Arch Biol Sci. 2019;71(3):489-500 https://doi.org/10.2298/ABS181226032Z

RNA-Seq for excavation of genes involved in the biosynthesis of primary active components and identification of new EST-SSR markers in medicinal chrysanthemum

Xiting Zhao[1,2,3,*], Lingyu Song[1], Mengdan Ma[1], Ke Liu[1], Meng Lv[1], Yuting Zhu[1] and Liwei Jiang[1,#]

[1]*College of Life Sciences, Henan Normal University*, Xinxiang 453007, China

[2]*Engineering Technology Research Center of Nursing and Utilization of Genuine Chinese Crude Drugs in Henan Province*, Xinxiang 453007, China

[3]*Engineering Laboratory of Biotechnology for Green Medicinal Plant of Henan Province*, Xinxiang 453007, China

Corresponding authors: *zhaoxt0411@126.com; #041121@htu.cn

Received: December 26, 2018; **Revised**: April 19, 2019; **Accepted**: May 17, 2019; **Published online**: May 29, 2019

Abstract: *Chrysanthemum morifolium* cv. 'Huaihuang' has great medicinal and commercial value. However, limited genomic and transcriptional information restricts the further research of 'Huaihuang'. In this study, using 'Huaihuang' leaf as the sample, we performed RNA-Seq and obtained numerous unigene sequences for functional characterization, and developed new EST-SSR markers. The results show that 176 unigenes were assigned to the 'antioxidant activity' in GO annotations, 998 unigenes were assigned to 'Secondary metabolite biosynthesis, transport and catabolism' in the KOG database, 381 unigenes were assigned to 'Biosynthesis of other secondary metabolites', and 377 unigenes were assigned to 'Metabolism of terpenoids and polyketides' in the KEGG database. A number of genes involved in the biosynthesis of flavones and terpenes, such as *CHS*, *CHI*, *FLS*, *HMGR* and *MVD*, were found in the transcription data of 'Huaihuang'. Also, 36 new polymorphic EST-SSR markers were developed and validated. Using them, the genetic diversity was analyzed and a dendrogram of six clades was constructed among 15 medicinal chrysanthemum varieties. The excavated unigenes and identified new EST-SSR markers will provide useful resources for the study of biosynthesis of active components, genetic diversity analysis, population structure identification and in molecular breeding research of 'Huaihuang' and related species.

Keywords: *Chrysanthemum morifolium* cv. 'Huaihuang'; medicinal chrysanthemum; RNA-Seq; functionally related genes; EST-SSR markers

INTRODUCTION

Chrysanthemum morifolium Ramat. is a perennial herb and flowering plant in the Asteraceae family, and globally, it is the second most commercially valuable ornamental species after rose [1,2]. The plant originates from Asia and northeastern Europe, most species originate are from East Asia, and the center of diversity is in China [3]. Medicinal chrysanthemum 'Huaihuang' cultivated in the Jiaozuo area of Henan Province has a long history and is a well-known medicinal plant of China. 'Huaihuang' not only has aesthetically pleasing traits, including beautiful flowers and a pleasant smell, it also contains multiple active components and trace elements, such as flavonoids, organic acids, volatile oils, amino acids and carbohydrates, and it is thought to "improve eyesight, purge liver toxins and dispel heat" [4]. However, the genetic background, evolutionary origin and transcriptional information of 'Huaihuang' are still unknown.

There is a long history of the use of active components from plants in the treatment of diseases. Flavonoids, as a type of active component, have been widely studied [5] and possess a variety of presumed health-promoting functions, such as antioxidative, antiinflammatory, anticancer and antiviral activities, prevention of osteoporosis, inhibition of alcohol addiction disorder, and protection of the cardiovascular system and the liver [6-9]. Terpenoids are also important active components in Chinese traditional herbs. For example, catalpol from *Rehmannia glutinosa* has a neuroprotective effect [10], andrographolide is a potential cancer therapeutic agent [11], and artemisinin in *Artemisia carvifolia* has a rapid-

acting effect on falciparum malaria [12]. Therefore, it is of great significance to excavate the genes involved in the biosynthesis of flavonoids and terpenes in 'Huaihuang' for further understanding of the mechanism of synthesis of these classes of compounds.

Transcriptome deep sequencing not only provides comprehensive information of the plant genome, but also contributes to the development of numerous unigene-based simple sequences repeat (SSR) markers [13,14]. Recent progress in RNA-seq technology and bioinformatics has opened up a new pathway for genetic research [15,16]. Compared to other molecular markers such as ISSR, RAPD, AFLP and SNP, SSR markers are more suitable for use in genetic linkage analysis, molecular breeding, genetic diversity analysis, and in the determining the location of target genes [17,18]. SSR markers are identified from entire genomic sequences while expressed sequence tags (EST) are derived from coding sequences. EST-SSR markers can be directly used to obtain the gene expression information of an organ or tissue at a certain period [19,20]. Furthermore, EST-SSR markers are characterized by abundant information, co-dominance and site-specificity and can effectively excavate genes [21]. At present, EST-SSR markers have been acquired and utilized in *Neottopteris nidus* [22], *Amentotaxus* [23], *Amorphophallus* [24] and *Apium graveolens* L. [25]. As regards chrysanthemum, available EST-SSR markers based on the transcriptome are rare. Two species of chrysanthemum, *C. nankingense* [26] and *C. lavandulifolium* [27], have been sequenced using Illumina HiSeq™ 2000. Twenty new EST-SSR markers were identified only in ornamental *C. nankingense*. Considering the existence of multiple varieties and the complicated genetic background of chrysanthemum, the available 20 EST-SSR markers are not enough. Moreover, EST-SSR markers for medicinal chrysanthemum have not been identified yet. Therefore, to enrich the resource of available EST-SSR markers in chrysanthemum, the development of transcriptome deep sequencing and EST-SSR markers in medicinal chrysanthemum 'Huaihuang' is essential. The transcriptome information would provide valuable information for exploring the key biosynthetic genes of primary active components in chrysanthemum, and EST-SSR markers obtained by RNA-seq would provide references for later studies and the construction of chrysanthemum fingerprint maps, gene mapping and molecular breeding.

In the present study, the transcriptome data of the medicinal chrysanthemum 'Huaihuang' were obtained using the Illumina HiSeq4000 paired-end sequencing platform. In addition, we developed new EST-SSR markers (distinct from those of *C. nankingense*) from transcriptome sequencing of 'Huaihuang', which were used to assess the genetic diversity and population structure of 15 medicinal chrysanthemum varieties.

