Transcriptomic Analysis, Genic SSR Development, and Genetic Diversity of Proso Millet (Panicum miliaceum; Poaceae)

Authors: Hou, Siyu, Sun, Zhaoxia, Li, Yaoshen, Wang, Yijie, Ling, Hubin, et al.

Source: Applications in Plant Sciences, 5(7)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1600137

Applications in Plant Sciences

Applications in Plant Sciences 2017 5(7): 1600137

Genomic Resources Article

Transcriptomic analysis, genic SSR development, and genetic diversity of proso millet (*Panicum miliaceum*; Poaceae)[1]

Siyu Hou[2,3,6], Zhaoxia Sun[2,6], Yaoshen Li[2], Yijie Wang[2], Hubin Ling[2], Guofang Xing[2,3], Yuanhuai Han[2,4,5], and Hongying Li[2,5]

[2]College of Agriculture, Institute of Agricultural Bioengineering, Shanxi Agricultural University, Taigu, Shanxi 030801, People's Republic of China; [3]Shanxi Key Laboratory of Genetic Resources and Genetic Improvement of Minor Crops, Taigu, Shanxi 030801, People's Republic of China; and [4]Key Laboratory of Crop Gene Resources and Germplasm Enhancement on the Loess Plateau, Ministry of Agriculture, Institute of Crop Genetic Resources, Shanxi Academy of Agricultural Sciences, Taiyuan, Shanxi 030031, People's Republic of China

- *Premise of the study:* Proso millet (*Panicum miliaceum*; Poaceae) is a minor crop with good nutritional qualities and strong tolerance to drought stress and soil infertility. However, studies on genetic diversity have been limited due to a lack of efficient genetic markers.
- *Methods:* Illumina sequencing technology was used to generate short read sequences of proso millet, and de novo transcriptome assemblies were used to develop a de novo assembly of proso millet. Genic simple sequence repeat (SSR) markers were identified and used to detect polymorphism among 56 accessions. Population structure and genetic similarity coefficient were estimated.
- *Results:* In total, 25,341 unique gene sequences and 4724 SSR loci were obtained from the transcriptome, of which 229 pairs of SSR primers were validated, which resulted in 14 polymorphic genic SSR primers exhibiting 43 total alleles. According to the ratio of polymorphic markers (6.1%, 14/229), there are potentially 288 polymorphic genic SSR markers available for genetic assay development in the future. Bayesian population analyses showed that the 56 accessions comprised two distinct groups.
- *Discussion:* A genetic structure and cluster assay indicated that the accessions from the Loess Plateau of China shared a high genetic similarity coefficient with those from other regions and that there was no correlation between genetic diversity and geographic origin. The transcriptome sequencing data and millet-specific SSR markers developed in this study establish an excellent resource for gene discovery and may improve the development of breeding programs in proso millet in the future.

Key words: de novo transcriptome; genetic diversity; genic SSR; germplasm resources; *Panicum miliaceum*; Poaceae; proso millet.

Proso millet, also named broomcorn or common millet (*Panicum miliaceum* L.; Poaceae), was domesticated in Neolithic China ca. 10,000 BP (Lu et al., 2009) and grown across Eurasia as a minor crop. The cultivation area of proso millet is about 0.32 million ha in China, with a total grain production of approximately 0.3 million tons per year (Diao, 2017). Although the grain yield of proso millet is lower than that of major crops such as maize (*Zea mays* L.), wheat (*Triticum aestivum* L.), and rice (*Oryza sativa* L.), it was once cultivated as a large crop in northern China because of its much greater resistance to drought and unfertile soil than these major crops (Diao, 2017). It was introduced to North America in the 18th century and used mainly for animal fodder and birdseed (Bagdi et al., 2011). Across Eurasia, however, proso millet was important in the human diet before the introduction of other crops such as wheat, barley, and potatoes. More attention has been paid to this crop in China, India, Nepal, Pakistan, and Southeast Asian countries, where it is considered a functional food due to its high protein content (11.3–17% of grain dry matter), abundant antioxidant components, and dietary fiber (Kalinova and Moudry, 2006). A few reports have shown that the phytochemical components of proso millet can prevent certain cancers, heart disease, and relieve liver disease and diabetes (Srivastava et al., 2001; Choi et al., 2005; Shimanuki et al., 2006). The crop has high-water use efficiency, a short growing period (60–90 d), and a high adaptability and tolerance to semiarid climates (Agdag et al., 2001; Graybosch and Baltensperger, 2009). Hence, it is ideally suited to cultivation in areas with hot, dry, and short summer seasons (Saseendran et al., 2009).

[1] Manuscript received 13 November 2016; revision accepted 2 May 2017.

The authors thank Professor Donald Grierson (University of Nottingham, Nottingham, United Kingdom) and Dongmei Xu (Chief Scientist, Altria Group Inc., Richmond, Virginia, USA) for discussions and help with the manuscript. This work was supported by the Science and Technology Project of Shanxi Province (20150311007-1), the National Natural Science Foundation of China (no. 31301385, 31471556, 31200914), the Basic Facility Platform Project of Shanxi Provincial Department of Science and Technology (2012091004-0103), and the College Key Discipline Construction Special Foundation of Shanxi Province.

[5] Authors for correspondence: hongylswgctd@163.com (H.Y.L.), swgctd@163.com (Y.H.H.)

[6] These authors contributed equally to this work.

doi:10.3732/apps.1600137

Table 1. Characterization of 14 polymorphic genic SSR markers in proso millet.

Marker ID	Repeat motif	A_e	PIC value	Gene annotation
SXAU12	$(AAC)_5$	2	0.5663	Misc RNA
SXAU17	$(GCA)_6$	2	0.6497	Hypothetical protein, mRNA
SXAU32	$(GGC)_6$	4	0.8084	Hypothetical protein, mRNA
SXAU33	$(GT)_9$	3	0.5865	Unknown
SXAU63	$(CT)_9$	3	0.7169	Hypothetical protein, mRNA
SXAU92	$(GCA)_6$	2	0.6847	Eukaryotic translation initiation factor 4 gamma-like
SXAU95	$(GCG)_6$	3	0.6999	Unknown
SXAU100	$(AG)_{10}$	1	0.1014	Zinc finger protein ZAT3-like
SXAU118	$(TCT)_5X_{86}(TTC)_6$	3	0.7496	Uncharacterized protein
SXAU106	$(AC)_{10}$	1	0.5762	Uncharacterized protein
SXAU119	$(CTG)_5X_{86}(GCT)_5$	2	0.6875	Uncharacterized protein
SXAU132	$(CCG)_5X_{76}(CGC)_5$	2	0.6248	Uncharacterized protein
SXAU138	$(AT)_8X_{60}(AAGG)_5$	4	0.7667	WRKY transcription factor 6
SXAU227	$(GCGAT)_5$	3	0.7402	B3 domain-containing protein

Note: A_e = effective number of alleles; PIC = polymorphism information content.

Although proso millet is important for both its nutritional and economic value as a source of food diversity, its genetics and breeding research have been largely neglected. Previously, 25 polymorphic microsatellite markers were developed through construction of a simple sequence repeat (SSR)–enriched library from genomic DNA of proso millet to assess 50 accessions of proso millet (Cho et al., 2010). Another study documented that 4.67% (46 out of 983) of genomic SSR markers derived from rice, wheat, oat, and barley were successfully transferred into proso millet (Hu et al., 2009). Although these available SSR markers from the genomic library, or other species, could be used for genetic analyses and marker-associated breeding in proso millet, most of them are constituted of noncoding DNA, involving a relatively small number and type of SSR repeats with low transferability across species. As next-generation sequencing costs decrease, many new SSR loci generated from these published transcriptomes in green plants could be derived from microsatellites in translated regions of the genome (Hodel et al., 2016). Such genic SSRs offer advantages over genomic SSRs because they detect variation in the expressed portion of the genome, so that gene tagging should give "perfect" marker-trait associations and, once developed, these markers, unlike genomic SSRs, may be used across a number of related species (Gupta et al., 2003; Kalia et al., 2011). Therefore, we used an economical and rapid strategy to develop an abundant and efficient number and type of SSR markers located in coding regions as a necessary and useful supplement to the previous research on genetic diversity in proso millet.

