RESEARCH ARTICLE

Transcriptome analysis and development of EST-SSR markers in *Anoectochilus emeiensis* (Orchidaceae)

Song Lu*

Sichuan Natural Resources Academy, Chengdu, Sichuan, China

* lusong06718@126.com

OPEN ACCESS

Citation: Lu S (2022) Transcriptome analysis and development of EST-SSR markers in *Anoectochilus emeiensis* (Orchidaceae). PLoS ONE 17(12): e0278551. https://doi.org/10.1371/journal.pone.0278551

Editor: Branislav T. Šiler, Institute for Biological Research, University of Belgrade, SERBIA

Received: June 7, 2022

Accepted: November 18, 2022

Published: December 6, 2022

Data Availability Statement: All relevant data are within the article and its Supporting information files.

Funding: The research was supported by Sichuan Natural Science Foundation Project (Grant No. 2022NSFSC0134) and Basic Scientific Research Programs Foundation of Sichuan Province (Grant No. 2021JDKY0023). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Anoectochilus emeiensis K. Y. Lang, together with other *Anoectochilus* species, has long been used as the main source of many traditional Chinese medicines. Owing to the shortcomings of molecular markers, the study of the genetic diversity and medicinal component synthesis mechanism of the endemic *Anoectochilus* species has been delayed. In this study, I carried out a transcriptome analysis of *A. emeiensis*. A total of 78,381 unigenes were assembled from 64.2 million reads, and 47,541 (60.65%) unigenes were matched to known proteins in the public databases. Then, 9284 expressed sequence tag-derived simple sequence repeats (EST-SSRs) were identified, and the frequency of SSRs in the *A. emeiensis* transcriptome was 9.88%. Trinucleotide repeats (3699, 39.84%) were the most common type, followed by dinucleotide (3251, 35.02%) and mononucleotide (1750, 18.85%) repeats. Based on the SSR sequence, 6683 primer pairs were successfully designed, 40 primer pairs were randomly selected, and 10 primer pairs were identified as polymorphic loci from 186 individuals of *A. emeiensis*. The EST-SSR markers examined in this study will be informative for future population genetic studies of *A. emeiensis*.

1. Introduction

Anoectochilus emeiensis K. Y. Lang, an endemic ethnomedicinal plant, is distributed on Mount Omei, southwestern China. It is a traditional Chinese perennial herbal medicine, similar to other species of the genus *Anoectochilus* (Orchidaceae), including *A. formosanus*, *A. roxburghii*, and *A. koshunensis*. In Chinese folk, it was named "King Medicine" owing to its diverse pharmacological effects such as the remedy for diabetes, hypertension, fever, tumor, liver disease, snakebite, and lung disease [1]. *Anoectochilus* species are usually diploid, with 40 chromosomes [2]. The whole diploid genome DNA content has been found to be approximately 6.83 pg based on flow cytometry analysis [3]. Overexploitation and depletion have occurred with the constantly increasing demand for wild *A. emeiensis* and other *Anoectochilus* species for their medicinal properties and ornamental value. Currently, *A. emeiensis* is considered "critically endangered" by the China Species Red List because of its habitat degradation or loss, narrow distribution, few populations, and small population sizes [4].

The revelation of population structure, local adaptation, and differences in diversity between populations is vital for understanding the genetic variation within a species [5]. Previous studies on *A*. *emeiensis* and other *Anoectochilus* species have mainly focused on the evaluation of germplasm resources, seed breeding, artificial cultivation techniques, chemical composition, and pharmacological effects [1]. However, molecular research and wild population protection of *A*. *emeiensis* have fallen behind.