MATERIALS AND METHODS

Plant material

The medicinal chrysanthemum cultivar 'Huaihuang' and 14 other medicinal chrysanthemum individuals (Supplementary Table S1) were used in this study. The 15 varieties are from the Germplasm Resource Nursery (Wen County Institute of Agricultural Science, China) of the Engineering Technology Research Center of Nursing and Utilization of Genuine Chinese Crude Drugs in Henan Province, Henan Normal University, China. Plants were grown in a 1:1 mixture of peat and vermiculite in greenhouses. The cultured conditions were a 10 h light/14 h dark light cycle, 18-22°C, 40 to 60% relative humidity, and 60 $mE \cdot s^{-1} \cdot m^{-2}$ light intensity. 'Huaihuang' was cultured to obtain 8-10 leaves. Leaves 3-5 below the top were removed from the plants and then rapidly placed in liquid nitrogen and stored at -80°C until RNA extraction. The Novogene Biological Information Technology Company conducted the construction of cDNA libraries and deep sequencing of the transcriptome. RNA extraction and transcriptome sequencing were performed on three biological replicates of leaves.

RNA extraction

Total RNA was extracted from leaves using the MiniBEST Plant RNA Extraction Kit (TaKaRa, DaLian, China). Agarose gels (1%) were used to detect contamination or degradation of RNA. RNA purity was checked using a NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). RNA concentration was measured using a Qubit® RNA Assay Kit in a Qubit® 2.0 fluorometer (Life Technologies, CA, USA). The quality of the three RNA samples (the RNA integrity number (RIN) ≥6.5, 28S:18S>1.5) was assessed using the RNA

Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 (Agilent Technologies, CA, USA). Total RNA with appropriate quality was subsequently used in cDNA library construction and Illumina deep sequencing.

cDNA library construction and sequencing

Using 1.5 μg RNA per sample, the cDNA libraries were constructed using NEBNext® Ultra™ RNA Library Prep Kit, Illumina® (NEB, USA) according to the procedure described by Zhong et al. [28]. Three cDNA libraries were sequenced on an Illumina HiSeq4000 platform and paired-end reads were generated. All reads generated in this research are visible from the NCBI Sequence Read Archive database [http://www.ncbi.nlm.nih.gov/sra/] under the project accession number SRP137033.

Data filtering and *de novo* assembly

Following CASAVA base recognition, the resulting image data were transformed into raw reads, and the raw reads were stored in FASTQ format [29]. Adapters, poly-Ns, and low-quality reads were removed to obtain clean reads by the program SeqPrep (https://github.com/jstjohn/SeqPrep). In addition, the resulting clean data were *de novo* assembled by the short read assembly program Trinity [30]. Then, high quality clean data were used to analyze the functional annotation of unigenes.

Functional annotation of unigenes

We compared the unigenes assembled with the known public databases NCBI Nr (1e-5), euKaryotic Orthologous Groups (KOG) (1e-5), NT (1e-5), Swissprot (1e-5), gene ontology (GO) (1e-6), Kyoto Encyclopedia of Genes and Genomes (KEGG) (1e-10), and Pfam (1e-5) to obtain unigene function and classification. GO annotation information of these unigenes was obtained from the NCBI Nr and Pfam database using Blast2GO v2.5 [31]. Additionally, these unigenes were classified and analyzed using the KOG database based on eukaryote gene homology. The function of the gene products and compounds and their metabolic pathways involved in cells were analyzed by the KEGG database using KAAS (KEGG Automatic Annotation Server) [32]. The supplementary annotation files as well as the Trinity assembled transcripts were uploaded to Zenodo (DOI: 10.5281/zenodo.2579584).

Detection of new EST-SSR markers and primer design

Using MISA software (http://pgrc.ipk-gatersleben.de/misa/misa.html), potential SSRs were found. The mono-, di-, tri-, tetra-, penta- and hexanucleotides with minimum repeat numbers of 10, 6, 5, 5, 5 and 5, respectively, were picked out. The selection criteria of primer pairs were as follows: primer length 18-24 bp, Tm of 50-62°C, amplicon length in the range 100-400 bp. Specific primers were designed by Primer 3 software (http://sourceforge.net/projects/primer3). The primers as well as Tm used for the 36 EST-SSR markers are shown in Supplementary Table S4. These primers were synthesized by GENEWIZ (SuZhou, China).

Development and validation of new EST-SSR markers

In order to develop new EST-SSRs, a total of 15 varieties of medicinal chrysanthemum were selected for analysis of polymorphism. The young developing leaves of each variety were used to extract their whole genomic DNAs using the modified sodium dodecyl sulfate (SDS) method [33]. The extracted genomic DNA was diluted to 50 ng·μL^{-1}. PCR reactions were conducted in a 20 μL reaction volume containing DNA (2 μL), each primer (0.5 μL), 2×Taq Master Mix (CWBIO) (10 μL) and double-distilled water (7 μL). The following conditions were used in each reaction: pre-denaturing for 5 min at 94°C, followed by 35 cycles of 30 s of denaturation at 94°C, 30 s of annealing at the Tm of specific primers, and 30 s of extension at 72°C, and lastly, by extension for 10 min at 72°C. Amplified PCR products were separated on an 8% non-denaturing polyacrylamide gels electrophoresis (PAGE) using a vertical electrophoresis device. Detection of EST-SSR bands was performed by silver staining [34].

The number of alleles (Na), expected heterozygosity (He), observed heterozygosity (Ho), genetic differentiation coefficient (Fst), Shannon's information index (I), gene flow (Nm), Nei's genetic identity and genetic distance of all EST-SSR loci were calculated using POPgene ver. 1.32 [35], and the polymorphism

information content (PIC) values were determined using PIC_CALC version 0.6 [36]. We analyzed genetic relationships and principal coordinate analysis (PCoA) between varieties using the NTSYS-pc 2.1 software package. A dendrogram was constructed also using the NTSYS-pc 2.1 software package by the unweighted pair group method of arithmetic average algorithm (UPGMA; Exeter Software, Setauket, NY, USA).

RESULTS

Sequencing and *de novo* assembly

As shown in Table 1, the sequencing error rate was low and the result satisfied the quality requirements of the subsequent assembly analysis. We identified 162,252 (46.25% of transcripts) unigenes with a total of 113,777,555 bp, and the average unigene length was 701 bp. Of the 162,252 unigenes, the length of 55,439 unigenes (34.17%) was within 1-300 bp; the length of 97,706 unigenes (60.22%) was within 301-2000 bp; and the length of 9106 unigenes (5.61%) was more than 2000 bp (Fig. 1A).

Table 1. Transcriptome data and assembled EST-SSRs for medicinal chrysanthemum 'Huaihuang'.