According to cytological investigations, proso millet is an allotetraploid cereal ($2n = 4x = 36$) (Hamoud et al., 1994). Its genome progenitors are not clear, although phylogenetic data suggest *P. capillare* L. or a close relative as the maternal ancestor, with the other genome being shared with *P. repens* L. (Hunt et al., 2014). It has been reported that the variation among these proso millet accessions is low when studied by isozyme and microsatellite molecular markers, which likely reflects the double-bottleneck of polyploidization and domestication (Warwick, 1987; Hu et al., 2009; Hunt et al., 2011).

Landraces of the crop and related species are inadequately used as genetic resources, although they may possess the potential for significant improvements in seed and forage yield as well as seed quality in the future. Here, we report on the transcriptome of proso millet. Using short reads from the Illumina sequencing platform, a nonredundant set of transcripts was generated for this species. Analyses of guanine-cytosine (GC) content, sequence similarity, functional categorization, and identification of SSRs provide a useful resource for future studies in proso millet.

MATERIALS AND METHODS

Plant material and sample collection—Fifty-six accessions of proso millet were grown in the field at the agricultural experiment station of Shanxi Agricultural University (northern China, 37°25′N, 112°29′E) during summer 2015. These accessions were derived from 14 geographic regions of China (Appendix 1) and donated from the Chinese Crop Germplasm Resources Information System (CGRIS) and the Institute of Crop Genetic Resources (Shanxi Academy of Agricultural Sciences, Shanxi, China). The Neimenggu-Y1 cultivar, which is widely planted in Shanxi Province, was chosen for Illumina sequencing and de novo transcriptome assembly. The accessions in this study were used to evaluate and identify the suitability of the SSRs for further analysis of genetic distance. To ensure as many genes as possible were included in the transcriptome, six tissues of Neimenggu-Y1 (corresponding to no. 46 in Appendix 1) were harvested and pooled together for RNA-Seq. The six tissues included roots, young leaves, flag leaves, the fifth leaves numbered from top to bottom, young spikes, and mature spikes. There are three independent biological replicates for each tissue. These samples were snap-frozen in liquid nitrogen and kept at −80°C until further analysis.

RNA extraction, RNA-Seq, and transcriptome analyses—Total RNA from each tissue of Neimenggu-Y1 (no. 46 in Appendix 1) was isolated for RNA-Seq

Table 2. Summary of RNA-Seq data in proso millet.

Sample name	Raw data (bp)	Clean data (bp)	Useful data (%)	GC (%)	Q20 (%)
Roots	2,382,926,800	2,072,638,120	86.98	52.63	95.01
Seedling leaves	2,758,786,400	2,410,330,025	87.37	53.34	95.17
Fifth leaf (from bottom to top)	2,862,374,400	2,497,088,922	87.24	52.51	95.17
Flag leaves	2,847,722,600	2,486,485,369	87.31	52.78	95.16
Young inflorescences	2,344,015,800	2,056,826,835	87.75	52.37	95.35
Mature spikes	2,224,403,600	1,941,752,916	87.29	53.38	95.20

analysis according to the RNA quick extraction kit method (Tiandz, Beijing, China). RNA purity and concentration were checked using a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA). RNA integrity was assessed using an Agilent Analyzer IIx 2000 (28S : 18S ≥ 1.5, RNA integrity number [RIN] ≥ 7.0) (Agilent Technologies, Santa Clara, California, USA). A total of 10 µg of RNA per tissue sample was pooled and used as input material for generating RNA-Seq libraries with a 175-bp cDNA insert fragment, according to the Illumina mRNA sequencing sample preparation guide. According to the manufacturer's instructions, these libraries were sequenced on the Illumina HiSeq 2000 platform (Illumina, San Diego, California, USA) using a paired-end (2 × 100 bp) sequencing method (Personalbio Corporation, Shanghai, China).

***Sequence assembly and functional gene classification*—**The raw reads from the images obtained after sequencing were transformed into the FASTQ format. Clean reads were obtained by removing reads containing adapter, low-quality sequences with raw read quality <Q20 and final read sequence length <50 bp with a Perl script (Patel and Jain, 2012). Meanwhile, the Q20, Q30, GC content, and level of sequence duplication of the clean data were calculated with the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Subsequently, Trinity (Grabherr et al., 2011) was used to assemble the sequencing information for the clean reads from all libraries to obtain nonredundant unigenes with default parameters (K-mer: 25, minimum contig length: 60). The minimum length of the assembled unigenes selected for further study was 200 nucleotides. The raw reads were submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (accession no. SRP083017, release date August 2017).

All assembled unigenes were compared to the NCBI nonredundant (Nr) protein databases (with *Zea mays* as reference protein) and Swiss-Prot by using BLASTX with a standard cut-off *E*-value of $1e^{-5}$ to search for homologs (Conesa et al., 2005). Based on the results of the Nr database annotation, BLAST2GO (Conesa et al., 2005) was used to obtain gene ontology (GO) annotations of assembled unigenes to better understand the distribution of gene functions at a macro level. The unigene sequences were also aligned to the Clusters of Orthologous groups of proteins (COG) database (http://www.ncbi.nlm.nih.gov/COG/) to predict and classify possible functions (Tatusov et al., 2000). Meanwhile, the unigenes were assigned to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations to analyze inner-cell metabolic pathways and the related gene functions using BLASTX (Kanehisa et al., 2016). Reads per kilobase of transcript per million mapped reads (RPKM) values were used to test expression of all unigenes and compare gene expression differences between different samples with RSEM software (Li and Dewey, 2011). Differential gene expression analysis was conducted with the RSEM-EBSeq pipeline with specifications and control of false discovery rate (FDR) at level 0.05, as described (Li and Dewey, 2011).

***DNA extraction*—**Young leaves from 56 accessions were collected and frozen in liquid nitrogen prior to genomic DNA extraction using cetyltrimethylammonium bromide (CTAB) methods, to evaluate and identify the suitability of the SSRs for genetic distance analysis (Porebski et al., 1997). DNA concentrations and purity were detected by the NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific), and their integrity was checked by gel electrophoresis analysis with the Horizontal Electrophoresis Systems (Bio-Rad Laboratories, Hercules, California, USA). The purified DNAs were diluted to 100 ng/µL and stored at −20°C.

***Genic SSR loci identification and primer analysis*—**The unigene data of the proso millet transcriptome were modified as a standard output format (FASTA) and deposited to NCBI (Bioproject no. PRJNA339701). Microsatellites in the proso millet transcriptome were identified using the Perl script MIcroSAtellite identification tool (MISA; http://pgrc.ipk-gatersleben.de/misa/) according to the method of analysis for microsatellite distribution in monocots (*Brachypodium*, *Sorghum*, and *Oryza*) and dicots (*Arabidopsis*, *Medicago*, and *Populus*) (Sonah et al., 2011). To identify the presence of SSRs, only two to six nucleotide motifs were considered, excluding redundant single nucleotides such as $(A)_n$ or $(T)_n$, and the minimum repeat unit was defined as eight for dinucleotides and five for tri-, tetra-, penta-, and hexanucleotides. Compound SSRs were defined as ≥2 SSR loci interrupted by ≤100 bases. All of the SSR loci information and flanking sequences were generated by MISA. Then, according to the flanking sequence of the microsatellite loci, two perl scripts were used to design SSR primers by program invocation and interchange between MISA with Primer3 (Untergasser et al., 2012) according to the following parameters: 100–280 bp amplicon size, 20 bp optimal primer length, and 50% optimal GC content; the annealing temperature was set at 50–65°C. Primers were synthesized by Sangon Co., Ltd. (Shanghai, China) (Appendix S1). Fourteen cultivars (shown in Appendix 1) derived from different geographic regions were selected to test the polymorphism of 229 random SSR loci. Each PCR product was studied using polyacrylamide gel electrophoresis (PAGE) at 60 W constant gain, and polymorphisms were detected by silver staining. To evaluate genetic diversity among all 56 accessions, 14 pairs of polymorphic SSR primers were chosen (Table 1, Appendix S1). Each PCR product was run on a 2% agarose gel at 120 V for a quality check. Subsequently, the PCR products were run on a Fragment Analyzer Automated CE system (Advanced Analytical Technologies, Ankeny, Iowa, USA), and polymorphisms were detected by fluorescence signal.

