frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 July 2022
DOI 10.3389/fpls.2022.959042

OPEN ACCESS

EDITED BY
Xiaoyu Weng,
University of Texas at Austin,
United States

REVIEWED BY
Liangliang Chen,
Howard Hughes Medical Institute
(HHMI), United States
Sajeevan Radha Sivarajan,
Swedish University of Agricultural
Sciences, Sweden
Likai Wang,
Nanjing Agricultural University, China

*CORRESPONDENCE
Shiqie Bai
baishiqie@126.com
Xiao Ma
maroar@126.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to
Plant Systematics and Evolution,
a section of the journal
Frontiers in Plant Science

RECEIVED 01 June 2022
ACCEPTED 28 June 2022
PUBLISHED 25 July 2022

CITATION
Xiong Y, Yang J, Xiong Y, Zhao J, Liu L, Liu W, Sha L, Zhou J, You M, Li D, Lei X, Bai S and Ma X (2022) Full-length transcriptome sequencing analysis and characterization, development and validation of microsatellite markers in *Kengyilia melanthera*. *Front. Plant Sci.* 13:959042.
doi: 10.3389/fpls.2022.959042

Full-length transcriptome sequencing analysis and characterization, development and validation of microsatellite markers in *Kengyilia melanthera*

Yanli Xiong[1†], Jian Yang[1†], Yi Xiong[1], Junming Zhao[1], Lin Liu[1], Wei Liu[1], Lina Sha[1], Jiqiong Zhou[1], Minghong You[2], Daxu Li[2], Xiong Lei[2], Shiqie Bai[2*] and Xiao Ma[1*]

[1]College of Grassland Science and Technology, Sichuan Agricultural University, Chengdu, China,
[2]Sichuan Academy of Grassland Science, Chengdu, China

As a typical psammophyte of the Triticeae, *Kengyilia melanthera* possesses high feeding potential and great utilization values in desertification control in the Qinghai-Tibet Plateau. However, few gene function and genetic studies have been performed in *K. melanthera*. In this study, single-molecule real-time sequencing technology was used to obtain the full-length transcriptome sequence of *K. melanthera*, following the functional annotation of transcripts and prediction of coding sequences (CDSs), transcription factors (TFs), and long noncoding RNA (lncRNA) sequences. Meanwhile, a total of 42,433 SSR loci were detected, with 5′-UTRs having the most SSR loci and trinucleotide being the most abundant type. In total, 108,399 SSR markers were designed, and 300 SSR markers were randomly selected for diversity verification of *K. melanthera*. A total of 49 polymorphic SSR markers were used to construct the genetic relationships of 56 *K. melanthera* accessions, among which 21 SSR markers showed good cross-species transferability among the related species. In conclusion, the full-length transcriptome sequence of the *K. melanthera* will assist gene prediction and promote molecular biology and genomics research, and the polymorphic SSR markers will promote molecular-assisted breeding and related research of *K. melanthera* and its relatives.

KEYWORDS

***Kengyilia melanthera*, full-length transcriptome, SMRT sequencing, SSRs development, transferability**

Background

The *Kengyilia* is a perennial genus in the tribe Triticeae of Poaceae with the genome constitution of StStYYPP ($2n = 6 \times = 42$; Yen and Yang, 2020). It is only distributed in the temperate zone of the northern hemisphere with the distribution range roughly between 29° to 48° N and 57° to 104° E (Bing, 2001). The genus *Kengyilia* can provide important genetic resources for forage breeding and cereal improvement owing to its high feeding values and excellent resistance to biotic and abiotic stresses (Yen and Yang, 2020). As a typical psammophyte species of *Kengyilia* genus, *Kengyilia melanthera* mainly distributed in the Qinghai-Tibet Plateau (QTP) areas at an altitude of 3,300–4,750 m and possessed great utilization potential in desertification control and ecological restoration (Yen and Yang, 2020). Furthermore, the strong drought resistance and high carbohydrate content of *K. melanthera* make it suitable for artificial grassland building in QTP areas to cope with the seasonal shortage of forage.

Kengyilia is a new genus established in 1990 (Yen and Yang, 1990). Therefore, more attention has been paid to the genetic diversity and taxonomy of *Kengyilia* species in previous studies. The application of gene sequences in genetic diversity detection of *Kengyilia* including nuclear genes, namely, *Pgk1* (Fan et al., 2012) and *DMC1* (Gao et al., 2016), and chloroplast genes, namely, *trnL-F* (Zhang et al., 2009), *trnT-trnL* (Gao et al., 2014), and *matK* (Luo et al., 2012), was reported. Considering the simplicity, polymorphism, and stability, some molecular markers (Zhou et al., 2000; Zhang et al., 2005, 2008) have been applied to the phylogenetic relationship analysis of *Kengyilia*. Compared to those markers, SSR markers have greater advantages including their easy development, codominance, and high polymorphism (Wu et al., 2020).

Compared to G-SSR (genomic SSR), EST-SSR (expressed sequence tag SSR) possesses higher cross-species transferability and the ability to mark functional genes (Karcı et al., 2020). Recently, the increasing utilization of transcriptome sequencing makes it an efficient and cost-effective tool for EST-SSR marker development. Although there have been many reports about SSR identification and development *via* second-generation transcriptomic sequencing, little research is concerned with the SSR development using full-length transcriptome sequencing technology (Ge et al., 2019; Wu et al., 2020). Compared to the second-generation transcriptome sequencing, the SMRT sequencing technology hold more advantages including longer read length, less sequencing, and assembly error, which is more beneficial for the study of plants without a reference genome (Thomas et al., 2014; Gordon et al., 2015). The acquisition of full-length transcripts contributed to the gene annotation and identification of isoforms, fusion transcripts, and long noncoding RNA (lncRNA) (Zhao et al., 2019). Therefore, more SSRs can be discovered, and complete functional genes can be explored in this case.

In this study, first, the full-length transcriptome of the *K. melanthera* was obtained, and the functional annotation was performed to better understand its functional classification. Second, we characterized the SSRs of the *K. melanthera* by analyzing the frequency, distribution, and function of SSRs in the transcriptome. Finally, newly developed EST-SSR markers were used for genetic diversity and structure study in 56 *K. melanthera* accessions and other *Kengyilia* populations.

Materials and methods

Plant material collection and DNA and RNA extraction

The *K. melanthera* was collected from the nursery base of Sichuan Academy of Grassland Sciences (Chengdu, China) in Qiongxi Town, Hongyuan County (32°48′N, 102°33′E) (**Supplementary Figure 1**). To obtain as many transcripts as possible, the roots, stems, leaves, and inflorescences of a single plant were collected and stored in liquid nitrogen rapidly. A total of mixed 2 g sample containing the equal amounts of each tissue was used for RNA extraction and full-length transcription sequencing. The wild germplasms used to identify the polymorphism of the developed SSRs were collected from the nursery base of Sichuan Academy of Grassland Sciences in Hongyuan County, the sandy land in Waqie town (33°10′N, 102°37′E) and the sandy land in Dazhasi town (33°40′N, 102°56′E). The young leaves were stored in silica gel (**Supplementary Table 1**). The seeds of a further four related species, *Kengyilia alatavica*, *Kengyilia batalinii*, *Kengyilia mutica*, and *Kengyilia rigidula*, all came from the U.S. National Plant Germplasm System and were planted in the greenhouse of Sichuan Agricultural University (**Supplementary Table 2**). Total DNA and RNA were extracted by DNA extraction kit and RNA extraction kit (Tiangen Biotech, Beijing, China), respectively. Their concentration and quality were checked using NanoDrop1 ND-1000 Spectrophotometer (NanoDrop Technologies, United States) and agarose gel electrophoresis, respectively.