Category	Items	Number
Raw reads	Total raw read	59,556,729
Clean reads	Total clean reads	57,566,460
	Total clean nucleotides (G)	7.20
	Q20 percentage (%)	95.16
	Q30 percentage (%)	90.57
	GC percentage (%)	42.80
	Error (%)	0.03
Unigenes	Total sequence number	162,252
	Total sequence base (bp)	113,777,555
	MaxLength (bp)	106,811
	MinLength (bp)	201
	Mean Length (bp)	701
	N50 (bp)	1093
	N90 (bp)	281
EST-SSR	Total number of identified SSRs	16,581
	Number of SSR-containing sequences	14,705
	Number of sequences containing more than one SSR	1624
	Number of SSRs present in compound formation	657

Functional annotation and classification of unigenes

Of the 162,252 unigenes, 57,890 (35.67%) were successfully annotated in 7 databases. The success rate in Nr database (30.07%) was the highest (Supplementary Table S2). Functional analysis of 'Huaihuang' helps to understand the function of genes and gene expression regulation.

In the Nr database, among the annotated 48,792, 19,078 (39.1%) unigenes of the sequences, 80% similarity was observed (Fig. 1B). Significant homology (<1e-15) was found between 37,960 (77.8%) unigenes of the control sequences with entries in the Nr database (Fig. 1C). Based on the Nr data base, the species distribution of annotated unigenes was mapped and the highest proportion of species was *Vitis vinifera* (11.7%), followed by *Coffea canephora* (7.8%), *Nicotiana tomentosiformis* (7.3%), *Theobroma cacao* (4.7%) and *Citrus sinensis* (3.9%) (Fig. 1D).

A total of 37,239 (22.95%) unigenes were assigned GO annotations and grouped into 'Biological process', 'Cellular component' and 'Molecular function', and further subdivided into 56 subclasses. 'Biological process' included 23 subgroups: of them, 'cellular process' with 20,485 unigenes was the largest, the second was 'metabolic process' with 19,868 unigenes, and the third was 'single-organism process' with 15,024 unigenes. The top three abundant subcategories of 'Cellular component' were 'cell', 'cell part' and 'macromolecular complex', respectively. The major proportions of the 'Molecular function' category belonged to 'binding' and 'catalytic activity' (Fig. 2A).

A total of 16,034 (9.88%) unigenes were assigned into 25 KOG categories. Of these classifications, the most frequently identified classes were 'General function prediction only' (3037 unigenes), followed by 'Posttranslational modification, protein turnover, chaperones' (2167 unigenes), and 'Signal transduction mechanisms' (1394 unigenes) (Fig. 2B).

The first level of the KEGG pathway for metabolism pathways is divided into five branches: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Metabolism, and Organismal Systems. These five branches were divided into 32 pathways at the second level. Of these pathways, the most frequently identified were 'Signal transduction' (1673 unigenes were found), followed

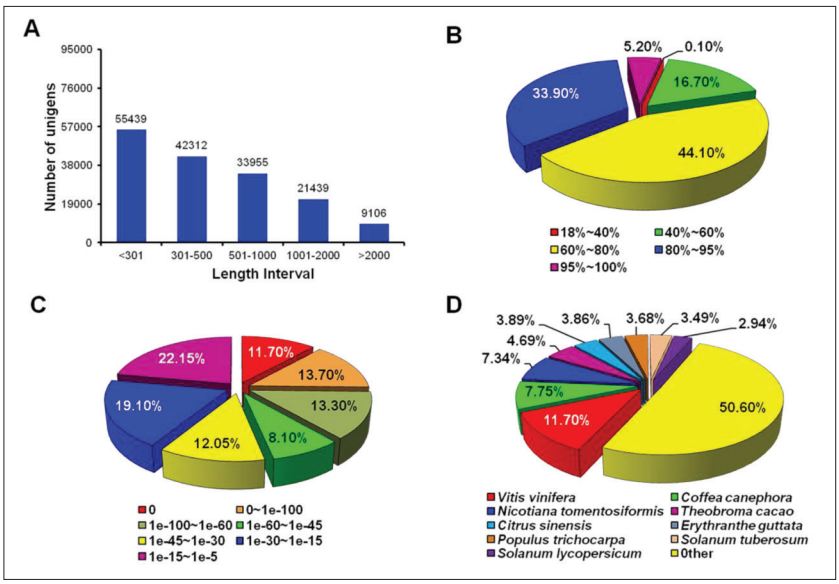

Fig. 1. Transcriptome unigenes for medicinal chrysanthemum 'Huaihuang'. **A** – Distribution of unigene length. **B** – Comparison of similarity distributions. **C** – E-value distribution map. **D** – Classification map of species distribution.

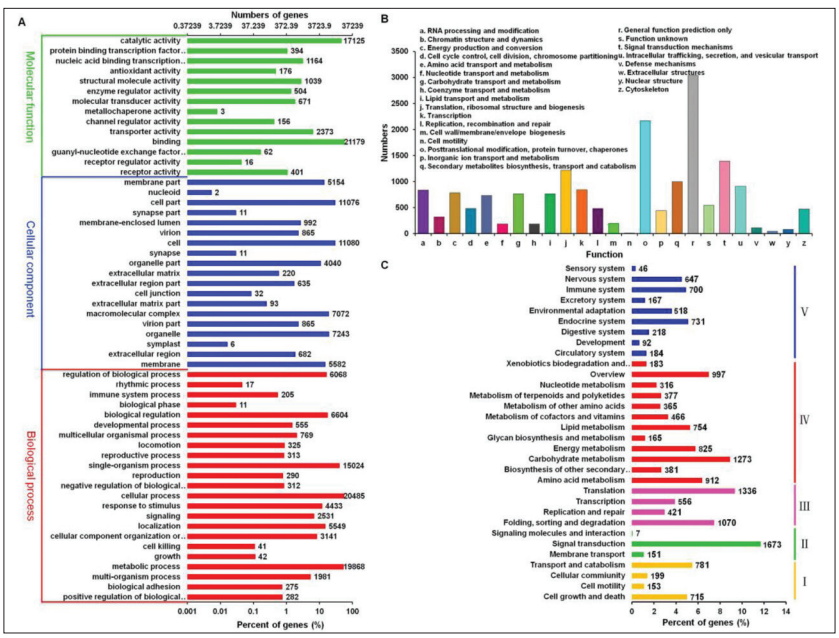

Fig. 2. Annotation of transcriptome unigenes for medicinal chrysanthemum 'Huaihuang'. **A** – Summary of GO analysis of the unigene sequences in medicinal chrysanthemum 'Huaihuang'. **B** – KOG classification of the unigene sequences in medicinal chrysanthemum 'Huaihuang'. The y-axis shows the number of annotated unigenes in a functional category; the x-axis shows functional of cluster. **C** – KEGG metabolic pathway of medicinal chrysanthemum 'Huaihuang'. The x-axis above shows unigene number in one classification. The x-axis below the graph shows the unigene percentage in the main classification.

by 'Translation' (1336 unigenes), 'Carbohydrate metabolism' (1273 unigenes), and 'Folding, sorting and degradation' (1070 unigenes) (Fig. 2C). At the third level, 14,356 unigenes were annotated into 276 KEGG pathways in a more specific classification. Among them, the highest number of unigenes was observed in 'Carbon metabolism' (537), followed by 'Ribosome' (510), and 'Biosynthesis of amino acids' (490) (Supplementary Table S3).

