 albumin of sorghum, 34 albumin of maize, and 19 albumin, 3 avenin, 1 farinin and 1 glutelin protein of barley were identified. The number and type of SSPs in each of these crops are shown in Table 1. The new proteins identified were again used to generate hmm profiles and were subsequently used as query to identify SSPs of foxtail millet. In this way, 47 SSPs consisting of 14 albumin, 18 globulin (11-7S-globulin, 2-11S globulin, 4-globulin-1S-allele, 1-19-kDa-globulin), 12 prolamin (4-prolamin and 8 Zein like) and 3 glutelin proteins were identified in foxtail millet.

Determination of percent individual essential amino acid content in foxtail millet seed storage proteins

Percent essential amino acid content in the albumins, prolamins, glutenins and globulins of foxtail millet and other cereals is shown in the heatmaps (Figs. 1b, 2b, 3b, 4b). Complete data on percentage of essential amino acids in individual proteins is given in the supplementary file 1. The analysis revealed several foxtail millet SSPs containing high amounts of essential amino acids. Two foxtail millet albumin proteins showed more than 10% (10.9% and 11.92%) lysine content while one albumin protein was found to contain 9.8% lysine residues (Table 2). One 10-kDa prolamin (XP_004983434.1) was found to contain 20% methionine (Table 3). Foxtail millet SSPs were also found to contain high percentage of tryptophan residues. Five foxtail millet SSPs exhibited more than 2% tryptophan residues (Table 4). The foxtail millet alpha-zein protein (XP_004962290.1) was found to contain 2.5% tryptophan residues. Further, a highest phenylalanine (9.4%) containing 10-kDa prolamin (XP_004965361.1) of foxtail millet was also detected. Besides these important

Fig. 1 **a** Phylogenetic relationship of foxtail millet albumin proteins with the albumin proteins of other cereals. **b** Heatmap showing the individual essential amino acid content in the albumin proteins of different cereals

Fig. 2 **a** Phylogenetic relationship of foxtail millet prolamins with the prolamins of other cereals. **b** Heatmap showing the individual essential amino acid content in the prolamins of different cereals

Fig. 4 **a** Phylogenetic relationship of foxtail millet globulin proteins with the globulin proteins of other cereals. **b** Heatmap showing the individual essential amino acid content in the globulin proteins of different cereals

Table 2 Seed storage proteins of different cereals containing highest percentages of lysine residues (shown in descending order). The protein sequences of these SSPs were further analysed through the online motif finding tool “MEME” to find repeating motifs within the protein sequence as well as conserved motifs across SSPs

Amino acid	Crop	Type	Seq ID	% of EAA
Lysine	Sorghum	Albumin	Sb08g005340.1	14.1
	Sorghum	Albumin	Sb08g005360.1	12.1
	Foxtail	Albumin	NP_001274467.1	11.9
	Barley	Albumin	BAK05648.1	11.2
	Barley	Albumin	Hv_MLOC_1972.1	11.2
	Sorghum	Albumin	Sb03g037220.1	10.8
	Maize	Albumin	GRMZM2G087413_P01	10.3
	Foxtail	Albumin	XP_004970265.1	10.1
	Maize	Albumin	GRMZM2G091054_P01	9.9
	Foxtail	Albumin	XP_012702999.1	9.8
	Barley	Albumin	Hv_MLOC_62188.1	9.4
	Maize	Albumin	GRMZM2G094632_P01	8.6

Foxtail millet proteins included in the study are indicated in bold

Table 3 Seed storage proteins of different cereals containing highest percentages of methionine residues (shown in descending order). The protein sequences of these SSPs were further analysed through the online motif finding tool “MEME” to find repeating motifs within the protein sequence as well as conserved motifs across SSPs

Amino acid	Crop	Type	Seq ID	% of EAA
Methionine	Maize	Zein	U31541.1	24.9
	Maize	18kD delta zein	AF371265.1	23.7
	Maize	10kD delta zein	AF371266.1	20.7
	Maize	delta zein	AY104139.1	20.7
	Maize	10 kDa zein	M23537.1	20.0
	Foxtail	10 kDa Prolamin	XP_004983434.1	20.0
	Sorghum	Delta kafirin	ACM68402.1	19.9

Foxtail millet proteins included in the study is indicated in bold

Table 4 Seed storage proteins of different cereals containing highest percentages of tryptophan residues (shown in descending order). The protein sequences of these SSPs were further analysed through the online motif finding tool “MEME” to find repeating motifs within the protein sequence as well as conserved motifs across SSPs