PacBio library construction and sequencing

The qualified mixed RNA samples of roots, stems, leaves, and inflorescences were used to construct cDNA library. SMARTer PCR cDNA Synthesis Kit (Clontech, Mountain View, CA, United States) was used to synthesize cDNA, and then PCR amplification, quality control, and purification were performed. The 1–6 kb cDNA fragments were generated using BluePippin Size Selection System (Sage Science, United States). The SMRT sequencing libraries was constructed using the Pacific Biosciences DNA Template Prep Kit 2.0. Qubit 2.0 and Agilent 2100 were used to detect the concentration and quality of cDNA libraries, respectively. Finally, SMRT sequencing was performed on the PacBio RS II platform (Pacific Biosciences, Menlo Park, CA, United States).

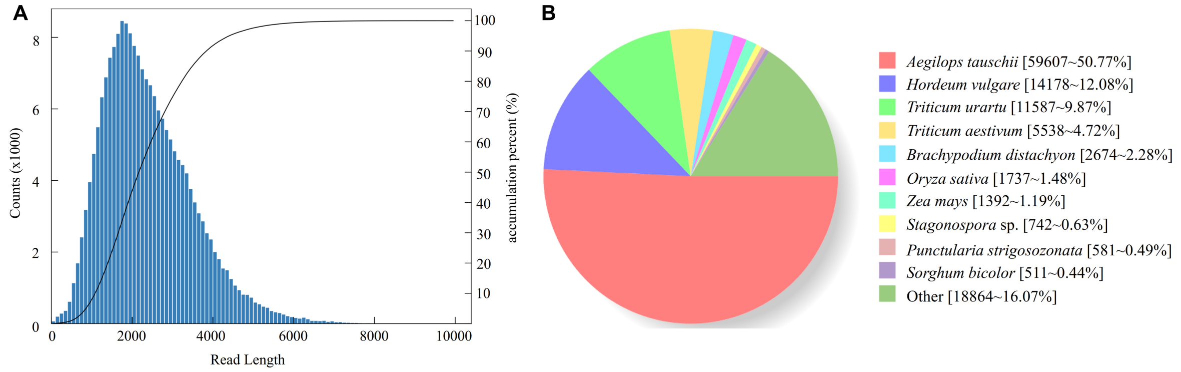

FIGURE 1
Length distribution of 204,028 consensus isoforms in *K. melanthera* **(A)** and homologous species distribution of *K. melanthera* based on the NCBI non-redundant protein sequences (NR) database **(B)**.

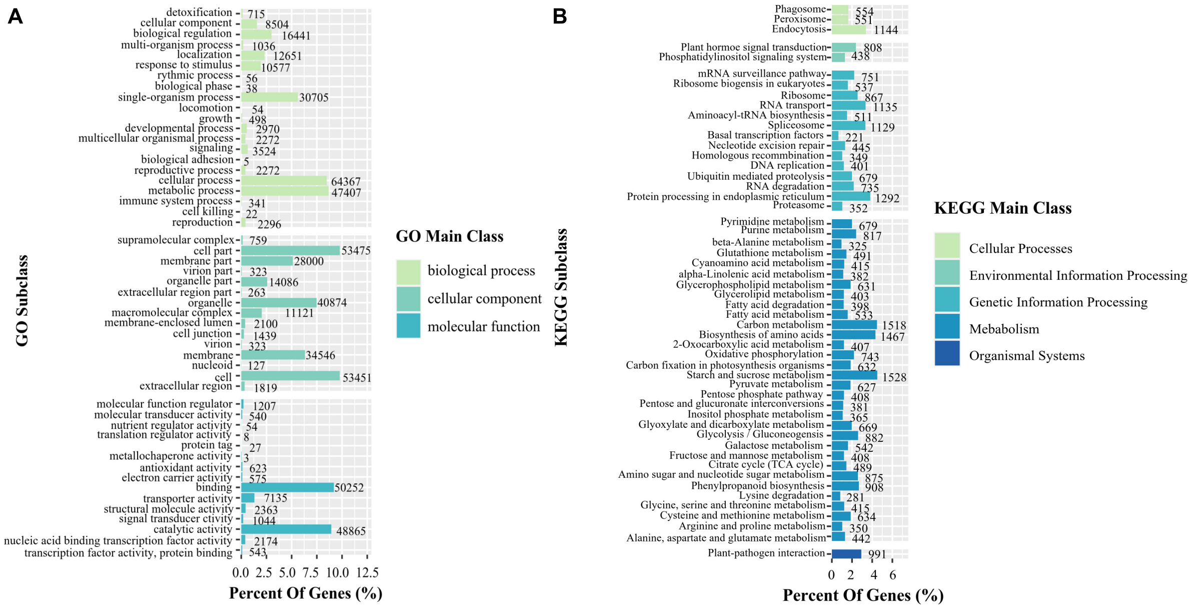

FIGURE 2
Functional classification of transcripts based on the GO **(A)** and KEGG **(B)** in *K. melanthera*.

Quality filtering and error correction of PacBio long reads

Raw reads were processed into circular consensus sequences (CCSs) according to the adaptor, with full pass ≥3 and sequence accuracy >0.9. Next, full-length, non-chimeric transcripts were detected by searching the polyA tail signal and the 5′ and 3′ cDNA primer sequences in CCSs. The IsoSeq module in SMRT Link version 5.0.1 software (Pacific Biosciences of California, Inc., Menlo Park, CA, United States) was used to group the full-length sequences of the same transcript, and the similar full-length sequences were grouped into a cluster. Each cluster contained a consistent sequence, and the corrected consistent sequences were used to obtain high-quality sequences (accuracy >99%) for subsequent analysis. After removing the low-quality and high-quality redundant sequences (identity >0.99) using CD-HIT software (Li and Godzik, 2006), non-redundant high-quality transcripts were obtained.

Functional annotation

The obtained transcript sequences were aligned to the NCBI non-redundant protein sequences (NR) (Deng et al., 2006), Protein family (Pfam) (Finn et al., 2014), Clusters of Orthologous Groups of proteins (COG) (Tatusov et al., 2000), euKaryotic Ortholog Groups (KOG) (Koonin et al., 2004), Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (eggNOG), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Minoru et al., 2004), Gene Ontology (GO) (Ashburner et al., 2000) using BLAST v2.2.26 (Altschul et al., 1997) (*E*-value < 10-5).

Predictions of coding sequences, transcription factors, and long noncoding RNA

TransDecoder v3.0.0[1] was used to identify coding sequences (CDSs) based on Pfam database. ITAK v1.2 (Yi et al., 2016) was used to identify transcription factors (TFs), transcription regulators (TRs), and protein kinases (PKs).

Coding Potential Calculator (CPC) (Kong et al., 2007), Coding-Non-Coding Index (CNCI), Coding Potential Assessment Tool (CPAT) (Wang et al., 2013), and Pfam database were used to screen nonprotein-coding RNA candidates. The lncRNA candidates were predicted with the following criterion: the transcripts longer than 200 nt and possessing more than two exons. The predicted lncRNA candidates were further screened in CPC/CNCI/CPAT/Pfam.

SSRs identification and primer design

Transcripts with the length more than 500 bp were used to identify SSRs based on MISA (Beier et al., 2017) software. The SSR loci were identified with the following criteria: repeat numbers of mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide motifs greater than or equal to 10, 6, 5, 5, 5, and 5, respectively. Finally, Primer3.0 (Untergasser et al., 2012) was used to design primers with the primer size of 18–25 bp, annealing temperature of 55–65°C, GC content of 30–70%, and product length of 100–300 bp.