Genes involved in biosynthesis of primary active components

Since 'Huaihuang' has many health promoting functions, we paid more attention to its medicinal value and hence the biosynthesis of primary active components. Therefore, we further excavated some genes related to the biosynthesis of primary active components. Among the 'Molecular function' category of GO annotations, 176 unigenes were assigned to 'antioxidant activity'. For KOG categories, the group of 'Secondary metabolite biosynthesis, transport and catabolism' (998 unigenes) was fifth among the frequently identified classes in 'Huaihuang'. As for the KEGG pathway, at the second level, 381 unigenes were mapped to 'Biosynthesis of other secondary metabolites', and 377 unigenes were mapped to 'Metabolism of terpenoids and polyketides'. At the third level, 241 genes were annotated in 'Phenylpropanoid biosynthesis'. We delved deeper into the transcriptome data of 'Huaihuang', and identified groups of genes associated with the biosynthesis of flavones and terpenes, such as chalcone synthase (CHS), chalcone isomerase (CHI),

Table 2. Repeat units of the EST-SSRs in medicinal chrysanthemum 'Huaihuang'.

Number of repeats	Mono-	Di-	Tri-	Tetra-	Penta-	Hexa-	Total	Percentage (%)
5			1958	143	17	8	2126	12.82
6		1447	747	16	2	3	2215	13.35
7		600	201			1	802	4.83
8		324	13			1	338	2.03
9		171					171	1.03
10	6251	86					**6337**	**38.21**
11	2198	39					2237	13.49
12	870						870	5.24
13	443						443	2.67
14	277						277	1.67
15	181						181	1.09
16	117						117	0.71
17	109						109	0.66
18	83						83	0.50
19	82						82	0.49
20	97						91	0.55
21	64						64	0.39
22	19						19	0.11
23	16						16	0.10
24	3						3	0.02
Total	10,804	2,667	2,919	159	19	13	16,581	
Percentage (%)	65.16	16.09	18.04	0.96	0.11	0.08		

The most abundant EST-SSR repeat unit was the ten tandem repeat (6337, 38.21%, in bold).

flavanone 3-hydroxylase (F3H), flavonol synthase (FLS), anthocyanidin synthase (ANS) and dihydroflavonol 4-reductase (DFR), MYB and MYC transcription factors, 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGR) and mevalonate decarboxylase (MVD).

Frequency and distribution of EST-SSR markers

For 'Huaihuang', 16,581 potential EST-SSRs were found in 14,705 unigenes, with 1624 unigenes containing more than one EST-SSR locus. The frequency of SSRs was 10.22%, with an average of one EST-SSR every 6.86 kbp (Table 1). As can be seen in Table 2, the most abundant was the ten tandem repeat of the EST-SSR (6337, 38.21%). These identified EST-SSR markers significantly enriched the resources of available SSR markers for the development in 'Huaihuang' and related species.

As di- and trinucleotide repeats are the key for developing EST-SSR markers [26,37], we further analyzed them. Of the dinucleotide repeats, AC/GT was the most plentiful motif (1222, 45.82%), followed by AT/AT (743, 27.86%), AG/CT (696, 26.10%), and CG/CG (6, 0.22%). Of the trinucleotide repeats, the predominant repeat motif was ATC/ATG (643, 22.03%), followed by AAC/GTT (581, 19.90%), ACC/GGT (570, 19.46%), and AAG/CTT (411, 14.08%) (Table 3).

Table 3. Different repeat motifs in the medicinal chrysanthemum 'Huaihuang' transcriptome.

Repeat type	Repeat motif	Number	Proportion (%)
Dinucleotide	AC/GT	1,222	45.82
	AG/CT	696	26.10
	AT/AT	743	27.86
	CG/CG	6	0.22
Trinucleotide	AAC/GTT	581	19.90
	AAG/CTT	411	14.08
	AAT/ATT	382	13.08
	ACC/GGT	570	19.46
	ACG/CGT	13	0.45
	ACT/AGT	67	2.30
	AGC/CTG	134	4.60
	AGG/CCT	95	3.25
	ATC/ATG	643	22.03
	CCG/CGG	23	0.79

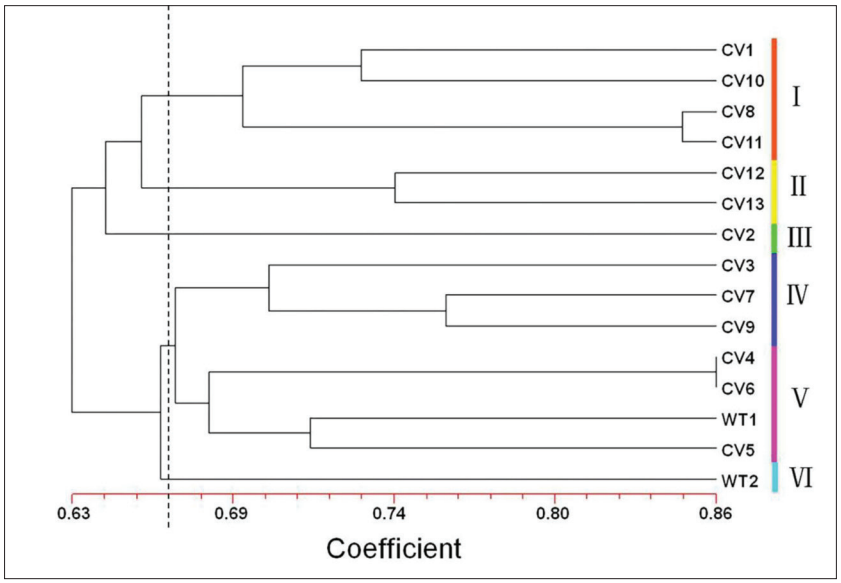

Fig. 3. Dendrogram of different medicinal chrysanthemum varieties. The tree was derived from genotype data from 36 EST-SSR loci. (I-VI) six clades were recognized. The details of 15 medicinal chrysanthemum varieties are shown in Supplementary Table S1.