Amino acid	Crop	Type	Seq ID	% of EAA
Tryptophan	Rice	Albumin	LOC_Os03g55730.1	2.6
	Sorghum	Albumin	Sb08g005340.1	2.6
	Foxtail	Zein-alpha	XP_004962290.1	2.5
	Maize	Alpha globulin	AF371278.1	2.4
	Maize	Albumin	GRMZM2G094632_P02	2.4
	Maize	Albumin	GRMZM2G130454_P01	2.3
	Barley	Albumin	Hv_MLOC_67523.1	2.3
	Foxtail	Zein beta	XP_004960933.1	2.2
	Foxtail	Prolamin	XP_004960634.1	2.2
	Foxtail	10 kDa prolamin	XP_004965361.1	2.2
	Barley	Vicilin like	Hv_MLOC_57363.1	2.1
	Sorghum	Globulin	XP_002441331.1	2.1
	Foxtail	Alpha-zein	XP_004979702.1	2.1

Foxtail millet proteins included in the study are indicated in bold

which showed species-specific sequence conservation, the foxtail millet prolamins showed similarities with prolamins of other cereals (Fig. 2a). Similar to clustering of high lysine-containing albumin proteins, a separate and distinct group of prolamin proteins having high methionine content was also formed. The 10-kDa foxtail millet prolamin protein Si_10_kDa_XP_004983434.1 containing 20%, methionine was also present in this cluster. The glutelin proteins also showed similar species-specific clustering (Fig. 3a). In contrast to the albumin, prolamin and glutelin proteins, which showed species-specific clustering, the globulin proteins were somewhat evenly distributed in 5 distinct groups, each containing at least one globulin protein of the cereals under study (Fig. 4a).

Table 5 Foxtail millet seed storage proteins containing highest percentages of essential amino acids

Species	Gene bank accession no.	Type	Dominant amino acid (%)	Overall total essential amino acid (%)	Hydropathy index
<i>S. italica</i>	NP_001274467.1	2S-albumin	Lysine (11.92%) Valine (11.92%)	47.7	0.2
<i>S. italica</i>	XP_004970265.1 (EST-JK572994.1)	2S-albumin	Lysine (10.9%)	46.3	0.3
<i>S. italica</i>	XP_012702999.1	2S-albumin	Lysine (9.82%)	38.3	0.12
<i>S. italica</i>	XP_004983434.1	10-kDa prolamin	Methionine (20.0%)	45	0.432
<i>S. italica</i>	XP_004965528.1	7S-globulin	Threonine (9.16%)	42.5	0.316
<i>S. italica</i>	XP_004965361.1	10-kDa prolamin	Phenylalanine (9.35%)	44.6	0.192
<i>S. italica</i>	XP_004962290.1	Alpha-zein	Tryptophan (2.5%)	36.2	0.143

Fig. 5 Amino acid sequences of all the identified foxtail millet seed storage proteins having high ratios of essential amino acids. The highest occurring amino acid is highlighted with bold font. The protein

domains present in each protein as detected by the protein domain finding tool “SMART” is also shown

Motif analysis of SSPs containing high ratios of essential amino acids

In case of high lysine-containing albumin proteins, 7 out of 10 detected conserved motifs (PFEHGYKCGSYT, VCMEKVVYVANY, NPKJPPSDACC, CLCSKVTKIEI, LIRECKQYVMFP, MAKLLLLFLIJL, VWKKABIP) were lysine-containing motifs (Fig. 7). The positions and composition of motifs across the proteins were highly

conserved. The two foxtail millet high lysine-containing albumin proteins Si_Albumin_NP_001274467.1 (11.9% K) and Si_Albumin_XP_004970265.1 (10.1% K) had similar motif compositions while in the third albumin Si_Albumin_XP_012702999.1 (9.8% K) many of the conserved motifs were missing. The conserved motifs containing methionine residues detected in high methionine-containing SSPs were CMQSCMMQQLLA, PSMMPMM, AAKMLALFALLA, PQCHDAVSQIM,

Fig. 6 Clustering of cereal seed storage proteins according to the percentages of essential lysine, methionine, tryptophan and total essential amino acid content. Clustering was performed using R-packages

“psych”, “stats”, “mclust” and “rattle”. *k*-mean clustering was carried out at $k=4$. The clusters are significantly different at $p<0.001$. Members of each cluster is given in the supplementary file 2

Fig. 7 Repeated motifs within individual SSPs and conserved motifs across different seed storage proteins rich in lysine