Validation and evaluation of SSR markers

A total of 300 SSR primer pairs were randomly selected for synthesis and further validation. The PCR system included 10 μl of 2× Taq Master Mix (Tiangen Biotech, Beijing, China), 1 μl of forward/reverse primer (20 ng/μl), 2 μl of genomic DNA (20 ng/μl), and 6 μl of ddH_2O. PCR was performed on Biometra Tadvanced under the following process: 94°C for 5 min, followed by 35 cycles including 30 s at 94°C, 30 s at 56 or 58°C, and 30 s at 72°C, and then extension at 72°C for 5 min.

Given that the heterohexaploid feature of *K. melanthera*, it is difficult to record the allelic variation. Therefore, the amplified bands were recorded as the 0/1 (presence/absence of bands) binary matrix. GenAlEx 6.503 (Peakall and Smouse, 2010) was used to calculate the number of alleles (Na), number of effective alleles (Ne), and Shannon's information index (I). The polymorphic information (PIC) and expected heterozygosity (He) were calculated based on the formula of $\text{PIC} = 1 - p^2 - q^2$, where *p* and *q* are frequency of present/absent band; $\text{He} = 1 - \Sigma pi^2$, where *pi* is frequency of the *i*-th allele (Zhang et al., 2019; Zheng et al., 2020). The Nei genetic distance (GD) matrix was obtained by Freetree (Hampl et al., 2001), the UPGMA (unweighted pair-group method with arithmetic means) dendrogram was constructed, and the visualization was performed on Figtree (Hampl et al., 2001). Population structure was speculated using STRUCTURE v2.3.4 (Falush et al., 2007) with 50,000 burn-in and 100,000 Monte Carlo Markov chain (MCMC). Then the optimal *K* value was determined using STRUCTURE HARVESTER (Earl and Vonholdt, 2012). The principal coordinate analysis (PCoA) was carried out by NTSYS v2.2 (Rohlf, 1987).

Results

General properties and functional annotations of full-length transcriptome

A total of 96.1 GB raw data was obtained (GenBank database accession number: PRJNA735213). There were 542,441 CCSs (read bases of 1,399,665,069) with an average length of 2,397 bp and mean number of passes of 35. We obtained 491,001 full-length non-chimeric (FLNC) reads, and the percentage of FLNC is 90.52%. After clustering the FLNC sequences, we obtained 4,580 polished low-quality isoforms and 199,134 (97.60%) polished high-quality isoforms. Finally, 204,028 consensus isoforms were obtained, with the average length of 2,399 bp (**Figure 1A**). After removing the low-quality and redundant high-quality transcripts, 126,410 transcripts were obtained. In order to access the completeness and accuracy of transcripts obtained in this study, we also aligned the transcripts to OrthoBD database. The results showed that 82.57% (1,189 of 1,440) of transcripts were completed with only 4.52 and 12.91% being fragmented and missing.

A total of 118,341 transcripts of *K. melanthera* were perfectly matched with the COG, GO, KEGG, KOG, Pfam Annotation, Swissprot Annotation, eggNOG, and Nr databases (**Supplementary Table 3**). Sequence alignment based on NR database showed that three species with the highest homology with *K. melanthera* all belonged to the Triticeae (**Figure 1B**), among which *Aegilops tauschii* (50.77% of transcripts) had the highest homology with *K. melanthera*, followed by *Hordeum*

[1] https://transdecoder.github.io/

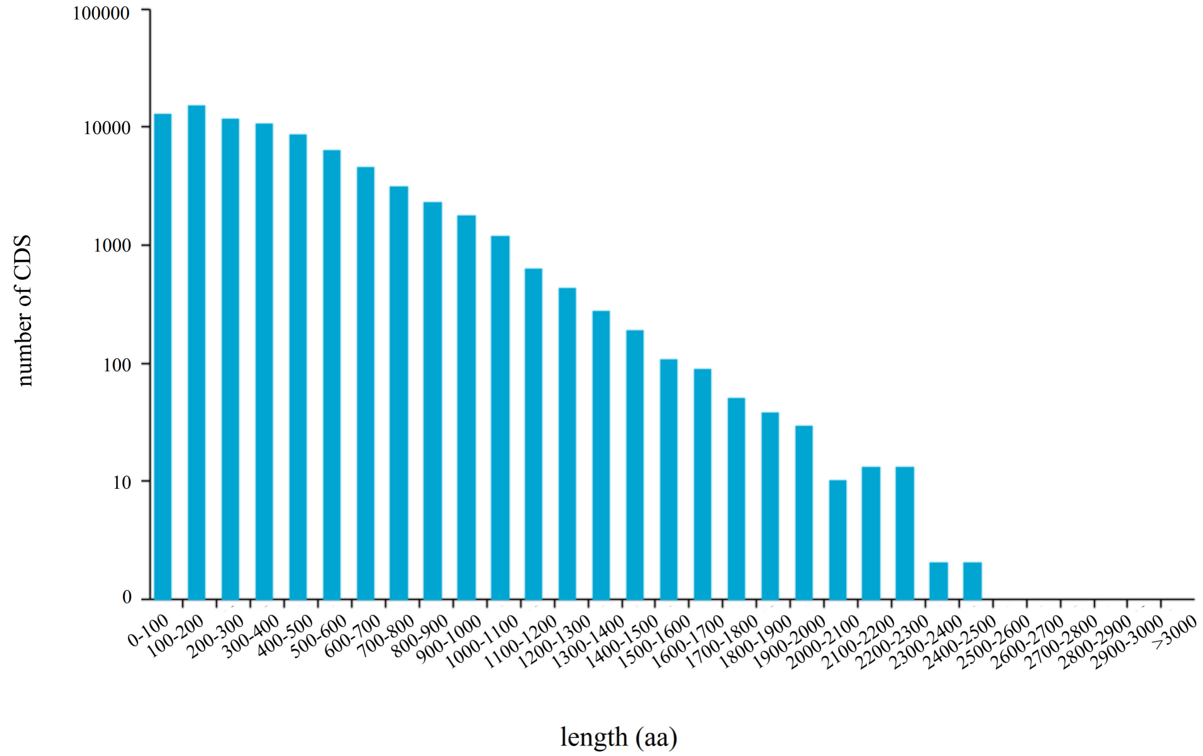

FIGURE 3
The length distribution of 79,159 CDS-encoded protein sequences.

vulgare (12.08% of transcripts) and *Triticum urartu* (9.87% of transcripts). A total of 91,155 and 44,928 transcripts were assigned into 51 subcategories of three major GO functional categories and 130 KEGG pathways (**Figures 2A,B**), respectively. The pathways involving the largest number of transcripts were *starch and sucrose metabolism* (1,528), *carbon metabolism* (1,518), and *biosynthesis of amino acids* (1,467).

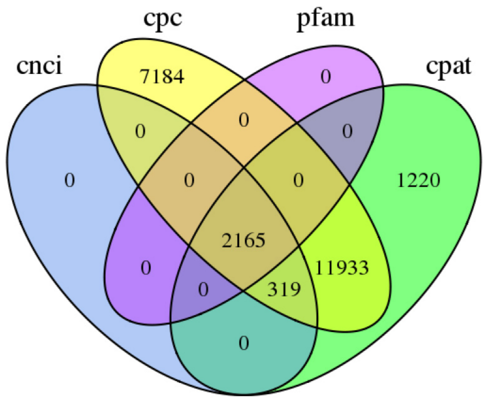

FIGURE 4
The number of long noncoding RNA transcripts predicted according to the CPC, CNCI, CPAT, and Pfam databases.