Optimal PCR conditions and programs were as follows. Each 25-µL reaction mixture contained 2.5 µL of 10× PCR buffer (plus Mg^{2+}), 2 µL of dNTP (2.5 mM), 1 µL of each reverse and forward primer (10 pM), 0.5 µL of rTaq polymerase (2.5 U/µL, TaKaRa Bio, Otsu, Shiga, Japan), and 1 µL of genomic DNA template (100 ng/µL). DNA amplification was accomplished in a PCR system (PTC-200; MJ Research, Waltham, Massachusetts, USA) programmed at 94°C for 5 min for initial denaturation, then 35 cycles at 94°C (30 s)/55°C (30 s)/72°C (30 s), followed by an extension step for 10 min at 72°C.

***SSR data analysis*—**Fourteen highly polymorphic SSR primer combinations were selected as core SSR markers and used to analyze 56 proso millet varieties (Table 1). The effective number of alleles (A_e) and polymorphism information content (PIC) were calculated using POPGENE version 1.32 (Yeh et al., 1999), and population genetic structure was calculated using STRUCTURE 2.3.4 (Hubisz et al., 2009). The genetic relationships between the genotypes were calculated using the unweighted pair group method with arithmetic mean (UPGMA) cluster analysis. This analysis was carried out on the similarity matrix obtained from the proportion of shared fragments (Nei and Li, 1979) using the program NTSYS 2.1 (Rohlf, 2008).

RESULTS

***Functional annotation and characteristics of de novo–assembled transcriptomes of proso millet*—**Between 2.22 and 2.86 Gb of raw sequence data were obtained from each tissue sampled from Neimenggu-Y1. Filtering to a threshold of Q20 sequence quality produced 1.94–2.49 Gb of clean sequence data, accounting for 86.69–87.75% of the raw data, which was used to assemble contigs and transcripts for further analysis (Table 2). A total of 359,126 contigs and 178,020 transcripts were obtained from the processed reads using Trinity software (Grabherr et al., 2011). Using BLASTX searches, 42,364,858 bp of assembled sequences were obtained from all of the RNA samples, covering 25,341 single genes with an average nucleotide length of 1672 bp (Table 3). The average GC content of these transcripts and unigenes in proso millet were 48.64% and 50.80%, respectively. The length distributions of transcripts and unigenes are shown in Fig. 1. BLASTX searches with these unigenes showed 24,621 (97.16%) and 25,341 (100%)

Table 3. Statistics of transcriptome assembly in proso millet.

Assembly quality parameters	No.
Contigs generated	359,126
Maximum contig length (bp)	17,623
Average contig length (bp)	335.42
Contig N50 value	562
Transcript generated	178,020
Maximum transcript length (bp)	15,773
Average transcript length (bp)	908
Transcript N50 value	1546
Unigene generated	25,341
Maximum unigene length (bp)	15,773
Average unigene length (bp)	1672
Unigene N50 value	2199

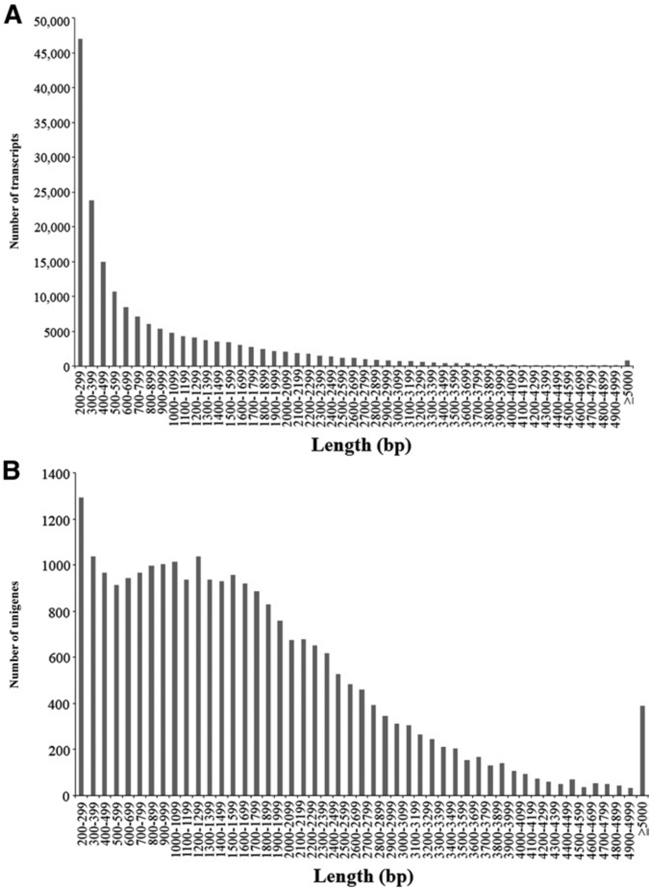

Fig. 1. Sequence length distribution frequency in proso millet, using transcriptome data collected from six different tissue types, for (A) assembled transcripts and (B) assembled unigenes.

nucleotide sequences were similar to those of *Setaria italica* (L.) P. Beauv. and *Zea mays*, respectively. From the combined total, only 7450 (29.39%) were highly homologous to proteins in the Swiss-Prot database. Of these, 5170 (20.4%) unigenes could be mapped to 146 KEGG pathways (Fig. 2A; Appendix S2), and the other unigenes showed only poor matches because of their short sequences. A significant similarity to the sequences available in the GO database was observed in 19,338 (76.31%) genes. Based on comparative transcriptome analysis, 2936 (11.59%) tissue-specific genes were identified from five tissue samples. These included 1058 genes from roots, 224 genes from the fifth leaves, 53 genes from mature leaves, and 133 and 128 genes from young and mature spikes, respectively (Fig. 2B). These genes were selected for GO annotation and classified into three subcategories, i.e., biological process (2779 genes in roots, 249 genes in mature spikes, 324 genes in the fifth leaves, 59 genes in flag leaves, and 204 genes in young spikes), molecular functions (1541 genes in roots, 144 genes in mature spikes, 251 genes in the fifth leaves, 63 genes in flag leaves, and 159 genes in young spikes), and cellular component (1425 genes in roots, 137 genes in mature spikes, 280 genes in the fifth leaves, 64 genes in flag leaves, and 174 genes in young spikes) (Fig. 2C).