As the most popular markers in population genetic studies, simple sequence repeats have been increasingly used in fingerprinting, parentage analysis, genetic mapping, and genetic structure analysis because of their codominant and highly polymorphic nature in recent years [6–8]. In the past two decades, next-generation sequencing technology has been developed so quickly that not only model but also non-model organisms without sequenced genomes have been successfully sequenced for their high-throughput, accurate, and cost-effective advantages in transcriptome analysis [9–13]. Transcriptome sequencing is an efficient process for producing expressed sequence tags that are valuable for the development of molecular markers, gene annotation and discovery, and expression profiling [9, 11, 14]. Because of the lengthy and costly development phase and relatively low throughput, the development of single sequence repeats (SSRs) based on genome sequencing has not been as frequently adopted as transcriptome sequencing. Expressed sequence tag-derived SSR (EST-SSR) markers have been successfully developed by *de novo* transcriptome sequencing and assembly in many crops, such as barley, rice, and wheat [15–17], and in some Chinese herbal medicines, such as *Dendrobium officinale*, *Dysosma versipellis*, *Panax ginseng*, *Panax vietnamensis*, and *Siberian ginseng* [18–21]. To better protect wild germplasm resources, it is imperative to sequence the transcriptome of *A*. *emeiensis*. In the present study, the transcriptome of *A*. *emeiensis* leaves was sequenced using the Illumina HiSeq 2000 platform, and several EST-SSRs were derived. To our knowledge, this is the first systematic report of the complete transcriptome of *A*. *emeiensis*. The aims of this study were to (1) characterize and analyze the transcriptome of *A*. *emeiensis*, (2) determine the frequency and distribution of SSRs, (3) establish EST-SSR markers and examine the level of polymorphism, and (4) assess the genetic diversity of *A*. *emeiensis* among six populations.

2. Materials and methods

2.1. Plant materials, DNA, and RNA isolation

A total of 186 samples of *A*. *emeiensis* were collected from six natural populations in Mount Omei, Sichuan Province, China, at Longdongcun (LDC), Longdonghu (LDH), Daping (DP), Wuhe (WH), Jinzhulin (JZL), and Chuanzhusi (CZS) (August 8, 2015). Field research was approved by the Emeishan Ecological Environment Bureau, Sichuan Province, China. The collected fresh leaves of *A*. *emeiensis* had a height of 8–10 cm. After freezing in liquid nitrogen, all leaves were stored at −80 ˚C until DNA and total RNA extraction. Genomic DNA was isolated using DNA Plantzol (QIAGEN, Dusseldorf, Germany), following the manufacturer's protocol. Voucher specimens were maintained in the herbarium of the Sichuan Institute of Natural Resources. Total RNA from ten different individuals was extracted using TRIzol reagent (Invitrogen Life Technologies, USA), following the manufacturer's instructions. The quality and concentration of RNA were determined using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Equal proportions of high-quality RNA were pooled from each individual to construct a cDNA library.

2.2. Transcriptome sequencing, assembly, and annotation

Following the methods described by Liu et al. [10, 14, 22], a normalized cDNA library of *A. emeiensis* leaves was constructed. Briefly, the total RNA was collected and the poly(A) mRNA was purified with magnetic oligo(dT) beads. Then the mRNA was cleaved to short fragments by an RNA Fragmentation Kit (Ambion, Austin, TX), which were then used as templates to reverse-transcribe first-strand cDNA using random hexamer primers and reverse transcriptase (Invitrogen, Carlsbad, CA). Second-strand cDNA was synthesized in a reaction containing buffer, dNTPs, RNase H (Invitrogen) and DNA polymerase I (Qingkeweiye Biological Technology, Chengdu). The library was paired-end synthesized according to Illumina Genomic Sample Preparation Kit manufacturer's instructions and sequenced using Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA) at One Gene Company Limited, Hangzhou, China. Image analysis, base calling, and quality value calculations were performed using the Illumina GA Pipeline version 1.6. The raw Illumina sequencing data were then submitted to the NCBI Sequence Read Archive (accession number: SRR8138703) (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra). Clean reads were obtained by removing those of low quality (more than 5% ambiguous bases "N") and reads with more than 20% low-quality bases (value of $\leq$10) from raw reads using FastQC Version 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). *De novo* assembly was performed using Trinity [12]. Generally, clean reads were combined to form longer fragments without Ns, termed contigs. The remaining isolated reads are described as singletons. Individual reads were then arranged back to the contigs using paired-end mapping. Subsequently, the order and distance between these contigs were determined from the same transcript. Finally, Trinity connected contigs and obtained sequences that could not be extended at either end, which were named as unigenes. These unigenes were further processed by sequence splicing and redundancy removal using TGICL version 2.1 to obtain nonredundant unigenes [23]. To intensively study the putative functions, the unigenes were compared using the BLASTX search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with a cut-off *E* value of 10^{-5} against protein databases such as NCBI nonredundant protein sequences (Nr), Swiss-Prot (http://www.expasy.ch/sprot/), the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (http://www.genome.jp/kegg), and the Clusters of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG). To further annotate the unigenes, gene ontology (GO) was performed using the Blast2GO program (B2G; http://www.blast2go.com) [24].