Development and validation of new polymorphic EST-SSR markers

In this study, according to the screening conditions, 121 primer pairs were selected and synthesized. Among them, 57 primer pairs amplified the expected size of the products. To further select polymorphic primer pairs, we performed PCR using DNA from 15 medicinal chrysanthemum individuals as templates, followed by 8% non-denaturing PAGE. The results showed that 36 primer pairs were polymorphic. Moreover, the 36 primer pairs are new and totally different from what has been reported [26]. These new EST-SSR markers will greatly enrich the resources of chrysanthemum molecular markers, and will be applicable as an effective tool for genetic diversity and evolutionary adaptation analysis among diverse chrysanthemum varieties.

The 36 new primer pairs were used to assess the genetic diversity of 15 medicinal chrysanthemum individuals. As can be seen in Table 4, 169 alleles were detected. The number of alleles at each locus was between 2 and 8 with an average of 4.6944. The value of Ho ranged from 0.0667 to 1.0000 with an average of 0.6352, and He was between 0.1862 and 0.8713 with a mean of 0.6439. The PIC ranged from 0.1638 to 0.8231, with a mean of 0.5801 (>0.5), which is a moderate level (Table 4). These results indicated that the new 36 primer pairs all had high polymorphism and could be used for the analysis of genetic diversity. Fst ranged from 0.2105 to 0.8148, with a mean of 0.4897. This indicated that 48.97% of the total variation occurred between different chrysanthemum varieties and 51.03% between different individual plants of the same cultivars. The high degree of genetic differentiation between populations (Fst >0.25) may be due to the long-term effects of positive artificial selection on populations. The average Nm was 0.2651 (<1), which indicated low gene exchange among 15 varieties of medicinal chrysanthemum, and showed differentiation of the population could be due to genetic drift. This may be due to the variety of chrysanthemum species in China, introduction disorders and other factors prone to genetic drift. To analyze the degree of genetic differentiation among different varieties, Nei's genetic identity and genetic distance of these different varieties were calculated (Table 5). As shown in Table 5, Nei's genetic identity of these different varieties ranged from 0.3178 to 1.1170, and the genetic distance ranged from 0.1866 to 1.1462. It is clear that the genetic background of chrysanthemum germplasm resources is complex. The genetic distance between *C. morifolium* cv. 'Huaibai' and other varieties was large (and the same as *C. morifolium* cv. 'Huaihuang').

Based on the 36 new EST-SSR markers polymorphisms, the phylogenetic relationships of 15 medicinal chrysanthemum varieties were analyzed. From the dendrogram (Fig 3), these 15 medicinal chrysanthemum varieties can be clearly distinguished, indicating

Table 4. Characteristics of 36 new EST-SSR markers in medicinal chrysanthemum 'Huaihuang'.

Primer	Na	Ne	Ho	He	PIC	Fst	I	Nm
Cm-2	2	1.2195	0.0667	0.1862	0.1638	0.8148	0.3251	0.0568
Cm-3	4	3.8136	0.7333	0.7632	0.6890	0.5030	1.3606	0.2470
Cm-6	2	1.8672	0.7333	0.4805	0.3566	0.2105	0.6572	0.9375
Cm-7	4	1.9149	0.6000	0.4943	0.4456	0.3721	0.9291	0.4219
Cm-15	2	1.3006	0.2667	0.2391	0.2044	0.4231	0.3927	0.3409
Cm-17	4	3.5714	0.6667	0.7448	0.6660	0.5370	1.3138	0.2155
Cm-20	5	3.1469	0.6667	0.7057	0.6452	0.5114	1.3630	0.2389
Cm-21	5	4.7368	0.9333	0.8161	0.7550	0.4085	1.5812	0.3621
Cm-24	4	2.2277	0.2667	0.5701	0.5115	0.7581	1.0484	0.0798
Cm-27	6	4.7872	0.8667	0.8184	0.7625	0.4522	1.6775	0.3028
Cm-31	4	3.1034	0.6667	0.7011	0.6260	0.5082	1.2474	0.2419
Cm-35	5	3.5156	0.8667	0.7402	0.6638	0.3944	1.3625	0.3839
Cm-41	4	3.3835	0.8667	0.7287	0.6563	0.3849	1.3048	0.3996
Cm-48	3	1.6245	0.4667	0.3977	0.3514	0.3931	0.7026	0.3860
Cm-49	8	6.3308	0.6667	0.8713	0.8231	0.6042	1.9509	0.1638
Cm-51	4	3.0612	0.8000	0.6966	0.6146	0.4059	1.2275	0.3659
Cm-59	6	4.1284	0.6667	0.7839	0.7227	0.5601	1.5803	0.1963
Cm-60	5	1.7928	04000	0.4575	0.4205	0.5477	0.9322	0.2064
Cm-61	6	4.0541	0.7333	0.7793	0.7129	0.5133	1.5327	0.2371
Cm-63	7	6.2500	1.0000	0.8690	0.8197	0.4048	1.8849	0.3676
Cm-64	3	2.6316	0.4000	0.6414	0.5419	0.6774	1.0222	0.1190
Cm-67	5	3.6585	0.6000	0.7517	0.6896	0.5872	1.4518	0.1758
Cm-71	3	2.7778	0.4667	0.6621	0.5632	0.6354	1.0549	0.1434
Cm-72	8	4.8387	0.8000	0.8207	0.7635	0.4958	1.7448	0.2542
Cm-81	8	2.3684	0.6000	0.5977	0.5605	0.4808	1.3515	0.2700
Cm-84	7	4.8387	0.5333	0.8207	0.7657	0.6639	1.7298	0.1266
Cm-85	6	4.6875	1.0000	0.8138	0.7542	0.3644	1.6349	0.4360
Cm-90	5	3.4884	0.9333	0.7379	0.6654	0.3458	1.3784	0.4730
Cm-91	5	2.8662	0.8667	0.6736	0.5895	0.3345	1.2173	0.4974
Cm-95	3	1.8672	0.6000	0.4805	0.4188	0.3541	0.8170	0.4561
Cm-96	5	2.7950	0.7333	0.6644	0.5975	0.4291	1.2497	0.3327
Cm-99	4	1.3196	0.1333	0.2506	0.2321	0.7248	0.5313	0.0949
Cm-102	4	2.9032	0.7333	0.6782	0.5906	0.4407	1.1794	0.3173
Cm-104	7	3.8462	0.9333	0.7655	0.7034	0.3694	1.5755	0.4268
Cm-108	3	1.7176	0.2667	0.4322	0.3697	0.6809	0.7299	0.1172
Cm-121	3	2.1127	0.3333	0.5448	0.4668	0.6835	0.8975	0.1157
Mean	4.6944	3.1821	0.6352	0.6439	0.5801	0.4897	1.2206	0.2605

Detailed information of these primers is shown in Supplementary Table S4.
Na – observed number of alleles; Ne – effective number of alleles; He – expected heterozygosity; Ho – observed heterozygosity; PIC – polymorphic information content; Fst – genetic differentiation coefficient; I – Shannon's information index; Nm – gene flow.

that 36 new EST-SSR markers can serve as the basis of molecular identification of their species. The results showed that the phylogenetic tree had six clades, namely I, II, III, IV, V and VL. What's more, PCoA analysis with the 36 new EST-SSR markers showed differences in phylogenetic relationship among the 15 medicinal chrysanthemum varieties (Supplementary Fig. S1).