QQLPFMFNPTAM, PLGTMTN and NFPHMMGIGMMD. The motif “PSMMPPMM” was repeated 9 times in the proteins Zm_Zein_U31541.1 (24.9% M) and Zm_18kD_delta_zein_AF371265.1 (23.7% M) while it was repeated 5 times in the proteins Zm_10kD_delta_zein_AF371266.1

(20.7% M), Zm_AY104139.1 (20.7% M), and Zm_10_kDa_zein_M23537.1 (20% M). The foxtail millet protein Si_10_kDa_Prolamin_XP_004983434.1 with 20% methionine however, showed a different pattern and was found to contain two copies of the motif “CMQSCMMQQLLA”,

one copy each of the motif “QQLPFMFNPTAM” and “NFPHMMGIGMMD” (Fig. 8). In SSPs containing high percentage of tryptophan residues, only 3 out of the 10 detected conserved motifs (NPAAFWQQQQLL, EEAMPPLEKGGW, FRWGTGLRMRCC) were found to contain tryptophan residues. Most striking was the motif “NPAAFWQQQQLL”, which was specifically present only in the foxtail millet proteins Si_Zein-alpha_XP_004962290.1 (2.5% trp residues) and Si_Zein-alpha_XP_004979702.1 (2.1% trp residues) in 6 and 8 copies, respectively (Fig. 9). Similar to motifs rich in methionine, motif rich in phenylalanine were also detected in the SSPs containing high percentage of phenylalanine (Supplementary Figure 1 & Supplementary Table 1). The foxtail millet 10 kDa prolamin protein (XP_004965361.1) containing 9.4% phenylalanine did not show any repeat motif. Only a single copy of phenylalanine containing motif

“QQPFPLQPQQPF” was detected. High leucine-containing SSPs showed the presence of many conserved motifs containing leucine residues. The motif “AYLQQQQLL-PFN” was repeated 5 times each in foxtail millet SSPs Si_Zein_alpha_XP_004979702 (containing 17.1% leu) and Si_Zein_alpha_XP00497904 (containing 18.4% Leu) (Supplementary Figure 2, Supplementary Table 2). Although, motifs containing multiple residues of valine and threonine were detected, the number of copies was restricted to one or two only (Supplementary Figure 3 & 4 and Supplementary Tables 3 & 4). Conserved motifs and motif repeats of SSPs rich in isoleucine and histidine are shown in the supplementary Figs. 5 & 6 and supplementary Table 5 and Table 6. Overall, it appears that the percentage of the essential amino acid in any protein does not strongly correlate with the number of repeated motifs containing more number of essential amino acid residues.

Fig. 8 Repeated motifs within individual SSPs and conserved motifs across different seed storage proteins rich in methionine

Fig. 9 Repeated motifs within individual SSPs and conserved motifs across different seed storage proteins rich in tryptophan

This indicates that probably the SSPs having high percentages of EAA either have random insertions of EAA in protein sequence or have unique EAA rich motifs.

Mapping and visualizing the identified seed storage protein genes onto foxtail millet genome

At present, the physical map and the genetic map of *Setaria italica* has not been integrated. Therefore, using the physical location (in terms of base pairs distance from the chromosome end) of the SSP genes resulting from the genome blast and the RFLP genetic map, all the SSP genes were mapped onto foxtail millet genome. Five albumin SSP genes were mapped on chromosome 5, while 3 were mapped on each chromosome 2, 7 and 9. Three, four and two prolamin SSP genes were mapped on chromosomes 3, 8 and 9, respectively, whereas one prolamin SSP was mapped on chromosome 4. One glutelin SSP was mapped on each chromosome 1 and 2, while two SSPs were mapped on chromosome 3. Compared to the other SSP gene families, the globulin SSP genes were spread throughout the genome. Five globulin

SSP genes were found to be located on the chromosome 5, while 4 genes were mapped each on chromosomes 3 and 9. Two globulin SSP genes were mapped on chromosome 4, while one each was located on chromosome 7 and 8. No SSPs were mapped on chromosome 6. The locations of the identified foxtail millet SSP genes are shown in Fig. 10.