In this study, a total of 48,095, 110,413, and 69,144 transcripts were assigned to 25 COG categories, 23 eggNOG categories, and 25 KOG categories. The largest number of transcripts were annotated in the *signal transduction mechanism* category (5,893) in COG categories (**Supplementary Figure 2a**). Among the 23 eggNOG categories, in addition to *function unknown*, most of the transcripts (8,403) were annotated with the *signal transduction mechanism* category (**Supplementary Figure 2b**). In KOG categories, most transcripts were assigned to the *general function prediction only* category (12,904), followed by the *signal transduction mechanism* (10,135) (**Supplementary Figure 2c**).

Analysis of coding sequences, transcription factors, and long noncoding RNAs

A total of 79,159 CDSs (**Figure 3**) and 12,355 TFs belonging to 222 families were identified (**Supplementary Table 4**). The most abundant TF family was *RLK-Pelle_DLSV* (1,122), followed by *C2H2* (412) and *RLK-Pelle_LRR-XII-1* (282). Four databases CNCI, CPC, Pfam, and CPAT were used to predict the lncRNAs. The results showed that there

TABLE 1 General information for the SSR analysis.

Items	Number
SSR number detected	42,433
Number of sequences containing SSR	31,862
Number of sequences containing ≥1 SSR	7,535
Number of compound SSRs	3,083
Frequency of SSRs loci (%)	35.87
Distribution density of SSRs loci (kb)	7.19

were 2,484, 21,601, 2,165, and 15,637 lncRNA candidates with length ≥200 bp and exon ≥2. Finally, 2,165 transcripts shared in four databases were considered as the potential lncRNA (Figure 4).

Characterization of SSRs in transcriptome

A total of 42,433 SSRs (Table 1) were identified from 31,862 transcripts in the *K. melanthera* transcriptome. The frequency of SSRs was 35.87%, and an average of one SSRs loci was found every 7.19 kb. The most abundant repeat motif type was trinucleotide (45.1%), followed by mononucleotide (30.1%) and dinucleotide (21.0%) (Supplementary Figure 3), while pentanucleotide and hexanucleotide accounted only for 0.6 and 0.5%. A/T (83.74%), AG/CT (54.75%), CCG/CGG (31.09%), ACAT/ATGT (11.28%), AGAGG/CCTCT (12.41%), and ACCGCC/CGGTGG (5.56%) were the most abundant motifs in mono- to hexanucleotide repeats, respectively (Supplementary Figure 3).

We also explored the position distribution feature of SSRs in *K. melanthera* transcriptome (Figure 5). A total of 25,944 SSRs were identified in the CDS and untranslated regions (UTRs), of which the 5′-UTR region had the most SSRs (13,234), followed by the 3′-UTR (8,270) and the CDS region (4,440). Totally, trinucleotide (85.4%), mononucleotide (53.3%), and dinucleotide SSRs (42.3%) were the most repeat types in the CDS, 3′-UTRs and 5′-UTRs regions.

In total, 8,458 transcripts holding SSRs were categorized into three functional categories (Supplementary Figure 4a). *Metabolic process* (5,227) and *cellular process* (4,285) were the two subcategories with the most transcripts in the "biological process category." In the "cellular component," *cell* (2,319) and *cell part* (2,322) are the two subcategories that involved the greatest number of transcripts. The two subcategories of *catalytic activity* (4,541) and *binding* (4,938) involved the most transcripts in the "molecular function category." In addition, based on KEGG pathway analysis, 8,833 transcripts were assigned to 124 pathways. The three most abundant transcript pathways are *Carbohydrate metabolism* (316), *Starch and sucrose metabolism* (294), and *Spliceosome* (262) (Supplementary Figure 4b).

Development and polymorphism identification of SSR markers

A total of 108,399 primer pairs were designed in 36,133 SSR loci (Supplementary Table 5), among which 300 SSR primer pairs from different SSR loci were selected to amplify eight selected DNA samples. A total of 208 primers pairs (69.3%) successfully produced an amplicon of the expected size and 49 primer pairs (16.3%) showing good polymorphism were used to amplify the 56 *K. melanthera* accessions (Supplementary Table 6). The gel figure of Km-eSSR231 was presented in Supplementary Figure 5. A total of 358 alleles (Na) and 285.68 effective alleles (Ne) were produced, with the average values of 7.31 and 5.83, respectively (Supplementary Table 7). The average values of I and He were 0.365 and 0.464, respectively. The PIC values ranged from 0.025 to 0.431 (marker Km-eSSR42).

The gene family annotations of the transcripts where the newly developed polymorphic SSR resides were listed in Supplementary Table 6. It is noteworthy that many SSRs were annotated to genes with the important functions in plant metabolism, growth and development, and resistance to adverse growth environments. For example, seven transcripts are annotated as PK genes, five as ABC transporter genes, and some as *POT* gene, *Spt20* gene, *Hsp70* gene, *F-box* gene, and cytochrome *P450* gene.

Population structure analysis using newly developed SSR markers

Unweighted pair-group method with arithmetic means cluster analysis was used to construct the phylogenetic tree using 49 pairs of SSR primer developed in this study. The results showed that 56 wild *K. melanthera* germplasms could be grouped into three clusters basically corresponding to their geo-locations (Figure 6). The Cluster I included 28 germplasms collected from Waqie town (WQ) and one from Dazhasi town (DZS). The Cluster II included 22 germplasms from Dazhasi town. The Cluster III included five germplasms collected from Sichuan Academy of Grassland Sciences (SAGS) and one from Dazhasi town. STRUCTURE analysis showed that the optimal *K* value was 3 (Supplementary Figure 6), which indicated that the tested germplasms possessed three potential genetic memberships. It is interesting to note that the germplasms in each cluster had the same main genetic background based on the STRUCTURE results. The results of PCoA analysis were similar with the UPGMA (Figure 7).

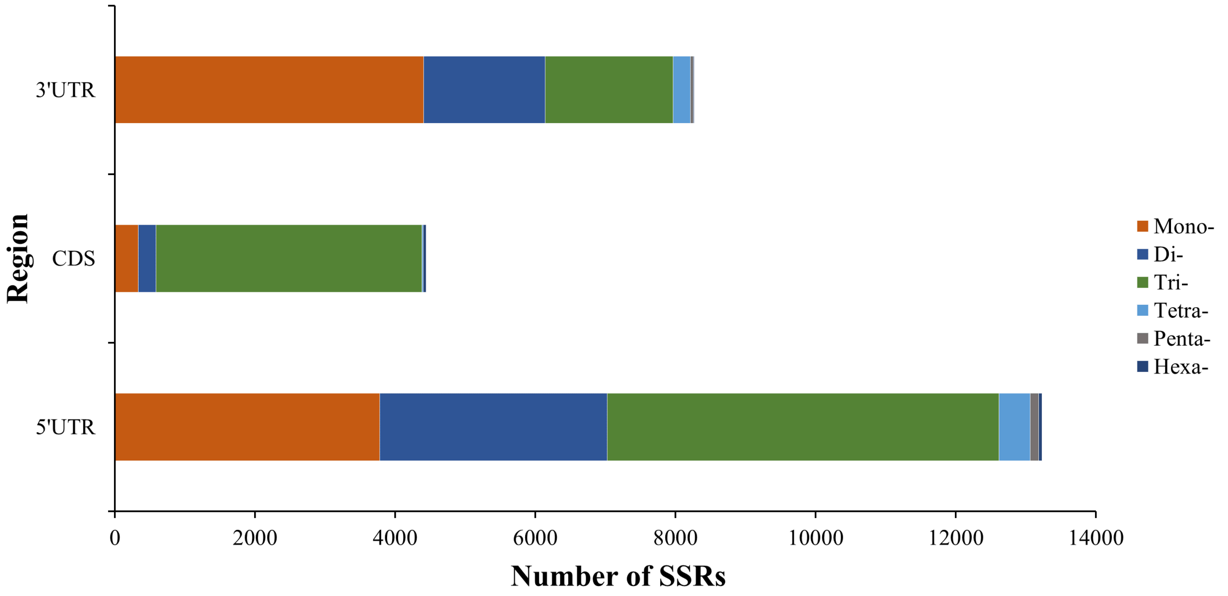

FIGURE 5
Distribution of six SSR repeat types in different genic regions.