DISCUSSION

'Huaihuang' has long been used as an herbal medicine in China. However, in chrysanthemum, RNA-seq data were only available for *C. nankingense* [26], *C. lavandulifolium* [27] and *C. morifolium* (Ramat.) cv. Fall Color [38]. Available EST-SSR sequences are also

Table 5. Nei's genetic identity and genetic distance in different varieties of chrysanthemum.

	CV1	CV2	CV3	CV4	CV5	CV6	CV7	CV8	CV9	CV10	CV11	CV12	CV13	WT1	WT2
CV1	****	0.5567	0.6278	0.6025	0.6083	0.5613	0.5723	0.6367	0.5594	0.6804	0.5818	0.4671	0.5686	0.4649	0.5060
CV2	0.5857	****	0.4905	0.4898	0.4796	0.4684	0.5555	0.5301	0.5111	0.4796	0.4849	0.3949	0.4971	0.5001	0.4816
CV3	0.4655	0.7032	****	0.6594	0.6108	0.5621	0.7083	0.5678	0.5605	0.5371	0.5734	0.4175	0.4840	0.5368	0.5169
CV4	0.5067	0.7137	0.4165	****	0.5430	0.8298	0.5797	0.5740	0.5778	0.5963	0.5270	0.4121	0.4600	0.5636	0.6030
CV5	0.4971	0.7348	0.4929	0.6106	****	0.5229	0.5555	0.5701	0.5430	0.5204	0.5253	0.3178	0.4690	0.5796	0.5201
CV6	0.5775	0.7585	0.5760	0.1866	0.6485	****	0.6163	0.5552	0.5683	0.5991	0.4960	0.3599	0.4205	0.6478	0.5963
CV7	0.5581	0.5878	0.3449	0.5453	0.5878	0.4841	****	0.4805	0.7047	0.4902	0.4745	0.3804	0.5206	0.8020	0.5244
CV8	0.4515	0.6347	0.5660	0.5551	0.5619	0.5884	0.7330	****	0.5845	0.5701	0.7624	0.4154	0.5424	0.6547	0.5759
CV9	0.5808	0.6712	0.5790	0.5486	0.6106	0.5651	0.3499	0.5370	****	0.5111	0.5165	0.4121	0.5383	0.8012	0.5628
CV10	0.3805	0.7348	0.6216	0.5170	0.6531	0.5123	0.7130	0.5619	0.6712	****	0.5556	0.4334	0.4877	0.7763	0.6915
CV11	0.5417	0.7239	0.5561	0.6405	0.6438	0.7011	0.7456	0.2713	0.6607	0.5877	****	0.4100	0.4921	0.8084	0.4939
CV12	0.7612	0.9292	0.8735	0.8866	1.1462	1.0221	0.9665	0.8785	0.8866	0.8361	0.8916	****	0.5046	1.0518	1.1170
CV13	0.5646	0.6990	0.7257	0.7766	0.7573	0.8663	0.6527	0.6117	0.6194	0.7180	0.7091	0.6840	****	1.0257	0.7758
WT1	0.7660	0.6929	0.6221	0.5734	0.5601	0.5232	0.4484	0.5196	0.4488	0.4601	0.4456	0.3493	0.3585	****	0.5287
WT2	0.6812	0.7307	0.6599	0.5058	0.6537	0.5170	0.6456	0.5519	0.5748	0.5008	0.6102	0.3273	0.4603	0.6374	****

Nei's genetic identity – above the diagonal; genetic distance – below the diagonal.

limited for chrysanthemum species, especially for medicinal chrysanthemum.

In the present paper, the transcriptome of medicinal chrysanthemum 'Huaihuang' is reported for the first time. There are more unigenes (162,252) assembled in the 'Huaihuang' transcriptome than ornamental *C. nankingense* (45,789) [26] or *C. lavandulifolium* (108,737) [27]. The mean length of the unigenes (701 bp) was longer than observed in transcriptome studies of *C. lavandulifolium* (349 bp) [27] and *C. nankingense* (585 bp) [26].Various factors, such as assembly program, parametric quantity assembly options and species differences can be ascribed to this result.

Due to the health-promoting functions of medicinal chrysanthemum 'Huaihuang', our focus was on the biosynthesis of primary active components. In the transcriptome data, we annotated many genes related to antioxidant, secondary and terpenoid metabolism. Compared to ornamental *C. nankingense* [26], there are more unigenes annotated into 'antioxidant activity' of GO annotations. Radicals are harmful compounds produced by the oxidation reaction and have strong oxidation properties. In physiological conditions, the excess of radicals has a damaging effect on the organism and is strongly associated with inflammation, cancer and other diseases. Plant metabolites with antioxidant activity can reduce and remove radicals, and have the effect of delaying organism aging and preventing and controlling diseases [39]. The group of 'Secondary metabolite biosynthesis, transport and catabolism' was fifth among the identified KOG categories in 'Huaihuang', while the same group was twelfth in *C. nankingense* [26] and thirteenth in *C. lavandulifolium* [27]. A number of unigenes were mapped to 'Biosynthesis of other secondary metabolites' and 'Metabolism of terpenoids and polyketides' in the KEGG metabolic pathway. Aromatic oils and pigments produced by plants are rich in terpenoids, which are chemical raw materials and a source of spices with important commercial value. The oxygen derivatives of monoterpenes and sesquiterpenes have strong biological activity and aroma, and are important raw materials in medicine, cosmetics and food industry. Plant phenylpropanoid substances are common in plants, and there are thousands of different chemical structures, including total flavonoids, flavonols, coumarins and tannins [40]. The discovery of genes related to the biosynthesis of flavonoids and terpenoids will help in further studies of the mechanism of their biosynthesis. Collectively, these unigenes could be used to reveal the antioxidative mechanism and excavate genes related to the biosynthesis, transport and catabolism of primary active components. This provides a considerable amount of

information for studying the biosynthetic pathway of primary active components in chrysanthemum.