Discussion

Foxtail millet seed storage proteins

To identify seed storage proteins of foxtail millet using homology approach, we initially collected the available SSPs of rice, maize, barley and sorghum from various public databases. Although different seed storage family members (albumins, globulins, prolamins and glutenins) of rice were retrieved, only a few members of seed storage proteins of maize, barley and sorghum could be retrieved. Therefore, to generate more comprehensive HMM profiles of SSPs for detecting target SSPs of foxtail millet, additional seed

Fig. 10 Putative map positions of the identified foxtail millet seed storage protein (SSP) genes. RFLP markers and their map distance (in cm) on each chromosome is shown. The scale on the left side of

each chromosome shows the physical length of each chromosome (expressed in Mb—mega base pairs)

storage proteins of sorghum, maize and barley were identified. Using the HMM profiles of the initially identified 6 and 2 albumin proteins of rice and barley additional 16, 34 and 19 albumin proteins of sorghum, maize and barley, respectively, were identified. Similarly, three avenin-like and one farinin-like proteins of barley were identified. Finally, using HMM profiles of the newly identified SSPs, 47 non-redundant SSPs of foxtail millet consisting of 12 prolamin, 18 globulin, 14 albumin and 3 glutelin proteins were identified. The identified SSPs were single copy SSPs that mapped onto unique locations on the foxtail millet genome. The map locations (Fig. 10) of these SSPs might be useful in understanding the evolution of these unique proteins and to understand their syntenic relationships with the SSPs of other cereals.

Phylogenetic relationship of foxtail millet SSPs with the SSPs of other cereals

Sequence alignment of the foxtail millet SSPs to SSPs of rice, maize, barley and sorghum revealed that, in general, the strongest conservation occurred within species. Strongest within-species conservation was observed in the albumins of maize and rice and to some extent barley, and the prolamins of rice and barley. Unlike the albumins of rice and maize, the foxtail millet albumins did not cluster together and showed higher similarities with sorghum than barley or maize. Similarly, the foxtail millet prolamins showed closeness to the sorghum prolamins. The closeness of foxtail millet SSPs with sorghum and maize could be expected as foxtail millet is phylogenetically closer to sorghum and maize (Brutnell et al. 2010). Moreover, the genus *Setaria* is a member of the tribe Paniceae closely related to Andropogoneae, which includes maize and sorghum. *Setaria* last shared a common ancestor with maize and sorghum ~ 26 Myr ago, a rather more recent event than with rice and Brachypodium, which is ~ 52 Myr ago (Bennetzen et al. 2012). Further, like sorghum and maize, *Setaria* uses C4 type of photosynthesis. The high conservedness shown by the prolamins and albumins might be because of the fact that the seed storage protein gene families have expanded in a lineage-specific manner through gene duplication (Xu and Messing 2008). However, unlike the prolamins and albumins proteins, no species-specific cluster of globulin proteins were observed. The globulin proteins were distributed in different groups each containing globulin proteins of each cereal. Homologous foxtail millet globulin proteins could be observed in each cluster. It appears that the globulin proteins are more diverged compared to prolamins and albumins. This might be possible because the globulins proteins, which are a part of the cupin superfamily (Dunwell et al. 2004), are the most widely distributed group of storage proteins evolved from bacterial enzymes. The globulins are present in both dicots and monocots (Shewry et al. 1995). Separate groups

containing homologous globulin proteins of different cereals indicate the presence of a common ancestor before their divergence into separate lineages. The remarkable clustering of high lysine-containing albumins and high methionine-containing prolamins into a separate cluster suggests that at least one high lysine and methionine-containing protein is present in all the cereals, which might be evolutionarily related and share a common ancestor.

Essential amino acid and motif composition of foxtail millet compared to the SSPs of other cereals

The amino acid composition of many foxtail millet SSPs was found to be among the highest compared to the SSPs of other cereals. The present investigation shows that compared to other SSPs the albumin proteins contain relatively higher percentage of lysine residues. Except the albumins of rice, 8 albumin proteins of different cereals including the 2 from foxtail millet were found to contain lysine residues more than 10%. The motif analysis revealed that the high lysine percentage was as a result of the presence of different conserved motifs containing lysine residues. As naturally occurring SSPs rich in lysine have not been reported (Yue et al. 2014; Lang et al. 2004; Jiang et al. 2016), the high lysine-containing albumin proteins identified in the present study appear to be good alternatives. However, since albumin proteins have been reported to be allergic, it is necessary to assess the allergenicity of the protein first before using it as a transgene (Moreno and Clemente 2008). In addition to the nutritional aspects, presence of high-lysine albumins might provide additional advantage to biotic and abiotic stress tolerance in plants (Cândido and Pinto 2011; Galili et al. 2001). Unlike a number of high-lysine albumin proteins identified in foxtail millet, only one high methionine-containing SSP (10 kDa prolamin) was identified as compared to the high methionine-containing SSPs reported in other crops (Izquierdo and Godwin 2005). It appears that at least one high methionine-containing SSP is present in each cereal. Motif analysis revealed the absence of the characteristic repeated units of methionine-rich motif “PSMMPPMM” in the high methionine foxtail millet SSP, which was present predominantly in high methionine zeins and kafirins. This indicates that the high methionine-containing maize zeins are evolutionarily more conserved while the high methionine kafirins have diverged and lost some of the motif repeats. The absence of any of the repeats in the foxtail millet high methionine-containing foxtail millet indicates that the protein probably has diverged and is different compared to other high methionine-containing SSPs.