Verification of cross-species transferability

In this study, 22 newly developed SSR markers with high polymorphisms within *K. melanthera* were examined for cross-species transfer. Except for Km-eSSR193, 21 markers were successfully amplified in all four species (*K. batalinii*, *K. mutica*, *K. rigidula*, and *K. alatavica*) (**Supplementary Table 8**). The cluster analysis of five *Kengyilia* species was carried out based on the amplified data, and the results demonstrated that *K. alatavica* had closer genetic relationship with *K. batalinii* (**Supplementary Figure 7**).

Discussion

Transcriptome sequencing can quickly and efficiently present information with wide coverage and high accuracy, which has become a good tool for the development of molecular markers and has been widely utilized in the fields of genetics and breeding, germplasm resources protection and development (Wang et al., 2018). Compared to the next-generation sequencing, full-length transcriptome sequencing can provide an efficient and convenient way to obtain transcriptome information of non-model plants without reference genomes (Wang et al., 2019; Zhao et al., 2019). Here, 96.1 GB of raw data was obtained, with the average transcripts length of 2,399 bp. Compared with other gramineous plants, it is shorter than *Saccharum officinarum* (3,099 bp) (Piriyapongsa et al., 2018) and longer than *Carex breviculmis* (2,302 bp) (Teng et al., 2019), *Cynodon dactylon* (2,317 bp) (Zhang et al., 2018), and *Lolium perenne* (2,192 bp) (Xie et al., 2020). The full-length transcriptome obtained in our study could accelerate further related studies of *K. melanthera* and its relatives.

K. melanthera possesses excellent resistance to drought and wind erosion, which makes it an ideal material for desertification control (Yen and Yang, 2020). However, the allohexaploid nature makes the genome assembly difficult, which limits the mining and research of excellent resistance genes contained in *K. melanthera*. Benefiting from the long-read length of third-generation transcriptome sequencing, a total of 118,341 (93.62%) non-redundant transcripts were annotated based on the public databases. Those annotated sequences of *K. melanthera* will provide a reference for the assembly of short-read transcriptome sequencing, thus laying a foundation for the subsequent exploration of drought and barren tolerance genes of *K. melanthera*. In addition to functional annotation of transcripts, CDSs, TFs, and lncRNAs were also predicted, which could provide the data reference for later related researches.

Distribution of SSRs in transcriptome

SSRs continue to be the marker of choice for surveys of genetic diversity (Merritt et al., 2015). Compared to G-SSRs, EST-SSRs were intimately connected to the functional genes and a lot of EST-SSRs have been developed from the transcriptome data to perform the genetic diversity analysis. Our results predicted the abundant SSRs (42,433 SSRs) based on the transcriptome data of *K. melanthera*. The SSRs frequency was

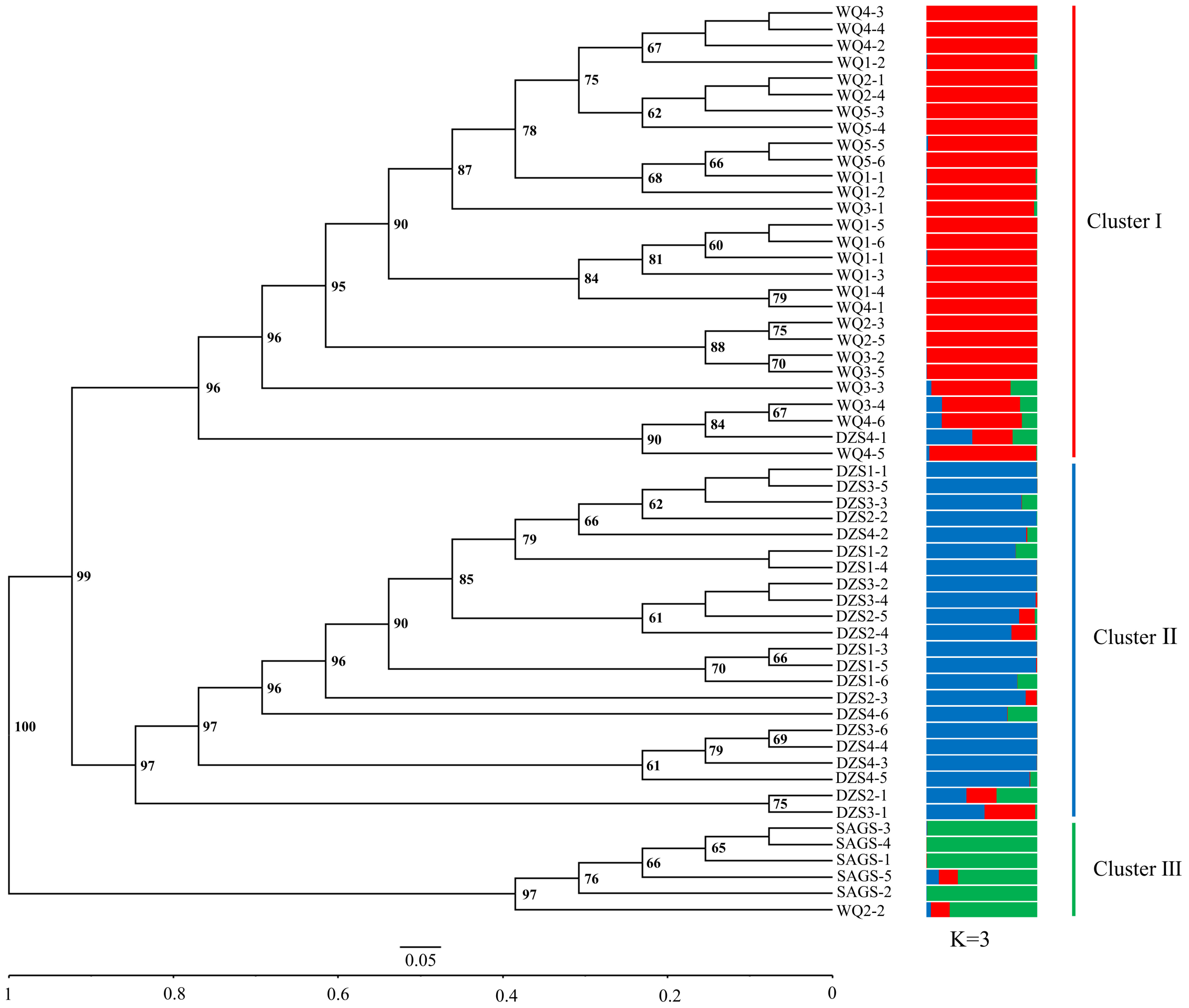

FIGURE 6
The UPGMA dendrogram and genetic structure of 56 *K. melanthera* individuals based on 49 developed SSR markers. WQ, Waqie town; DZS, Dazhasi town; SAGS, Sichuan Academy of Grassland Science.