2.3. Identification of EST-SSRs and primers design

SSR loci were detected using MIcroSAtellite software (MISA) (http://pgrc.ipkgatersleben.de/misa) [16]. The minimum number of repeats was six for dinucleotides, five for tri- and tetranucleotides, four for penta- and hexanucleotides, and twelve for mononucleotides. PRIMER version 3.0 [25] was used to design the primer pairs with the following settings: 18–28 bases (optimum 23 bp), product size of 80–160 bp, annealing temperature between 55 ˚C and 65 ˚C (optimum 60 ˚C), and CG content from 40% to 60% (optimum 50%). The potential primer pairs, hairpin structure, and occurrence of mismatches were checked using OLIGO (version 6.67; Molecular Biology Insights, Inc., Cascade, CO, USA). Finally, the specificity of the primer pairs was checked using BLAST against EST sequences.

2.4. EST-SSR analysis

The EST-SSR markers were initially tested for amplification using DNA from 24 *A. emeiensis* individuals to optimize the annealing temperature. In total, 40 SSRs were selected, and M13 (universal sequence-`TGTAAAACGACGGCCAGT`) [26] was added to the 5' end of each forward primer of pairs and synthesized by Qingkeweiye Biological Technology (Chengdu) Co., Ltd. Polymerase chain reaction (PCR) amplification was performed using a thermal cycler GeneAmp PCR System 9700 (Applied Biosystems, Foster City, USA) in a 10 μl reaction mixture containing 1 μl of genomic DNA (50 ng), 0.25 unit *Taq* DNA polymerase (TaKaRa, Dalian, Liaoning, China), 1 × PCR buffer, 1 μl of 2.5 mM $MgCl^2$, 1 μl of 2.5 mM dNTPs, 0.1 μl bovine serum albumin (TaKaRa), 0.5 μl of each 10 μM primer, and 0.5 μl of 10 μM fluorescent labeled M13 primer. The PCR amplification conditions were as follows: hot start at 95 ˚C for 5 min, followed by 35 cycles of denaturation at 94 ˚C for 45 s, annealing at various primer-specific temperatures for 45 s, extension at 72 ˚C for 1 min, and a final extension at 72 ˚C for 10 min. The PCR products were analyzed using a LI-COR 4300 DNA Analyzer (Gene Company Limited, USA), and their sizes were scored and compiled using the SAGA Generation 2 software (LI-COR).

2.5. Selective neutrality tests and population genetic analysis

The number of observed alleles (N_A), observed (H_O) and expected (H_E) heterozygosity, and polymorphism information content (PIC) in each locus were calculated to test the polymorphism level using CERVUS version 3.0.3 [27]. For each population and locus of *A*. *emeiensis*, the diversity and inbreeding parameters, such as N_A, H_E, allelic richness (R_S, standardized for the smallest number of individuals per unit using rarefaction), and the average inbreeding coefficient across all loci (F_{IS}) were estimated using FSTAT version 2.9.3 [28].