Tryptophan is an essential component of the human diet and plays a crucial role in many metabolic functions. Seed storage proteins are generally devoid of this amino acid. The

identification of a high tryptophan (2.5%)-containing foxtail millet 10-kDa prolamin protein in the present study appears to be promising as it offers a unique opportunity to increase tryptophan levels in cereal seeds. Identification of the unique repeating motif in this protein “NPAAF^WQQQQL” also opens new avenues to engineer SSPs for achieving high tryptophan levels during seed development. Likewise, another foxtail millet 10-kDa prolamin containing highest phenylalanine content (9.4%) offers opportunity to increase phenylalanine content in cereal seeds. Apart from these important amino acids discussed, the 2s albumin and the 7s globulin protein containing 11.92% valine and 9.16% threonine, respectively, with high overall essential amino acid content (Table 5) could be promising candidate genes for cereal nutritional quality improvement programmes. The coding region of the foxtail millet genes encoding these proteins could be inserted into suitable plant gene delivery vectors under the control of a strong seed-specific promoter for its high expressions in developing seed to increase the content of the desired amino acid in the seed. More research endeavors are required to fully understand the nutritional capability of foxtail millet.

Conclusion

Foxtail millet is known as a nutritious cereal crop for centuries. The present study highlights the presence of unique seed storage proteins containing high amounts of essential amino acids identified so far in any other cereals and legumes. The earlier studies on biofortification of major food grains with essential amino acids have used genes encoding non-seed storage proteins (other than seed storage proteins) from non-cereal sources. Hence, the seed storage protein with highest lysine, tryptophan and other essential amino acid content could be potential candidates for future biofortification studies in major cereals. Further, along with the strategy of over producing free amino acids, the genes encoding these proteins could be used as transgenes to generate effective sinks for production of nutritious grains rich in essential amino acids.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Agizzio A, Da P, Cunha M, Carvalho AO, Oliveira MA, Ribeiro SFF, Gomes V (2006) The antifungal properties of a 2S albumin homologous protein from passion fruit seeds involve plasma membrane permeabilization and ultrastructural alterations in yeast cells. *Plant Sci* 171:515–522
- Bennetzen JL, Schmutz J, Wang H, Percifield R (2012) Reference genome sequence of the model plant *Setaria*. *Nat Biotechnol* 30:555–561
- Brutnell TP, Wang L, Swartwood K, Goldschmidt A, Jackson D, Zhu XG, Kellogg E, Van Eck J (2010) *Setaria viridis*: a model for C4 photosynthesis. *Plant Cell* 22:2537–2544
- Cândido ES, Pinto MFS (2011) Plant storage proteins with antimicrobial activity: novel insights into plant defense mechanisms. *FASEB J* 25(10):3290–3305
- Casey R (1999) Distribution and some properties of seed globulins. In: Shewry PR, Casey R (eds) *Seed proteins*. Kluwer Academic Publishers, Dordrecht, pp 159–169
- Chakraborty S, Chakraborty N, Datta A (2000) Increased nutritive value of transgenic potato by expressing a non-allergenic seed albumin gene from *Amaranthus hypochondriacus*. *Proc Natl Acad Sci USA* 97(7):3724–3729
- Dunwell JM, Purvis A, Khuri S (2004) Cupins: The most functionally diverse protein superfamily? *Phytochemistry* 65:7–17
- Galili G, Tang G, Zhu X, Gakiere B (2001) Lysine catabolism: a stress and development super-regulated metabolic pathway. *Curr Opin Plant Biol* 4:261–266
- Gupta SM, Arora S, Mirza N, Pande A, Lata C, Puranik S, Kumar J, Kumar A (2017) Finger millet: a “certain” crop for an “uncertain” future and a solution to food insecurity and hidden hunger under stressful environments. *Front Plant Sci* 8:643
- Holding DR (2014) Recent advances in the study of prolamin storage protein organization and function. *Front Plant Sci* 5:276
- Izquierdo L, Godwin ID (2005) Molecular characterization of a novel methionine-rich δ -kafirin seed storage protein gene in *Sorghum* (*Sorghum bicolor* L.). *Cereal Chem* 82(6):706–710
- Jiang S, Ma A, Xie L, Ramachandran S (2016) Improving protein content and quality by over-expressing artificially synthetic fusion proteins with high lysine and threonine constituent in rice plants. *Sci Rep* 6:34427
- Kumar A, Pathak RK, Gupta SM, Gaur VS, Pandey D (2015) Systems biology for smart crops and agricultural innovation: filling the gaps between genotype and phenotype for complex traits linked with robust agricultural productivity and sustainability. *Omics J Integr Biol* 19(10):581–601
- Kumar A, Metwal M, Kaur S, Gupta AK, Puranik S, Singh S, Singh M, Gupta S, Babu BK, Sood S, Yadav R (2016a) Nutraceutical value of finger millet [*Eleusine coracana* (L.) Gaertn.], and their improvement using omics approaches. *Front Plant Sci* 7:934
- Kumar S, Stecher G, Tamura K (2016b) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874
- Lang Z, Zhao Q, Yu J, Zhu D, Ao G (2004) Cloning of potato *SBGLR* gene and its intron splicing in transgenic maize. *Plant Sci* 166:1227–1233
- Letunic I, Doerks T, Bork P (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res* 43(Database issue):D257–D260
- Mandal S, Mandal RK (2000) Seed storage proteins and approaches for improvement of their nutritional quality by genetic engineering. *Curr Sci* 79:5
- Matta NK, Singh A, Kumar Y (2009) Manipulating seed storage proteins for enhanced grain quality in cereals. *Afr J of Food Sci* 3(13):439–446
- Mitsukawa N, Konishi R, Uchiki M, Masumura T, Tanaka K (1999) Molecular cloning and characterization of a cysteine-rich 16.6-kDa prolamin in rice seeds. *Biosci Biotechnol Biochem* 63(11):1851–1858
- Molvig L, Tabe LM, Eggum BO, Moore AE, Craig S, Spencer D, Higgins TJV (1997) Enhanced methionine levels and increased