35.87%, which was much higher than that of SSR frequency obtained from NGS sequencing in related *Elymus sibiricus* (8.19%, 1/6.95 kb) (Zhou et al., 2016), *Leymus chinensis* (4.38%, 1/10.78 kb) (Chen et al., 2013), and *Pennisetum purpureum* (10.89%, 1/6.45 kb) (Wang et al., 2017). The A/T and CCG/CGG rich tendency of mononucleotide and trinucleotide motifs was also consistent with the eukaryotes (Tóth et al., 2000). The most abundant dinucleotide repeat motif was AG/CT (54.75%), which was also the same as that of *E. sibiricus* (Zhou et al., 2016) and *Lolium multiflorum* (Pan et al., 2018).

Microsatellites were found to be non-randomly distributed in gene regions, including CDS, UTRs, and introns (Li et al., 2004). The results of this study showed that UTR possessed a higher number of SSRs compared to CDS regions, which was also found in other species (Vieira et al., 2016). The possible reason is that SSRs have a high mutation rate, and the structure and function of genes will be severely changed if mutation occurs in CDS regions (Xu et al., 2020). Among the six repeat types, SSRs located in the CDS region are dominated by trinucleotide repeat motifs, and their distribution proportion was much higher than that of UTRs, because trinucleotide repeat motifs are less likely to cause frameshift mutations (Metzgar et al., 2000). At the same time, 5′-UTRs contained more trinucleotide repeat motifs than 3′-UTRs, because the SSR variations in 5′-UTRs could affect gene expression (Li et al., 2004).

Development and transferability of polymorphic SSR markers

At present, only a few RAPD (Fan et al., 2012), RAMP (Gao et al., 2016), and ISSR (Zhang et al., 2009) markers have been applied in phylogeny and genetic variation analysis of

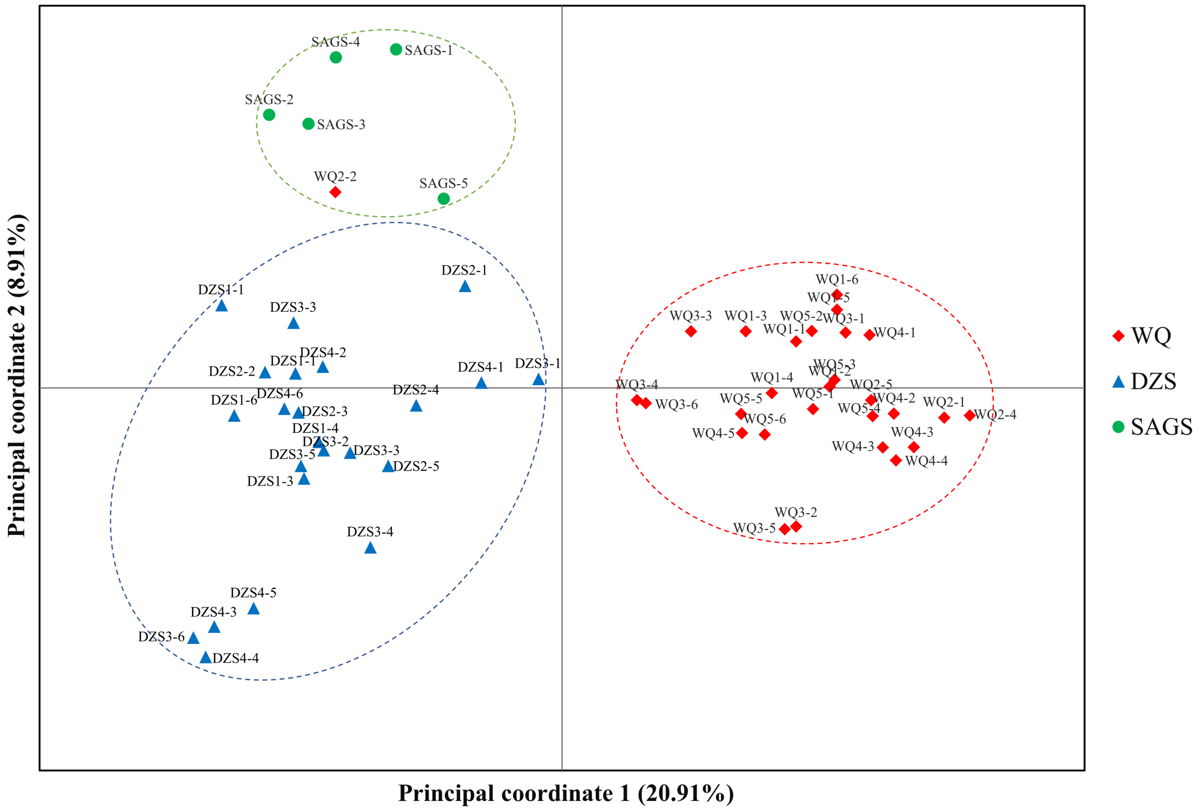

FIGURE 7
The PCoA analysis of 56 *K. melanthera* individuals based on 49 developed SSR markers.

Kengyilia, and few reports are available on the development of SSR markers. Through transcriptome sequencing, we were able to find a large number of SSRs residing on gene sequences. In this study, 49 SSR markers with good polymorphism were selected from 300 candidate SSR markers with an average PIC value of 0.24, which was close to the *E. sibiricus* (0.25) (Hampl et al., 2001). Considering the PIC value ranged from 0 to 0.5 for SSR markers, the EST-SSR markers developed in this study possessed the potential for further genetic study of *K. melanthera* and its relatives.

It has been found that mutations in SSR motifs can affect gene regulation, transcription, and protein function (Kashi and King, 2006). The transcripts of 49 newly developed SSR markers are involved in the most important life activities of plants, such as cell signal transduction, material transport, response to abiotic stress, multiple biosynthetic pathways, and biological detoxification pathways. Given that *K. melanthera* mainly grows in high-altitude environments with insufficient water and desertification (Yen and Yang, 2020), the polymorphic SSRs located in these genes may be the result of adaptive evolution of *K. melanthera* to the environment.

Both UPGMA and PCoA analyses divided 56 *K. melanthera* accessions into three clusters, which coincided with their geographical origin, indicating that there was a high degree of genetic variation among geo-populations of *K. melanthera*. This may be due to the geographical isolation and natural selection in different populations (Kashi and King, 2006). Nevertheless, the subpopulations of wild accessions collected from Waqie and Dazhasi town were not distinguished. This is partly attributed to the high level of gene flow caused by close geographical distance between the subpopulations.

Of the 22 newly developed markers, 21 were successfully amplified in the other four *Kengyilia* species, and the transferability ratio was as high as 95%, indicating that the newly developed markers had good transferability. The SSR markers developed in this study showed a high level of transferability among the related species. This phenomenon was also found in many other species, such as *E. sibiricus* (Zhang et al., 2019), *Agropyron cristatum* (Ren et al., 2016), and *E. excelsus* (Xiong et al., 2019). In terms of geographical distribution, *K. alatavica* and *K. Batalinii* were all from Central Asia, while *K. Mutica*, *K. rigidula*, and *K. melanthera* were mainly distributed in the QTP (Yen and Yang, 2020), which was consistent with the results of clustering. Therefore, the novel markers are reliable and have wide application value in other *Kengyilia* species.