3. Results

3 1. Assembly of *A*. *emeiensis* transcriptome data, functional annotation, and categorization of unigenes

In total, 10.7 Gb, 71,389,320 raw reads of 90 bp in length and 64,177,276 clean reads (89.90%) with 97.72% Q20 bases (base quality greater than 20) were generated. The total length of the reads was approximately 9.6 Gb (Table 1). The percentage of ambiguous bases and GC percentage for *A*. *emeiensis* were 0 and 47.70%, respectively. After TRINITY assembly, high-quality reads were assembled into 115,283 contigs with a mean length of 852 bp and an N50 length of 1375 bp. These contigs were then stitched into 78,381 unigenes with a mean length of 1002 bp and an N50 length of 1602 bp. Of the 78,381 unigenes with a total length of approximately 78.5 Mb, 33,002 unigenes (42.10%) had a length range of 200–500 bp, 35,462 unigenes

Table 1. Summary of the analysis of *A*. *emeiensis* transcriptomic data.

Number of raw reads	71,389,320
Number of high-quality reads	64,177,276
Number of clean nucleotides (nt)	9,626,591,400
Total length of contigs (bp)	98,217,609
Total length of unigenes (bp)	78,535,310
Number of contigs	115,283
Number of unigenes	78,381
Number of singletons	45,700
Total number of identified SSRs	9284
Number of SSR-containing sequences	7,742
Number of sequences containing more than one EST-SSR	1,195
Number of SSRs present in compound formation	791

https://doi.org/10.1371/journal.pone.0278551.t001

(45.24%) had a length range of 501–1000 bp, and 9917 unigenes (12.65%) had a length of more than 2000 bp.

Sequence similarity was determined using the BLAST algorithm. Of the 78,381 unigenes, 47,541 (60.65%) matched the known proteins in the public databases searched. That is 45,228 (57.70%), 34,547 (44.08%), 29,347 (34.77%), 27,793 (35.46%), 19,203 (24.50%), and 28,464 (36.31%) unigenes in the Nr, Nt, SwissProt, KEGG, COG, and GO databases, respectively. Furthermore, 80.8% of the Nr mapped unigenes showed high homologous matches to available plant sequences (E value $<10^{-15}$), and 58.9% had significant conservation (E value $<10^{-45}$). Regarding the similarity distribution, 86.5% of the Nr assembled sequences showed more than 40% similarity to known sequences, and 13.9% of the annotated sequences showed more than 80% similarity. For the species distribution, 16.4% (7395) annotated sequences were hit on *Vitis vinifera*, while 3.5–6.7% orthologous genes were distributed in *Amborella trichopoda*, *Amygdalus persica*, *Chaetochloa italica*, *Zea mays*, *Japanese rice*, and *Theobroma cacao*.

A total of 19,203 unigenes were aligned to the COG database, and 39,622 matched unique sequences were clustered into 25 COG categories (Fig 1). The largest category among these classifications was the general function prediction only (6375, 16.09%), followed by transcription (3937, 9.94%); replication, recombination, and repair (3630, 9.16%); signal transduction mechanisms (2711, 6.84%); post-translational modification, protein turnover, and chaperones