Identification and functional annotation of genic SSR loci—A total of 4724 genic SSR loci were identified from 42,364,858-bp mRNA sequences. These loci were mainly distributed among 4055 gene sequences with one SSR locus, with 567 genes containing more than one SSR locus. The SSR repeat motifs consisted of 24.96% dinucleotide, 71.85% trinucleotide, 2.43% tetranucleotide, 0.66% pentanucleotide, and 0.11% hexanucleotide repeat units (Fig. 3A). The dinucleotide AG/CT

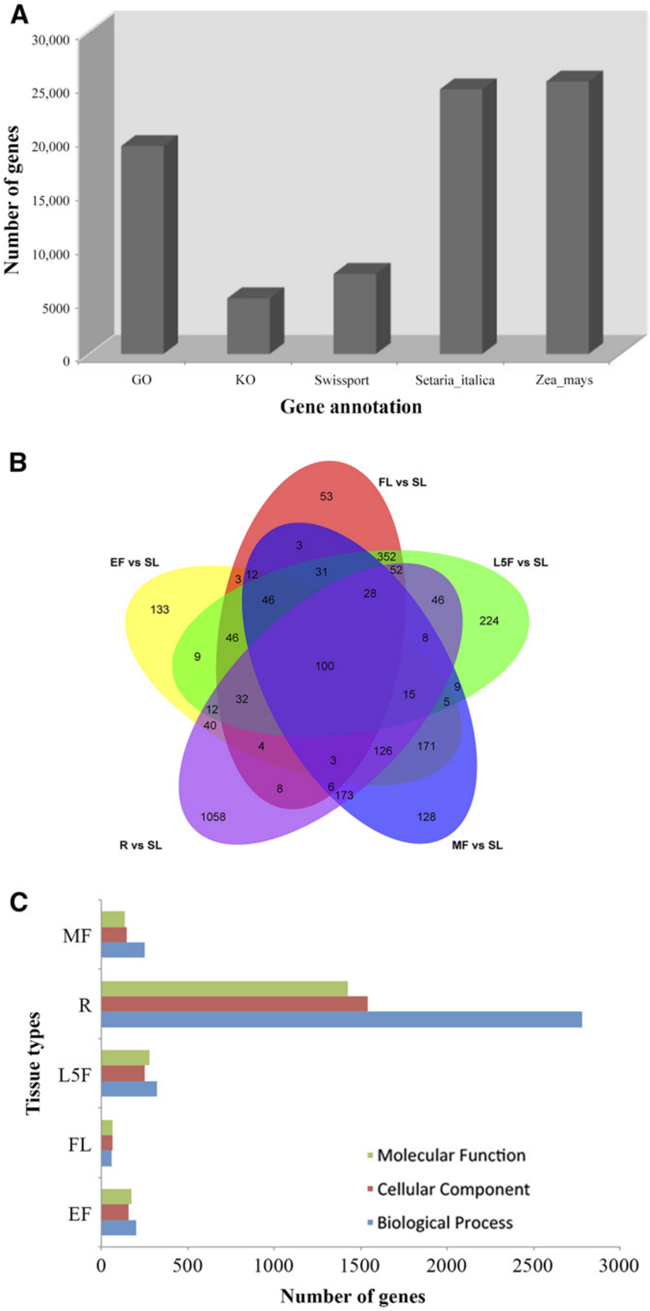

Fig. 2. Annotation and analyses of different genes expressed in proso millet, using transcriptome data collected from six different tissue types. (A) Genes annotated by GO functional category, KEGG pathway, Swiss-Prot protein data, *Setaria italica* genome data, and *Zea mays* genome data. (B) Venn diagram of differential expression gene analysis by comparing root (R), flag leaves (FL), the fifth leaves from top to bottom (L5F), young spike (EF), and mature spike (MF) with seedling leaves (SL) tissue transcriptome data. (C) GO functional category of differential gene expression in the above five tissues in this transcriptome data analysis.

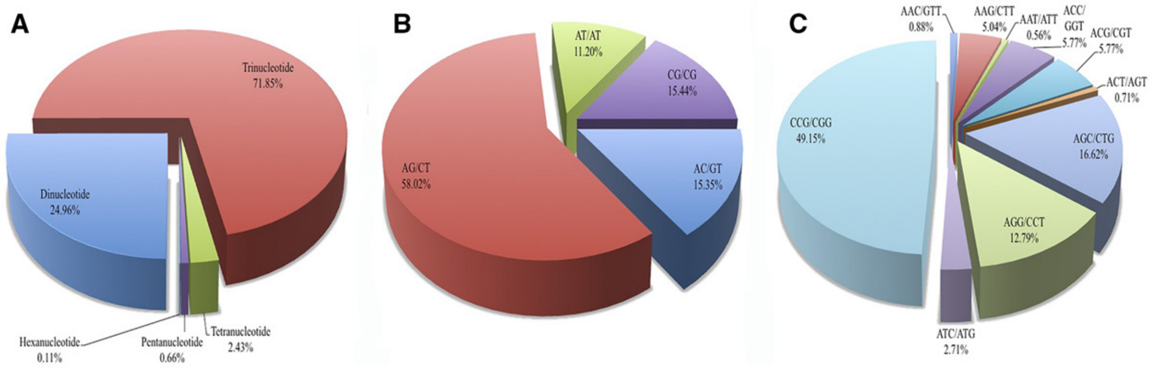

Fig. 3. Identification and characteristics of genic SSR motifs from RNA-Seq data. (A) Frequency of di-, tri-, penta-, tetra-, and hexanuclotide SSR motif repeats. (B) Frequency of different dinucleotide SSR motifs. (C) Frequency of different trinucleotide SSR motifs.

motif repeats were the most abundant type, accounting for 58.02% of all dinucleotide repeats (Fig. 3B). Of the trinucleotide SSR repeat motifs, the CCG/CGG repeats were the most abundant, accounting for 49.15% of all of trinucleotide repeat motifs, and AGC/CTG and AGG/CCT repeats accounted for 16.62% and 12.79%, respectively (Fig. 3C). The distribution frequency of dinucleotide repeats ranged from six to 12, and the trinucleotide repeat frequency ranged from five to 12. The frequency of other types of nucleotide repeats was generally five or six (Fig. 4). The most abundant SSR repeat motif was the $(CCG/GGC)_5$ type; there were 1225 of these repeat motifs, accounting for 25.93% of all SSR repeats. These were classified into 24 groups by functional annotation by comparing protein sequences encoded by the genes containing SSR motifs using the COG database. Of the COG categories, the signal transduction mechanisms (T) group was the largest, accounting for 23.99% (973 genes) of all genes, followed by the unknown function group (S), accounting for 14.57% (591 genes) (Fig. 5).

Genic SSR primer validation and polymorphism information—Using online primer design tools, 229 genic SSR markers with one SSR motif were obtained by random selection from 4055 gene sequences. Based on PCR validation, 189 and 14 genic SSR markers showed monomorphic and polymorphic amplification products, respectively, whereas the remaining 26 genic SSR markers failed to generate PCR products at various annealing temperatures (Table 2, Appendix S1). Fourteen genic SSR loci with repeat motifs contained four dinucleotide, five trinucleotide, one pentanucleotide, and four compound formations, which are made of two combinations of $(\text{trinucleotide repeat})_n$-$(\text{trinucleotide repeat})_n$ and $(\text{dinucleotide repeat})_n$-$(\text{tetranucleotide repeat})_n$. Their PCR products showed 43 alleles in total and PIC ranged from 0.5663 to 0.8084. Of these genic SSR markers, the SXAU32 marker containing the $(GGC)_6$ repeat motif displayed a higher number of polymorphic alleles (4) and a higher PIC value (0.8084) than other genic SSR markers (Appendix S3); the markers SXAU92, SXAU138, and SXAU227 belonged to the annotated function genes encoding *WRKY* transcript factor 6 (*WRKY6*), B3 domain-containing protein (*HIGH-LEVEL EXPRESSION OF SUGAR-INDUCIBLE GENE2-LIKE1* [*HSL1*]), and eukaryotic translation initiation factor 4 gamma-like (*EIF4G-like*), respectively. The rest of the markers belonged to hypothetical proteins or unannotated genes in the Nr database. We used the 14 genic SSR primer pairs as core primers to analyze genetic diversity among the proso millet germplasm collections from China.