Conclusion

Here, full-length transcriptome sequencing was performed in *K. melanthera* for the first time, and 126,410 non-redundant transcripts were annotated in multiple databases. In the absence of *K. melanthera* genome-wide information, these full-length transcriptome data will provide great help for future related research. We identified 42,433 SSR loci from the transcriptome and designed 108,399 primer pairs. In addition, the transcripts containing SSR was associated with some important biological processes. The 49 SSR markers obtained by screening showed good polymorphism, and some of the SSR markers had good transferability, which provided a basis for the genetic research in *K. melanthera* and related species.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA735213.

Author contributions

YLX and JY conceptualized the basic idea and plan the study, and contributed to writing—reviewing and editing. YX, JMZ, and LS helped in data collection and analyses. LL, WL, and YLX performed the statistical analyses. JQZ, MHY, and DXL helped in primary draft preparation. XL, XM, and SQB contributed to supervision. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (3177131276), the Key Laboratory of Superior Forage Germplasm in the Qinghai-Tibetan plateau (2020-ZJ-Y03), Sichuan Provincial Key Science and Technology Program (2019YFN0170), and Biological Genetic Resources Management Project of National Forestry and Grassland Administration (KJZXSA202102). These funding bodies did not conduct the design of the study and collection, analysis, and interpretation of data and writing the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.959042/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Images of *K. melanthera* with whole plant **(A)** and spike **(B)**.

SUPPLEMENTARY FIGURE 2
Functional classification of *K. melanthera* transcripts. The predicted functions were based on COG **(a)**, eggNOG **(b)**, and KOG **(c)**.

SUPPLEMENTARY FIGURE 3
Distributions of the major repeat types in the *K. melanthera* transcriptome.

SUPPLEMENTARY FIGURE 4
Classification of transcripts containing SSR loci based on GO **(a)** and KEGG **(b)** databases.

SUPPLEMENTARY FIGURE 5
The gel figure amplified by Km-eSSR231. The left band means 50 bp marker. The samples 20 and 43 were discarded for further analysis.

SUPPLEMENTARY FIGURE 6
Estimation of number of *K* value for 56 *K. melanthera* individuals.

SUPPLEMENTARY FIGURE 7
The UPGMA dendrogram of five *Kengyilia* species was studied based on 22 selected SSR markers.

SUPPLEMENTARY TABLE 1
Kengyilia melanthera accessions for SSR marker polymorphism validation and population analysis.

SUPPLEMENTARY TABLE 2
Accessions used to verify cross-species transferability of SSR markers.

SUPPLEMENTARY TABLE 3
Gene annotations of 118,341 transcripts in *K. melanthera*.

SUPPLEMENTARY TABLE 4
Transcription factor prediction in the transcriptome of *K. melanthera*.

SUPPLEMENTARY TABLE 5
Characterization of primers developed based on 36,133 SSR loci. LncRNAs were marked with red.

SUPPLEMENTARY TABLE 6
Summary of 49 selected SSR markers with high polymorphism.

SUPPLEMENTARY TABLE 7
The diversity parameters of 49 SSR markers amplified in 56 *K. melanthera* individuals.

SUPPLEMENTARY TABLE 8
The transferability of 22 selected SSR markers was validated in other four *Kengyilia* species.

References

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Webb, M., et al. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Bing, C. L. (2001). Geographical distribution of *Kengyilia* Yen et J. L. Yang (Poaceae). *J. Syst. Evol.* 39, 248–259.

Chen, S., Huang, X., Yan, X., Liang, Y., Wang, Y., and Li, X. (2013). Transcriptome analysis in sheepgrass (*Leymus chinensis*): a dominant perennial grass of the Eurasian steppe. *PLoS One* 8:e67974. doi: 10.1371/journal.pone.0067974

Deng, Y., Li, J., Wu, S., Zhu, Y., Chen, Y., and He, F. (2006). Integrated nr database in protein annotation system and its localization. *Comput. Eng.* 32, 71–72. doi: 10.1109/INFOCOM.2006.241

Earl, D. A., and Vonholdt, B. M. (2012). Structure harvester: a website and program for visualizing structure output and implementing the evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* 7, 574–578. doi: 10.1111/j.1471-8286.2007.01758.x

Fan, X., Sha, L. N., Zeng, J., Kang, H. Y., Zhang, H. Q., Wang, X. L., et al. (2012). Evolutionary dynamics of the Pgk1 gene in the polyploid genus *Kengyilia* (Triticeae: Poaceae) and its diploid relatives. *PLoS One* 7:e31122. doi: 10.1371/journal.pone.0031122

Finn, R. D., Alex, B., Jody, C., Penelope, C., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, 222–230. doi: 10.1093/nar/gkt1223

Gao, G., Tang, Z. L., Deng, J. B., Gou, X. M., Wang, Q., Zhang, Y., et al. (2016). Phylogenetic relationships and Y genome origin in *Kengyilia* (Triticeae: Poaceae) based on single copy gene DMC1. *Biologia* 71, 298–304. doi: 10.1515/biolog-2016-0041

Gao, G., Tang, Z. L., Wang, Q., Gou, X. M., Ding, C. B., Zhang, L., et al. (2014). Phylogeny and maternal donor of *Kengyilia* (Triticeae: Poaceae) based on chloroplast trnT–trnL sequences. *Biochem. Syst. Ecol.* 57, 102–107. doi: 10.1016/j.bse.2014.07.004

Ge, Y., Zang, X., Tan, L., Wang, J. S., Liu, Y. Z., Li, Y. X., et al. (2019). Single-molecule long-read sequencing of *Avocado* generates microsatellite markers for analyzing the genetic diversity in *Avocado germplasm*. *Agronomy* 9, 512–530. doi: 10.3390/agronomy9090512

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., et al. (2015). Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* 10:e0132628. doi: 10.1371/journal.pone.0132628

Hampl, V., Pavlicek, A., and Flegr, J. (2001). Construction and bootstrap analysis of DNA fingerprinting-based phylogenetic trees with the freeware program FreeTree: application to trichomonad parasites. *Int. J. Syst. Evol. Microbiol.* 51, 731–735. doi: 10.1099/00207713-51-3-731

Karcı, H., Paizila, A., Topçu, H., Ilikçioğlu, E., and Kafkas, S. (2020). Transcriptome sequencing and development of novel genic SSR markers from *Pistacia vera* L. *Front. Genet.* 11:1021. doi: 10.3389/fgene.2020.01021

Kashi, Y., and King, D. G. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* 22, 253–259. doi: 10.1016/j.tig.2006.03.005

Kong, L., Yong, Z., Ye, Z. Q., Liu, X. Q., and Ge, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, 345–349. doi: 10.1093/nar/gkm391

Koonin, E. V., Fedorova, N. D., Jackson, J. D., and Jacobs, A. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* 5, 1–28. doi: 10.1186/gb-2004-5-2-r7

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Li, Y. C., Korol, A. B., Tzion, F., and Eviatar, N. (2004). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 21, 991–1007. doi: 10.1093/molbev/msh073

Luo, X. M., Tinker, N. A., Fan, X., Zhang, H. Q., Sha, L. N., Kang, H. Y., et al. (2012). Phylogeny and maternal donor of *Kengyilia* species (Poaceae: Triticeae) based on three cpDNA (matK, rbcL and trnH-psbA) sequences. *Biochem. Syst. Ecol.* 44, 61–69. doi: 10.1016/j.bse.2012.04.004

Merritt, B. J., Culley, T. M., Avanesyan, A., Stokes, R., and Brzyski, J. (2015). An empirical review: characteristics of plant microsatellite markers that confer higher levels of genetic variation. *Appl. Plant Sci.* 3:1500025. doi: 10.3732/apps.1500025