EST-SSR markers have widespread applications in many areas of research such as the assessment of genetic diversity and population structure, the development of genetic maps, and in comparative genomics and molecular breeding [37,41]. In this research, the frequency of EST-SSR marker occurrence in 'Huaihuang' (1/6.86 kbp) was much higher than was obtained in *C. nankingense* (1/14.7 kbp) [26]. The differences in SSR locus frequency might be due to differences in the minimum length of SSR repeat motifs or genome size [42]. The identified EST-SSR significantly enriched the resources of available SSR for marker development in 'Huaihuang' and related species.

Among the dinucleotide and trinucleotide repeats, the frequency of a part of motifs was inconsistent with those reported in *C. nankingense* [26]. This suggests that the frequency may depend on species differences, SSR criteria for searching, database size, and on the database-mining tools used in different studies [20,43,44].

Among the six clades, clade I included CV1, CV10, CV8 and CV11, which was consistent with previous research [45]. *C. morifolium* cv. 'Boju' was formed by the gradual southward migration of *C. morifolium* cv. 'Huaiju' from Henan Province to Bozhou. CV12 and CV13 came from the same place (Wenxian, Henan, China) and all belonged to clade II, which is consistent with the phylogenetic relationship. Clade IV included CV3, CV7 and CV9, which is consistent with the fact that they have the same origin (Tongxiang, Zhejiang, China). These results are also in agreement with previous studies [45]. However, there are some exceptions, such as clade V that included CV4, CV6, WT1 and CV5, which may be due to the introduction of species among different regions. CV2 and WT2 belonged to the independent clades III and IV, respectively.

CONCLUSIONS

In this study, through RNA-Seq, a total of 162,252 unigenes was generated from 'Huaihuang' and 57,890 unigenes were successfully annotated into public databases. A number of genes associated with the biosynthesis of flavones and terpenes, such as *CHS*, *CHI*, *FLS*, *F3H*, *ANS*, *DFR*, *HMGR* and *MVD*, were found. The presented findings clarify the biosynthesis process of primary active components in 'Huaihuang'. In addition, 36 new EST-SSR markers were successfully identified, and were successfully used to assess the genetic diversity and to construct the dendrogram of 15 medicinal chrysanthemums varieties. These EST-SSR markers will not only enrich the molecular marker resources in 'Huaihuang', but can also be powerful molecular tools for population structure and evolutionary adaptation analysis, in genetic breeding and genetic relationship investigations in other chrysanthemum species.

Funding: This work was funded by the Joint Funds of the National Natural Science Foundation of China (No. U1704120) and the National Natural Science Foundation of China (No. 31372105).

Acknowledgments: The authors are grateful to Doctor Rika Judd (Department of Plant and Microbial Biology, North Carolina State University) for revising the paper.

Author contributions: The project was conceived and designed by X.Z. The preparation of material and experiments were carried out by Y.Z., L.S., L.J., K.L. and M.M. L.S. and M.L. analyzed the data and wrote the manuscript. X.Z. revised the manuscript. All authors have read and approved the final manuscript.

Conflict of interest disclosure: The authors declare that they have no conflict of interest, and the manuscript has been approved by all authors for publication.

REFERENCES

1. Li H, Chen S, Song A, Wang H, Fang W, Guan Z, Jiang J, Chen F. RNA-Seq derived identification of differential transcription in the chrysanthemum leaf following inoculation with *Alternaria tenuissima*. BMC Genomics. 2014;15(1):9.
2. Silva JATd, Shinoyama H, Aida R, Matsushita Y, Raj SK, Chen F. Chrysanthemum Biotechnology: *Quo vadis*? Crit Rev Plant Sci. 2013;32(1):21-52.
3. Liu PL, Qian W, Guo YP, Ji Y, Rao GY. Phylogeny of the Genus *Chrysanthemum* L.: Evidence from Single-Copy Nuclear Gene and Chloroplast DNA Sequences. PLoS One. 2012;7(11):e48970.
4. Zhang JY. Comparative Analysis of The Main Functional Components in Hang Chrysanthemum and Huai Chrysanthemum. Med Plant. 2010;1(8):84-6.
5. Sayre CL, Davies NM. Quantification of three chiral flavonoids with reported bioactivity in selected licensed Canadian natural health products and US marketed dietary supplements. J Pharm Pharm Sci. 2013;16(2):272.
6. Benaventegarcía O, Castillo J. Update on Uses and Properties of Citrus Flavonoids: New Findings in Anticancer, Cardiovascular, and Anti-inflammatory Activity. J Agric Food Chem. 2008;56(15):6185-205.