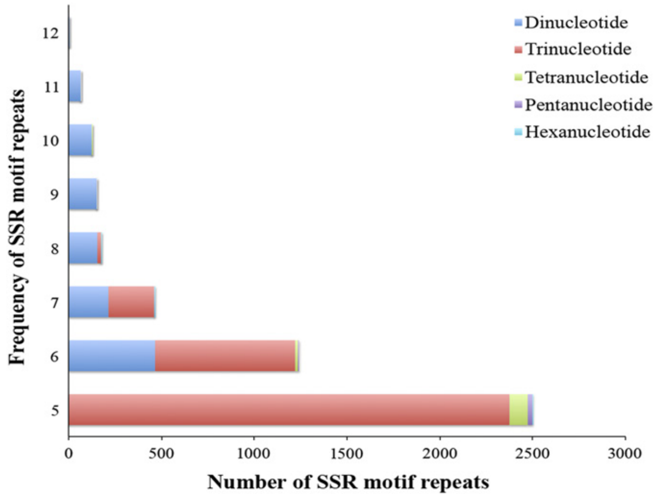

Fig. 4. The distribution and frequency of SSR motif repeats. Blue bars = dinucleotide repeat motifs; red bars = trinucleotide repeat motifs; green bars = tetranucleotide repeat motifs; purple bars = pentanucleotide repeat motifs; light blue bars = hexanucleotide repeat motifs.

Genetic population grouping of proso millet cultivars—Based on the Bayesian (model-based algorithm) method, we analyzed genetic population clusters with STRUCTURE 2.3.4 (Hubisz et al., 2009) to assess genetic diversity of the accessions. All estimated mean log probability data (Ln P(D)) with a *K* value from two to 15 were used to find the more likely number of clusters by following the formulas of Evanno et al. (2005). When $K = 2$ and $\Delta K = 149.34$, we found one clear peak value for the optimum choice of this researched population of proso millet (Fig. 6). The 56 accessions from 14 geographic regions were classified into two subgroups (Fig. 7). The first group contained 19 accessions from Heilongjiang (five accessions), Shaanxi (two accessions), Shanxi (two accessions), Jilin (two accessions), Ningxia (two accessions), Inner Mongolia (two accessions), Gansu (one accession), Hebei (one accession), Anhui (one accession), and Shandong (one accession). The second group consisted of 37 accessions from the Loess Plateau (five each from Ningxia and Shaanxi, four from Gansu, and three from Shanxi), the Mongolian Plateau (three from Inner Mongolia), the Northeast Plateau (three from Liaoning, two from Heilongjiang, and two from Jilin), the

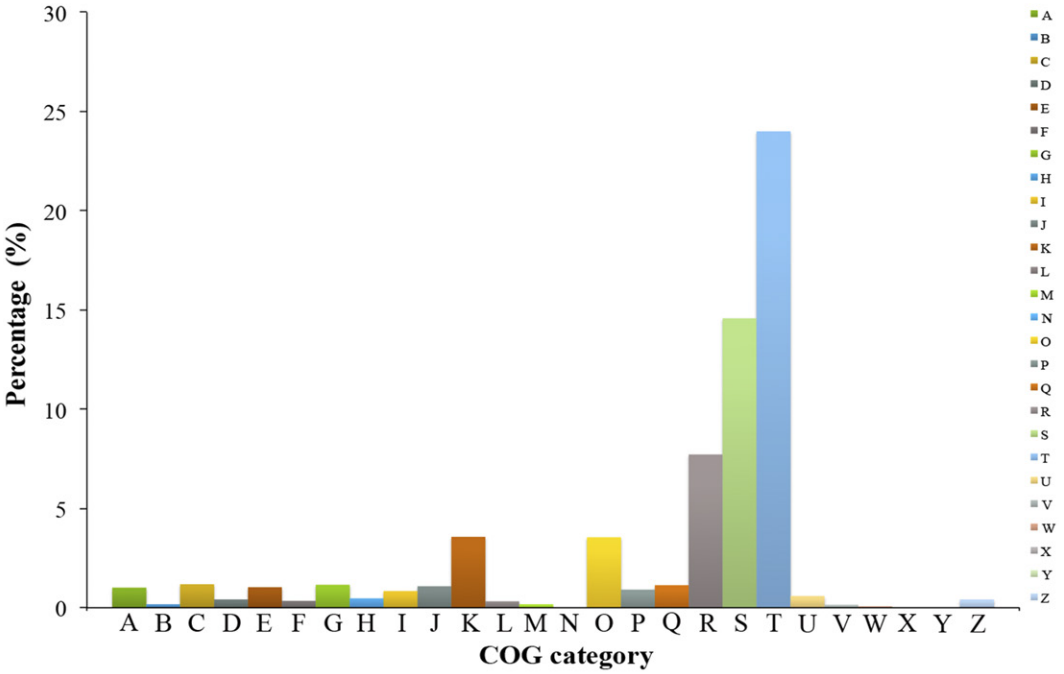

Fig. 5. Genes containing SSR motifs annotated by classification into clusters of orthologous groups (COGs). A = RNA processing and modification; B = chromatin structure and dynamics; C = energy production and conversion; D = cell cycle control, cell division, chromosome partitioning; E = amino acid transport and metabolism; F = nucleotide transport and metabolism; G = carbohydrate transport and metabolism; H = coenzyme transport and metabolism; I = lipid transport and metabolism; J = translation, ribosomal structure, and biogenesis; K = transcription; L = replication, recombination, and repair; M = cell wall/membrane/envelope biogenesis; N = cell motility; O = posttranslational modification, protein turnover, and chaperones; P = inorganic ion transport and metabolism; Q = secondary metabolites biosynthesis, transport, and catabolism; R = general function prediction only; S = function unknown; T = signal transduction mechanisms; U = intracellular trafficking, secretion, and vesicular transport; V = defense mechanisms; W = extracellular structures; X = mobilome, transposons; Y = nuclear structure; Z = cytoskeleton.

North China Plain (three from Hebei), and other geographic origin (four from Qinghai, one from Shandong, one from Guangdong, and one from Xinjiang Uyghur Autonomous Region). Using UPGMA cluster analysis, the 56 accessions were clustered into two groups with a genetic similarity coefficient of 0.771 (Fig. 8). Group I comprised 31 accessions from the Loess Plateau (five from Shaanxi, three from Shanxi, four from Ningxia, and three from Gansu), the Mongolian Plateau (two from Inner Mongolia), the Northeast Plateau (three from Heilongjiang, two each from Liaoning and Jilin), the North China Plain (three from Hebei), and other geographic origin (two from Qinghai, one from Shandong, and one from Xinjiang). The remaining 25 accessions belonged to group II. Except for six accessions (two from Qinghai and one each from Ningxia, Liaoning, Gansu, and Guangdong), the results of the cluster analyses were completely consistent with the population structure analyses.

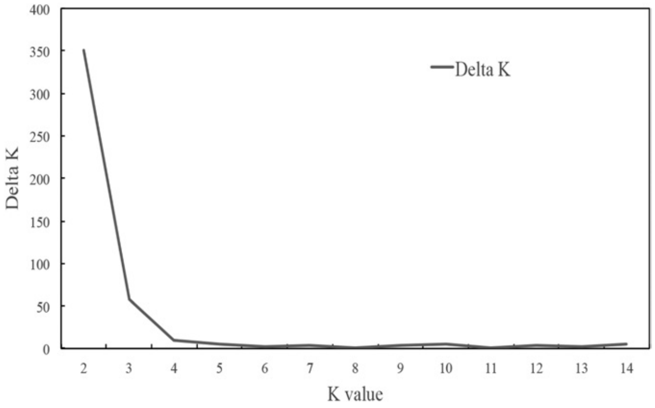

Fig. 6. The number of clusters (*K*) revealed in proso millet population genetic structure analysis using STRUCTURE 2.3.4 (Hubisz et al., 2009).