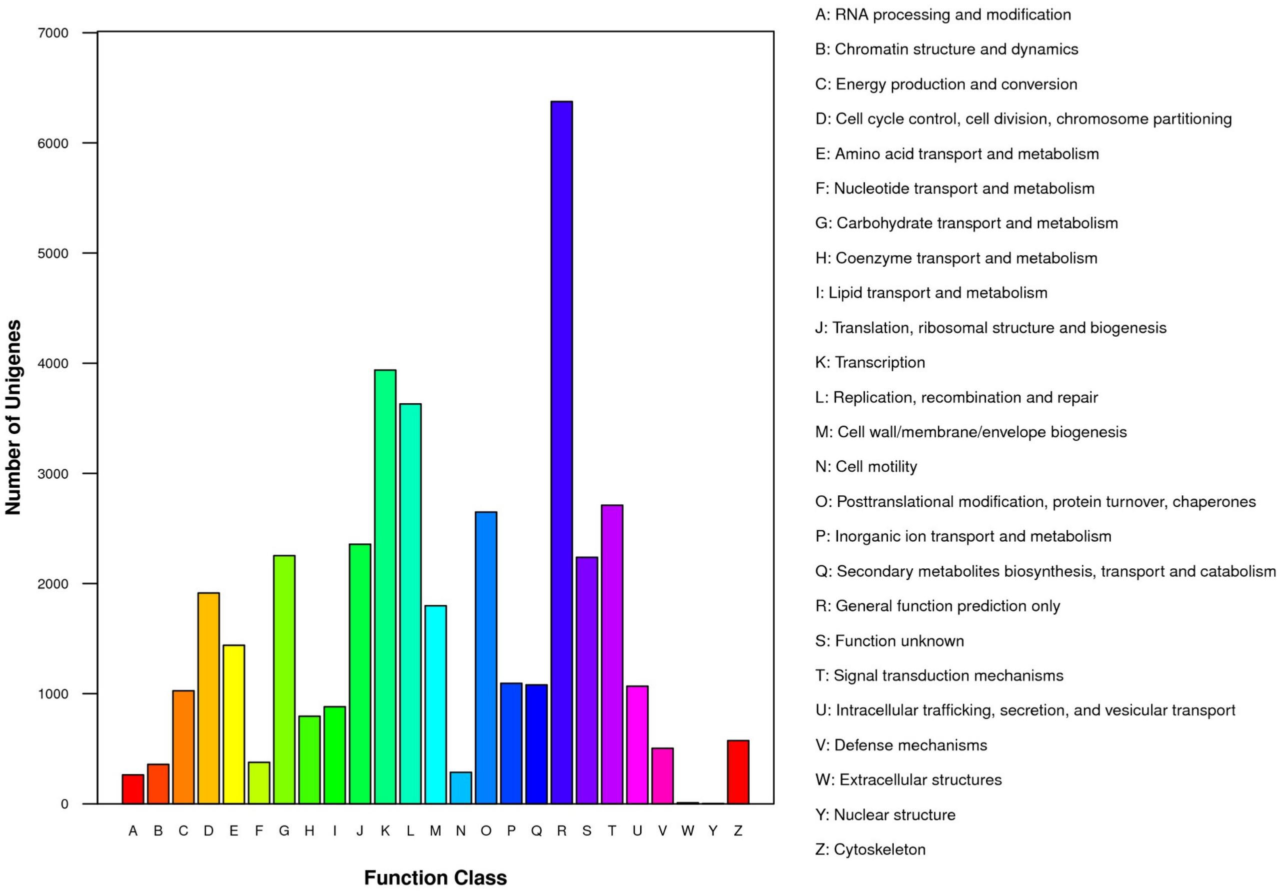

Fig 1. COG function classification by BLASTx with an E value threshold of 10¯5 against COG databases.

https://doi.org/10.1371/journal.pone.0278551.g001

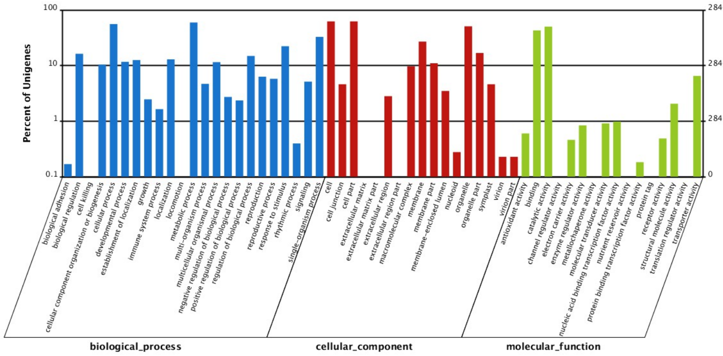

Fig 2. GO classification for *A. emeiensis*. by BLASTx with an E value threshold of 10¯5 against the Nr database.