7. Agati G, Azzarello E, Pollastri S, Tattini M. Flavonoids as antioxidants in plants: Location and functional significance. Plant Sci. 2012;196(3):67-76.
8. Fraga CG, Martino VS, Ferraro GE, Coussio JD, Boveris A. Flavonoids as antioxidants evaluated by *in vitro* and *in situ* liver chemiluminescence. Biochem Pharmacol. 1987;36(5):717.
9. George VC, Dellaire G, Hpv R. Plant flavonoids in cancer chemoprevention: role in genome stability. J Nutr Biochem. 2017;45:1-14.
10. Zhang X, Zhang A, Jiang B, Bao Y, Wang J, An L. Further pharmacological evidence of the neuroprotective effect of catalpol from *Rehmannia glutinosa*. Phytomedicine. 2008;15(6–7):484-90.
11. Rajagopal S, Kumar RA, Deevi DS, Satyanarayana C, Rajagopalan R. Andrographolide, a potential cancer therapeutic agent isolated from Andrographis paniculata. J Exp Ther Oncol. 2010;3(3):147-58.
12. Klayman DL. Qinghaosu (artemisinin): an antimalarial drug from China. Science. 1985;228(4703):1049-55.
13. Chen LY, Cao YN, Yuan N, Nakamura K, Wang GM, Qiu YX. Characterization of transcriptome and development of novel EST-SSR makers based on next-generation sequencing technology in *Neolitsea sericea* (Lauraceae) endemic to East Asian land-bridge islands. Mol Breed. 2015;35(9):187.
14. Hu YH, Zhao P, Zhang Q, Wang Y, Gao XX, Zhang T, Zhou HJ, Dang M, Woeste KE. De novo assembly and characterization of transcriptome using Illumina sequencing and development of twenty five microsatellite markers for an endemic tree *Juglans hopeiensis* Hu in China. Biochem Syst Ecol. 2015;63:201-11.
15. Rai A, Yamazaki M, Takahashi H, Nakamura M, Kojoma M, Suzuki H, Saito K. RNA-seq Transcriptome Analysis of *Panax japonicus*, and Its Comparison with Other *Panax* Species to Identify Potential Genes Involved in the Saponins Biosynthesis. Front Plant Sci. 2016;7:e0144.
16. Li J, Wang C, Han X, Qi W, Chen Y, Wang T, Zheng Y, Zhao X. Transcriptome Analysis to Identify the Putative Biosynthesis and Transport Genes Associated with the Medicinal Components of *Achyranthes bidentata* Bl. Front Plant Sci. 2016;7(S15):1860.
17. Pan L, Xia QJ, Quan ZW, Liu HG, Ke WD, Ding Y. Development of novel EST-SSRs from sacred lotus (*Nelumbo nucifera* Gaertn) and their utilization for the genetic diversity analysis of *N. nucifera*. J Hered. 2010;101(1):71.
18. Zhou Q, Luo D, Ma L, Xie W, Wang Y, Wang Y, Liu Z. Development and cross-species transferability of EST-SSR markers in Siberian wildrye (*Elymus sibiricus* L.) using Illumina sequencing. Sci Rep. 2016;6:20549.
19. Guo Q, Wang JX, Su LZ, Lv W, Sun YH, Li Y. Development and Evaluation of a Novel Set of EST-SSR Markers Based on Transcriptome Sequences of Black Locust (*Robinia pseudoacacia* L.). Genes. 2017;8(7):177.
20. Zhang Y, Zhang X, Wang YH, Shen SK. *De Novo* Assembly of Transcriptome and Development of Novel EST-SSR Markers in *Rhododendron rex* Lévl. through Illumina Sequencing. Front Plant Sci. 2017;8:1664.
21. Levinson G, Gutman GA. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. Nucleic Acids Res. 1987;15(13):5323-38.
22. Jia X, Deng Y, Sun X, Liang L, Su J. De novo assembly of the transcriptome of *Neottopteris nidus* using Illumina paired-end sequencing and development of EST-SSR markers. Mol Breed. 2016;36(7):94.
23. Li CY, Chiang TY, Chiang YC, Hsu HM, Ge XJ, Huang CC, Chen CT, Hung KH. Cross-Species, Amplifiable EST-SSR Markers for *Amentotaxus* Species Obtained by Next-Generation Sequencing. Molecules. 2016;21(1):67.
24. Zheng X, Cheng P, Ying D, You Y, Yang C, Hu Z. Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). BMC Genomics. 2013;14(1):490.
25. Li MY, Feng W, Qian J, Jing M, Xiong AS. Identification of SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens* L.) by deep transcriptome sequencing. Hortic Res. 2014;1(10):10.
26. Wang H, Jiang J, Chen S, Qi X, Peng H, Li P, Song A, Guan Z, Fang W, Liao Y. Next-generation sequencing of the *Chrysanthemum nankingense* (Asteraceae) transcriptome permits large-scale UniGene assembly and SSR marker discovery. PLoS One. 2013;8(4):e62293.
27. Wang Y, Huang H, Ma Y, Fu J, Wang L, Dai S. Construction and de novo characterization of a transcriptome of *Chrysanthemum lavandulifolium*: analysis of gene expression patterns in floral bud emergence. Plant Cell, Tissue and Organ Culture (PCTOC). 2014;116(3):297-309.
28. Zhong S, Joung JG, Zheng Y, Chen YR, Liu B, Shao Y, Xiang JZ, Fei Z, Giovannoni JJ. High-throughput illumina strand-specific RNA sequencing library preparation. Cold Spring Harb Protoc. 2011;2011(8):940.
29. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767-71.
30. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnol. 2011;29(7):644.
31. Götz S, Garcíagómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36(10):3420-35.
32. Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics. 2005;21(19):3787-93.
33. Yang L, Zhang J, Sun H, Wang Y. Study on Apricot Genomic DNA Extraction by Modified SDS Method. Chinese Agr Sci Bull. 2008;24(4):69-71.
34. Benbouza H, Jacquemin JM, Baudoin JP, Mergeai G. Optimization of a reliable, fast, cheap and sensitive silver staining method to detect SSR markers in polyacrylamide gels. Biotechnologie. 2006;10(2):77-81.

35. Noormohammadi Z, Rahnama A, Sheidai M. EST-SSR and SSR analyses of genetic diversity in diploid cotton genotypes from Iran. Nucleus. 2013;56(3):171-8.
36. Nagy S, Poczai P, Cernák I, Gorji AM, Hegedűs G, Taller J. PICcalc: An Online Program to Calculate Polymorphic Information Content for Molecular Genetic Studies. Biochem Genet. 2012;50(9-10):670-2.
37. Xing W, Liao J, Cai M, Xia Q, Liu Y, Zeng W, Jin X. De novo assembly of transcriptome from *Rhododendron latoucheae* Franch. using Illumina sequencing and development of new EST-SSR markers for genetic diversity analysis in *Rhododendron*. Tree Genet Genomes. 2017;13(3):53.
38. Xu Y, Shan G, Yang Y, Huang M, Cheng L, Qian W, Fei Z, Gao J, Bo H. Transcriptome sequencing and whole genome expression profiling of chrysanthemum under dehydration stress. BMC Genomics. 2013;14(1):662.
39. Markland FSJ. 16 Subtilisins: Primary Structure, Chemical and Physical Properties. Enzymes. 1971;3:561-608.
40. Boudet AM. Evolution and current status of research in phenolic compounds. Cheminform. 2008;39(17):2722.
41. Yan X, Zhang X, Lu M, He Y, An H. De novo sequencing analysis of the *Rosa roxburghii* fruit transcriptome reveals putative ascorbate biosynthetic genes and EST-SSR markers. Gene. 2015;561(1):54-62.
42. Varshney RK, Thiel T, Stein N, Langridge P, Graner A. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett. 2002;7(2A):537.
43. Biswas MK, Chai L, Mayer C, Xu Q, Guo W, Deng X. Exploiting BAC-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in Citrus. Mol Biol Rep. 2012;39(5):5373-86.
44. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 2005;23(1):48-55.
45. Shao QS, Guo QS, Deng YM, Guo HP. A comparative analysis of genetic diversity in medicinal Chrysanthemum morifolium based on morphology, ISSR and SRAP markers. Biochem Syst Ecol. 2010;38(6):1160-9.

Supplementary Data

Available at: http://serbiosoc.org.rs/NewUploads/Uploads/Zhao%20et%20al_3767_Supplementary%20Data.pdf