DISCUSSION

In China, proso millet is widely cultivated over a large area and is a popular functional food compared to other minor crops. The largest collection of proso millet germplasm is held by the N. I. Vavilov All-Russian Research Institute of Plant Industry in St. Petersburg, with about 8778 accessions. A further 6517 proso millet germplasm resources are maintained at the Chinese Crop Germplasm Resources Information System in China (Dwivedi et al., 2012). Preliminary diversity clustering was based on agronomic traits such as flowering time, seed color, plant height, and inflorescence length (Dwivedi et al., 2012). Due to the lack of molecular markers and genome sequences in the public domain, some molecular markers used for proso millet have been derived from the sequence data of related species including switchgrass, rice, wheat, barley, and oats, or from a SSR-enriched genomic library (Hu et al., 2009; Cho et al., 2010). However, further research is urgently needed to provide tools for studying genetic diversity and identification of core germplasm resources.

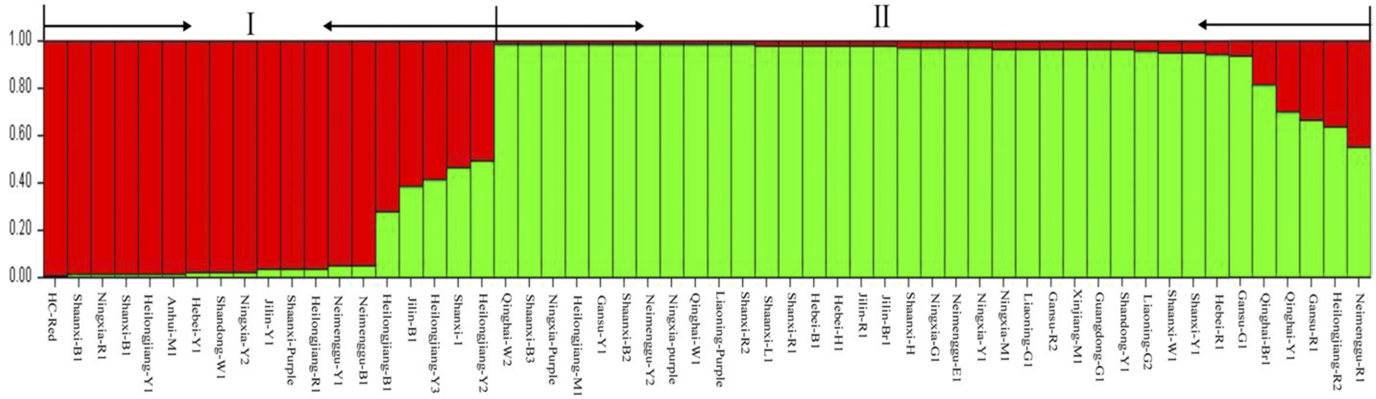

Fig. 7. Barplots of Q assignment for each of 56 proso millet accessions at *K* = 2. The *x*-axis shows the 56 accessions; the *y*-axis shows Q assignments for each sample.

The whole genome of foxtail millet (*Setaria italica*), which is a closely related species to proso millet, has been sequenced and published (Bennetzen et al., 2012; Zhang et al., 2012). Further efforts in genome, expressed sequence tag (EST), or RNA sequencing will help the development of molecular markers in these minor species. Recently, de novo transcriptome assembly from developing spikes in finger millet (*Eleusine coracana* (L.) Gaertn.) has produced 52,251 and 35,120 transcripts from two genotypes with different levels of calcium in seeds, from which 24,748 and 21,276 SSRs were identified and genic SSR markers developed for analysis of germplasm resources using the Illumina HiSeq 2000 platform (Kumar et al., 2015). This present study reported the de novo–assembled transcriptome analysis of proso millet without a reference genome, revealing 25,341 unique gene sequences and 178,020 transcripts. These sequences will be valuable for investigating gene structure and function and for identifying single nucleotide polymorphisms (SNPs), insertion/deletion polymorphisms (indels), and SSR loci to develop molecular markers. Compared to SNPs and indels, SSRs have the advantage of simple detection by PCR and direct gel electrophoresis, without the need for additional sample sequencing or resequencing. Therefore, in this study, we have developed a resource to detect genetic diversity of proso millet accessions in further analysis.

Previous analysis of genetic variation in proso millet, using 46 genomic microsatellites (SSRs) transferred from other cereal species and 25 SSRs developed from a genomic SSR-enriched library, showed highly polymorphic alleles with an average of 4.91 and 4.4 per locus, respectively (Cho et al., 2010). Our results identified 43 alleles from 14 polymorphic genic markers (6.1%, 14/229) with an average of 3.3 per locus, which is lower than the above results, probably due to differences in molecular marker types and species sequence resources. Research comparing genomic SSR and EST-SSR markers for genetic diversity analyses in wheat, barley, and cucumber indicated that genomic SSRs (gSSRs) were enriched in the plant genome, with a high level of polymorphism and codominant inheritance (Gadaleta et al., 2011). Nonetheless, the availability of data for EST-SSR markers, which belong to the transcribed regions of DNA, are expected to be more conserved and have a higher transferability rate across species than gSSR markers (Hu et al., 2011; Gadaleta et al., 2011; Zhang et al., 2014). We hope that these genic SSRs developed in proso millet can also be applied to assess genetic diversity of closely related species, such as foxtail millet (*Setaria italica*), pearl millet (*Pennisetum glaucum* (L.) R. Br.), and finger millet (*Eleusine coracana*).

Among cereal species, trinucleotide repeats are the most abundant (54–78%), followed by either dinucleotide or tetranucleotide repeats (Varshney et al., 2002). Our results in proso millet were consistent with previous results, with trinucleotide repeats accounting for 71.85%, dinucleotide repeats for 24.96%, and tetranucleotide repeats for 2.43%. Among the 4724 SSR loci derived from 4055 unique transcriptomic sequences, most (2059 genes) belonged to the unknown function group (S) in the COG classification, suggesting that (i) there is lower sequence variation in known genes in proso millet, or (ii) there has been lower evolutionary pressure because of shorter growth cycles and self-fertilization. Further investigation is needed to explore the correlation between genetic variation among these accessions and gene function related to physiological characteristics in proso millet. The SSR markers developed here from genes of known function displayed genetic variation between accessions and could be used as markers related to phosphorus utilization efficiency (*WRKY6*) and plant growth and development (*HSL1* and *EIF4G-like*) (Chen et al., 2009).

A previous report suggested that there is a positive correlation between the geographic origin and genetic distance among 118 accessions using genome SSR data, and a lower genetic similarity coefficient for the accessions in the Loess Plateau ecotype compared to those from other ecotypes (Hu et al., 2009). In contrast, our results indicated that these Chinese accessions showed no obvious correlation between genetic distance and geographic origin according to population structure and UPGMA cluster analysis. Based on the population structure analysis, 17 of 24 accessions from the Loess Plateau were classified into group I, accounting for 70.83% of the genetic variation, and the remaining accessions belonged to group II. These results suggest that the accessions from the Loess Plateau have a higher genetic similarity coefficient when compared to other ecotypes. Likewise, UPGMA cluster analysis resulted in a genetic similarity coefficient among 56 accessions that supported the same conclusion. The difference between our results and the previous reports may be due to differences in the accessions and types of marker used, or the properties of markers selected. The new set of genic SSR markers developed here should provide valuable genetic and genomic tools for further genetic research in proso millet,