- nutritive value of seeds of transgenic lupins (*Lupinus angustifolius* L.) expressing a sunflower seed albumin gene. *Proc Natl Acad Sci USA* 94(16):8393–8398
- Moreno FJ, Clemente A (2008) 2S albumin storage proteins: what makes them food allergens? *Open Biochem J* 2:16–28
- Osborne TB (1897) Amount and properties of the proteins of the maize kernel. *J Am Chem Soc* 19:525–532
- Shewry PR, Halford NG (2002) Cereal seed storage proteins: structures, properties and role in grain utilization. *J Exp Bot* 53(370):947–958
- Shewry PR, Mifflin BJ, Bright SWJ (1981) Conventional and novel approach to the improvement of the nutritional quality of cereal and legume seeds. *Sci Prog* 67:575–600
- Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: structures and biosynthesis. *Plant Cell* 7:945–956
- Tabe LM, Wardley-Richardson T, Ceriotti A, Aryan A, McNabb W, Moore A, Higgins TJ (1995) A biotechnological approach to improving the nutritive value of alfalfa. *J Anim Sci* 73(9):2752–2759
- Takaiwa F, Oono K, Wing D, Kato A (1991) Sequence of three members and expression of a new major subfamily of glutelin genes from rice. *Plant Mol Biol* 17:875–885
- Timothy LB, Michael G (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14:48–54
- Timothy LB, Mikael B, Fabian AB, Martin F, Charles EG, Luca C, Jingyuan R, Wilfred WL, William SN (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37:W202–W208
- Wallace NH, Kriz AL (1991) Nucleotide sequence of a cDNA clone corresponding to the maize *Globulin-2* gene. *Plant Physiol* 95:973–975
- 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 USA* 105(38):14330–14335
- Youle RJ, Huang AHC (1981) Occurrence of low molecular weight and high cysteine containing albumin storage proteins in oilseeds of diverse species. *Am J Bot* 68:44
- Yue J, Li C, Zhao Q, Zhu D, Yu J (2014) Seed-specific expression of a lysine-rich protein gene, *GhLRP*, from cotton significantly increases the lysine content in maize seeds. *Int J Mol Sci* 15:5350–5365
- Zhang G, Liu X, Quan Z, Cheng S et al (2012) Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat Biotechnol* 30:549–554