Metzgar, D., Bytof, J., and Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* 10, 72–80. doi: 10.1101/gr.10.1.72

Minoru, K., Susumu, G., Shuichi, K., Yasushi, O., and Masahiro, H. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, 277–280. doi: 10.1093/nar/gkh063

Pan, L., Huang, T., Yang, Z., Tang, L., Cheng, Y., and Wang, J. (2018). EST-SSR marker characterization based on RNA-sequencing of *Lolium multiflorum* and cross transferability to related species. *Mol. Breed.* 38, 1–12. doi: 10.1007/s11032-018-0775-4

Peakall, R., and Smouse, P. E. (2010). GENALEX 6: genetic analysis in excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Piriyapongsa, J., Kaewprommal, P., Vaiwsri, S., Anuntakarun, S., Wirojsirasak, W., Punpee, P., et al. (2018). Uncovering full-length transcript isoforms of sugarcane cultivar Khon Kaen 3 using single-molecule long-read sequencing. *PeerJ* 6:e5818. doi: 10.7717/peerj.5818

Ren, J. D., Li, F. Z., Xu, Y. J., Wang, X. X., and Ma, X. G. (2016). Genetic diversity of the major varieties of *Kengyilia* C. Yen et J. L. Yang in Qing Hai plateau based on EST-SSR markers. *J. Plant Genet. Resour.* 17, 663–670.

Rohlf, F. J. (1987). NTSYS-pc: microcomputer programs for numerical taxonomy and multivariate analysis. *Am. Stat.* 41, 330–330.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36. doi: 10.1093/nar/28.1.33

Teng, K., Teng, W., Wen, H., Yue, Y., Guo, W., and Wu, J. (2019). PacBio single-molecule long-read sequencing shed new light on the complexity of the *Carex breviculmis* transcriptome. *BMC Genomics* 20:789. doi: 10.1186/s12864-019-6163-6

Thomas, S., Underwood, J. G., Tseng, E., Holloway, A. K., and Bench To Basinet CvDC Informatics Subcommittee (2014). Long-read sequencing of chicken transcripts and identification of new transcript isoforms. *PLoS One* 9:e94650. doi: 10.1371/journal.pone.0094650

Tóth, G., Gáspári, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10, 967–981.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40:115. doi: 10.1093/nar/gks596

Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.* 39, 312–328. doi: 10.1590/1678-4685-GMB-2016-0027

Wang, B., Kumar, V., Olson, A., and Ware, D. (2019). Reviving the transcriptome studies: an insight into the emergence of single-molecule transcriptome sequencing. *Front. Genet.* 10:384. doi: 10.3389/fgene.2019.00384

Wang, F., Liu, L., Gao, T., Gao, L., Bao, A., and Wang, S. (2018). Research advances on transcriptomics in forage plants. *Pratacult. Sci.* 36, 402–413.

Wang, J., Chen, Z., Jin, S., Hu, Z., Huang, Y., and Diao, Y. (2017). Development and characterization of simple sequence repeat (SSR) markers based on a full-length cDNA library of Napier grass (*Pennisetum purpureum* Schum). *Genes Genomics* 39, 1297–1305. doi: 10.1016/j.ygeno.2014.10.009

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., and Li, W. (2013). CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41:74. doi: 10.1093/nar/gkt006

Wu, Q., Zang, F., Xie, X., Ma, Y., Zheng, Y., and Zang, D. (2020). Full-length transcriptome sequencing analysis and development of EST-SSR markers for the endangered species *Populus wulianensis*. *Sci. Rep.* 10, 1–11. doi: 10.1038/s41598-020-73289-5

Xie, L., Teng, K., Tan, P., Chao, Y., Li, Y., Guo, W., et al. (2020). PacBio single-molecule long-read sequencing shed new light on the transcripts and splice isoforms of the perennial ryegrass. *Mol. Genet. Genomics* 295, 475–489. doi: 10.1007/s00438-019-01635-y

Xiong, Y. L., Liu, W. H., Xiong, Y., Yu, Q. Q., Ma, X., Lei, X., et al. (2019). Revelation of genetic diversity and structure of wild *Elymus excelsus* (Poaceae: Triticeae) collection from western China by SSR markers. *PeerJ* 2019:e8038. doi: 10.7717/peerj.8038

Xu, R., Wang, Z., Su, Y., and Wang, T. (2020). Characterization and development of microsatellite markers in *Pseudotaxus chienii* (Taxaceae) based on transcriptome sequencing. *Front. Genet.* 11:1249. doi:10.3389/fgene.2020.574304

Yen, C., and Yang, J. L. (1990). *Kengyilia gobicola*, a new taxon from West China. *Can. J. Bot.* 68, 1894–1897. doi: 10.1139/b90-248

Yen, C., and Yang, J. L. (2020). *Biosystematics of Triticeae: Volume I. Triticum-Aegilops complex*. Berlin: Springer Nature.

Yi, Z., Chen, J., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhang, B., Liu, J., Wang, X., and Wei, Z. (2018). Full-length RNA sequencing reveals unique transcriptome composition in bermudagrass. *Plant Physiol.* 132, 95–103. doi: 10.1016/j.plaphy.2018.08.039

Zhang, C., Fan, X., Yu, H., Zhang, L., Wang, X., and Zhou, Y. (2009). Different maternal genome donor to *Kengyilia* species inferred from chloroplast trnL-F sequences. *Biol. Plant.* 53, 759–763. doi: 10.1007/s10535-009-0139-3

Zhang, L., Zheng, Y. L., Wei, Y. M., Liu, S. G., and Zhou, Y. H. (2005). The genetic diversity and similarities among *Kengyilia* species based on random amplified microsatellite polymorphism (RAMP). *Genet. Resour. Crop Evol.* 52, 1011–1017. doi: 10.1007/s10722-004-6087-z

Zhang, L., Zhou, Y., Ding, C., Yang, R., and Liu, S. (2008). Genetic variation of the genus *Kengyilia* by ISSR markers. *Front. Biol. China* 3:170–175.

Zhang, Z., Xie, W., Zhao, Y., Zhang, J., Wang, N., Ntakirutimana, F., et al. (2019). EST-SSR marker development based on RNA-sequencing of *E. sibiricus* and its application for phylogenetic relationships analysis of seventeen *Elymus* species. *BMC Plant Biol.* 19:235. doi: 10.1186/s12870-019-1825-8

Zhao, L., Zhang, H., Kohnen, M. V., Prasad, K. V. S. K., and Reddy, A. S. N. (2019). Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and nanopore-based direct RNA sequencing. *Front. Genet.* 10:253. doi: 10.3389/fgene.2019.00253

Zheng, Y., Zhang, Z., Wan, Y., Tian, J., and Xie, W. (2020). Development of EST-SSR markers linked to flowering candidate genes in *Elymus sibiricus* L. based on RNA sequencing. *Plants* 9:1371. doi: 10.3390/plants9101371

Zhou, Q., Luo, D., Ma, L., Xie, W., Wang, Y., and Wang, Y. (2016). Development and cross-species transferability of EST-SSR markers in Siberian wildrye (*Elymus sibiricus* L.) using Illumina sequencing. *Sci. Rep.* 6, 1–10. doi: 10.1038/srep20549

Zhou, Y. H., Zheng, Y. L., Yang, J. L., Yen, C., and Jia, J. Z. (2000). Relationships among *Kengyilia* species assessed by RAPD markers. *J. Syst. Evol.* 38, 515–521.