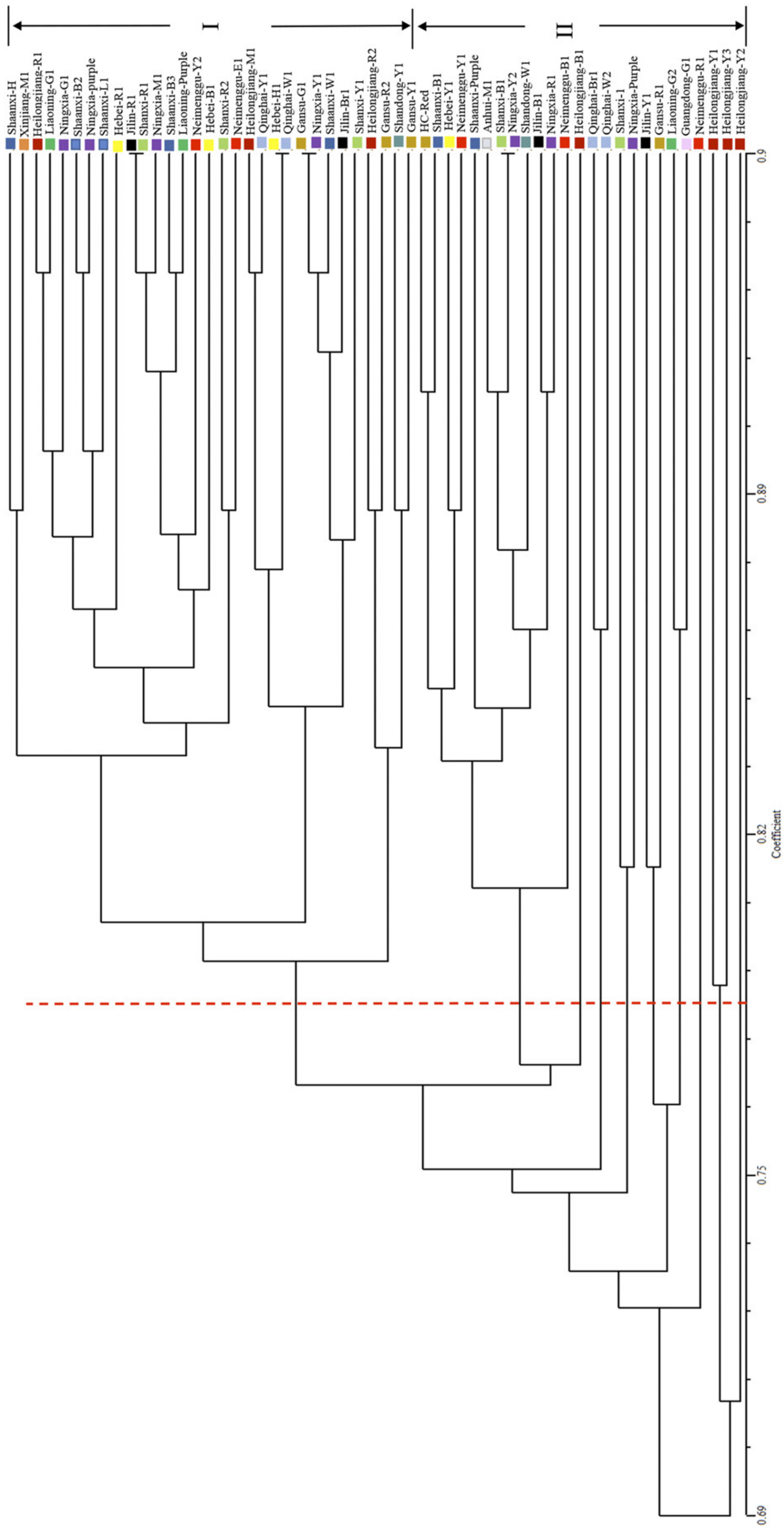

Fig. 8. Dendrogram of 56 proso millet accessions. This phylogenetic tree of proso millet accessions was constructed using unweighted pair-group method with arithmetic mean (UPGMA) cluster analysis. The accessions were clustered into two groups according to genetic similarity coefficient (0.771).

including genetic map construction, quantitative trait loci mapping, and marker-assisted selection.

LITERATURE CITED

Agdag, M., L. Nelson, D. Baltensperger, D. Lyon, and S. Kachman. 2001. Row spacing affects grain yield and other agronomic characters of proso millet. *Communications in Soil Science and Plant Analysis* 32: 2021–2032.

Bagdi, A., G. Balázs, J. Schmidt, M. Szatmári, R. Schoenlechner, E. Berghofer, and S. Tömösközia. 2011. Protein characterization and nutrient composition of Hungarian proso millet varieties and the effect of decortication. *Acta Alimentaria* 40: 128–141.

Bennetzen, J. L., J. Schmutz, H. Wang, R. Percifield, J. Hawkins, A. C. Pontaroli, M. Estep, et al. 2012. Reference genome sequence of the model plant *Setaria. Nature Biotechnology* 30: 555–561.

Chen, Y. F., L. Q. Li, Q. Xu, Y. H. Kong, H. Wang, and W. H. Wu. 2009. The WRKY6 transcription factor modulates *PHOSPHATE1* expression in response to low Pi stress in *Arabidopsis. Plant Cell* 21: 3554–3566.

Cho, Y. I., J. W. Chung, G. A. Lee, K. H. Ma, A. Dixit, J. G. Gwag, and Y. J. Park. 2010. Development and characterization of twenty-five new polymorphic microsatellite markers in proso millet (*Panicum miliaceum* L.). *Genes & Genomics* 32: 267–273.

Choi, Y. Y., K. Osada, Y. Ito, T. Nagasawa, M. R. Choi, and N. Nishizawa. 2005. Effects of dietary protein of Korean foxtail millet on plasma adiponectin, HDL-cholesterol, and insulin levels in genetically type 2 diabetic mice. *Bioscience, Biotechnology, and Biochemistry* 69: 31–37.

Conesa, A., S. Gotz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)* 21: 3674–3676.

Diao, X. M. 2017. Production and genetic improvement of minor cereals in China. *Crop Journal* 5: 103–114.

Dwivedi, S. L., H. D. Upadhyaya, S. Senthilvel, C. T. Hash, K. Fukunaga, X. Diao, D. Santra, et al. 2012. Millets: Genetic and genomic resources. *In* J. Jules [ed.], Plant breeding reviews, 247–375. Wiley-Blackwell, Hoboken, New Jersey, USA.

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* 14: 2611–2620.

Gadaleta, A., A. Giancaspro, S. Zacheo, D. Nigro, S. L. Giove, P. Colasuonno, and A. Blanco. 2011. Comparison of genomic and EST-derived SSR markers in phylogenetic analysis of wheat. *Plant Genetic Resources* 9: 243–246.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, and X. Adiconis. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644–652.

Graybosch, R. A., and D. D. Baltensperger. 2009. Evaluation of the waxy endosperm trait in proso millet (*Panicum miliaceum*). *Plant Breeding* 128: 70–73.

Gupta, P. K., S. Rustgi, S. Sharma, R. Singh, N. Kumar, and H. S. Balyan. 2003. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetics and Genomics* 270: 315–323.

Hamoud, M. A., S. A. Haroun, R. D. MacLeod, and A. J. Richards. 1994. Cytological relationships of selected species of *Panicum* L. *Biologia Plantarum* 36: 37–45.

Hodel, R. G. J., M. A. Gitzendanner, C. C. Germain-Aubrey, X. Liu, A. A. Crowl, M. Sun, L. B. Jacob, et al. 2016. A new resource for the development of SSR markers: Millions of loci from a thousand plant transcriptomes. *Applications in Plant Sciences* 4: 1600024.

Hu, J., L. Wang, and J. Li. 2011. Comparison of genomic SSR and EST-SSR markers for estimating genetic diversity in cucumber. *Biologia Plantarum* 55: 577–580.

Hu, X., J. Wang, P. Lu, and H. Zhang. 2009. Assessment of genetic diversity in *Proso millet* (*Panicum miliaceum* L.) using SSR markers. *Journal of Genetics and Genomics* 36: 491–500.

Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* 9: 1322–1332.

Hunt, H. V., M. G. Campana, M. C. Lawes, Y. J. Park, M. A. Bower, C. J. Howe, and M. K. Jones. 2011. Genetic diversity and phylogeography of broomcorn millet (*Panicum miliaceum* L.) across Eurasia. *Molecular Ecology* 20: 4756–4771.

Hunt, H. V., F. Badakshi, O. Romanova, C. J. Howe, M. K. Jones, and J. S. Heslop-Harrison. 2014. Reticulate evolution in *Panicum* (Poaceae): The origin of tetraploid broomcorn millet, *P. miliaceum. Journal of Experimental Botany* 65: 3165–3175.

Kalia, R. K., M. K. Rai, S. Kalia, R. Singh, and A. K. Dhawan. 2011. Microsatellite markers: An overview of the recent progress in plants. *Euphytica* 177: 309–334.

Kalinova, J., and J. Moudry. 2006. Content and quality of protein in proso millet (*Panicum miliaceum* L.) varieties. *Plant Foods for Human Nutrition* 61: 43.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* 44: D457–D462.

Kumar, A., V. S. Gaur, A. Goel, and A. K. Gupta. 2015. De novo assembly and characterization of developing spikes transcriptome of finger millet (*Eleusine coracana*): A minor crop having nutraceutical properties. *Plant Molecular Biology Reporter* 33: 905–922.

Li, B., and C. N. Dewey. 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323.

Lu, H. Y., J. P. Zhang, K. B. Liu, N. Q. Wu, Y. M. Li, K. S. Zhou, M. L. Ye, et al. 2009. Earliest domestication of common millet (*Panicum miliaceum*) in East Asia extended to 10,000 years ago. *Proceedings of the National Academy of Sciences, USA* 106: 7367–7372.

Nei, M., and W. H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, USA* 76: 5269–5273.

Patel, R. K., and M. Jain. 2012. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE* 7: e30619 https://doi.org/10.1371/journal.pone.0030619.

Porebski, S., L. G. Bailey, and B. R. Baum. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter* 15: 8–15.

Rohlf, F. J. 2008. NTSYSpc: Numerical Taxonomy System, ver. 2.20. Exeter Publishing, Ltd., Setauket, New York, USA.

Saseendran, S. A., D. C. Nielsen, D. J. Lyon, L. Ma, D. G. Felter, D. D. Baltensperger, G. Hoogenboom, and L. R. Ahuja. 2009. Modeling responses of dryland spring triticale, proso millet and foxtail millet to initial soil water in the High Plains. *Field Crops Research* 113: 48–63.

Shimanuki, S., T. Nagasawa, and N. Nishizawa. 2006. Plasma HDL subfraction levels increase in rats fed proso-millet protein concentrate. *Medical Science Monitor* 12: 221–226.

Sonah, H., R. K. Deshmukh, A. Sharma, V. P. Singh, D. K. Gupta, R. N. Gacche, J. C. Rana, et al. 2011. Genome-wide distribution and organization of microsatellites in plants: An insight into marker development in *Brachypodium. PLoS ONE* 6: e21298 https://doi.org/10.1371/journal.pone.0021298.

Srivastava, S., A. Thathola, and A. Batra. 2001. Development and nutritional evaluation of proso millet-based convenience mix for infants and children. *Journal of Food Science and Technology* 38: 480–483.

Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 28: 33–36.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen. 2012. Primer3—New capabilities and interfaces. *Nucleic Acids Research* 40: e115.

Varshney, R. K., T. Thiel, N. Stein, P. Langridge, and A. Graner. 2002. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cellular & Molecular Biology Letters* 7: 537–546.

Warwick, S. I. 1987. Isozyme variation in proso millet. *Journal of Heredity* 78: 210–212.

Yeh, F. C., R. C. Yang, and T. Boyle. 1999. POPGENE version 1.31: Microsoft Windows–based freeware for population genetic analysis, quick user guide. Center for International Forestry Research, University of Alberta, Edmonton, Alberta, Canada.

Zhang, G., X. Liu, Z. Quan, S. Cheng, X. Xu, S. Pan, M. Xie, et al. 2012. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology* 30: 549–554.

Zhang, M., W. Mao, G. Zhang, and F. Wu. 2014. Development and characterization of polymorphic EST-SSR and genomic SSR markers for Tibetan annual wild barley. *PLoS ONE* 9: e94881.

Appendix 1. The 56 proso millet cultivars and their geographic origins in China.

No.	Accession name	Cultivar area
1	Shaanxi-H	Shaanxi, China[a,c]
2	HC-Red	Gansu, China[a,c]
3	Shaanxi-B1	Shaanxi, China[a]
4	Shaanxi-Purple	Shaanxi, China[a]
5	Shanxi-1	Shanxi, China[b,c]
6	Heilongjiang-Y1	Heilongjiang, China[a,c]
7	Heilongjiang-Y2	Heilongjiang, China[a]
8	Heilongjiang-R1	Heilongjiang, China[a]
9	Heilongjiang-M1	Heilongjiang, China[a]
10	Heilongjiang-B1	Heilongjiang, China[a]
11	Heilongjiang-R2	Heilongjiang, China[a]
12	Heilongjiang-Y3	Heilongjiang, China[a]
13	Jilin-Y1	Jilin, China[a,c]
14	Jilin-R1	Jilin, China[a]
15	Liaoning-G1	Liaoning, China[a,c]
16	Liaoning-G2	Liaoning, China[a]
17	Neimenggu-B1	Inner Mongolia, China[a,c]
18	Ningxia-M1	Ningxia, China[a,c]
19	Shaanxi-B2	Shaanxi, China[a]
20	Shaanxi-B3	Shaanxi, China[a]
21	Neimenggu-Y2	Inner Mongolia, China[a]
22	Ningxia-Purple	Ningxia, China[a]
23	Gansu-G1	Gansu, China[a]
24	Gansu-R1	Gansu, China[a]
25	Shaanxi-L1	Shaanxi, China[a]
26	Guangdong-G1	Guangdong, China[a,c]
27	Anhui-M1	Anhui, China[a,c]
28	Gansu-R2	Gansu, China[a]
29	Shaanxi-W1	Shaanxi, China[a]
30	Jilin-Br1	Jilin, China[a]
31	Jilin-B1	Jilin, China[a]
32	Liaoning-Purple	Liaoning, China[a]
33	Hebei-R1	Hebei, China[a,c]
34	Hebei-B1	Hebei, China[a]
35	Shanxi-R1	Shanxi, China[b,c]
36	Shanxi-R2	Shanxi, China[b]
37	Shanxi-B1	Shanxi, China[b]
38	Ningxia-G1	Ningxia, China[a]
39	Ningxia-Y1	Ningxia, China[a]
40	Shanxi-Y1	Shanxi, China[b]
41	Shandong-Y1	Shandong, China[a,c]
42	Ningxia-Purple	Ningxia, China[a]
43	Hebei-H1	Hebei, China[a]
44	Hebei-Y1	Hebei, China[a]
45	Neimenggu-R1	Inner Mongolia, China[a]
46	Neimenggu-Y1	Inner Mongolia, China[a]
47	Ningxia-R1	Ningxia, China[a]
48	Neimenggu-E1	Inner Mongolia, China[a]
49	Ningxia-Y2	Ningxia, China[a]
50	Shandong-W1	Shandong, China[a]
51	Xinjiang-M1	Xinjiang Uyghur Autonomous Region, China[a,c]
52	Gansu-Y1	Gansu, China[a]
53	Qinghai-Y1	Qinghai, China[a]
54	Qinghai-Br1	Qinghai, China[a]
55	Qinghai-W1	Qinghai, China[a]
56	Qinghai-W2	Qinghai, China[a]

[a] Chinese Crop Germplasm Resources Information System (CGRIS).

[b] Institute of Crop Genetic Resources, Shanxi, China.

[c] Fourteen representative accessions used for detecting polymorphic microsatellite markers.