https://doi.org/10.1371/journal.pone.0278551.g002

(2649, 6.69%); translation, ribosomal structure, and biogenesis (2357, 5.95%); and carbohydrate transport and metabolism (2253, 5.69%). Approximately 2238 (5.65%) unigenes were poorly characterized as function unknown. Extracellular and nuclear structures contained the fewest unigenes.

Using the Blast2GO software, 28,464 (41.63%) matched unigenes were assigned to three main categories: biological process (83,272, 44.64%), cellular component (73,002, 39.14%), and molecular function (30,251, 16.22%) (Fig 2). Fifty-six subcategories were further subdivided into GO terms. The three major sub-categories of the biological process were "metabolic process" (16,960, 20.37%), "cellular process" (15,899, 19.09%), and "single-organism process" (9339, 11.22%). In terms of cellular components, "cell" (17,746, 24.31%), "cell part" (17,746, 24.31%), "organelle" (14,474, 19.83%), and "membrane" (7683, 10.52%) represented the most common terms. Catalytic activity (14,298, 47.26%), binding (12,202, 40.33%), and transporter activity (1853, 6.12%) were the dominant molecular functions. To further explore the biological pathways and gene interactions of *A. emeiensis*, a BLASTX search was performed against the KEGG database: 27,793 (35.46%) of the 78,381 unigenes were grouped into 128 KEGG pathways. Of these pathways, metabolic pathways (7072, 25.45%) were the most representative, followed by biosynthesis of secondary metabolites (3146, 11.32%), endocytosis (1482, 5.33%), and glycerophospholipid metabolism (1409, 5.07%). Besides, pathways related to terpenoid backbone (265, 0.95%), flavonoid (211, 0.76%), flavone and flavonol (137, 0.49%), steroid (136, 0.49%), isoflavonoid (80, 0.29%), and monoterpenoid (15, 0.05%) biosynthesis were also represented by unigenes. These pathways may be conducive to research on the metabolism of active ingredients in *A. emeiensis*.

3.2. SSRs identification and characterization

The results of the MISA search showed that 7742 unigenes contained 9284 EST-SSRs. There were 1195 unigenes containing more than one EST-SSR locus, and 678 SSRs were present during compound formation. The frequency of SSRs in the *A. emeiensis* transcriptome was 9.88%. Trinucleotide repeats (3699, 39.84%) were the most common type, followed by dinucleotides (3251, 35.02%), mononucleotides (1750, 18.85%), hexanucleotides (229, 2.47%), pentanucleotides (189, 2.04%), and tetranucleotides (166, 1.79%). In total, five tandem SSRs (22.27%) were the most common, followed by six (21.85%), seven (12.26%), and eight (8.61%). Subsequently, 355 different types of motifs were identified. Of the 3251 dinucleotide repeats, AG/CT was the most dominant motif (2170, 23.37%), followed by AT/AT (834, 8.98%), AC/GT (232, 2.50%), and CG/CG (15, 0.16%). In the 3699 trinucleotide repeats, the most abundant repeat motifs were AAG/CTT (1253, 13.50%), followed by ATC/ATG (720, 7.76%), CCG/CGG (492, 5.30%), AAT/ATT (426, 4.59%), AGC/CTG (298, 3.21%), AGG/CCT (277, 2.98%), ACC/GGT (117, 1.26%), AAC/GTT (77, 0.83%), ACG/CGT (27, 0.30%), and ACT/AGT (12, 0.13%) (Fig 3). The 14 repeat motifs mentioned above constituted 74.86% of the total, whereas the remaining 341 types constituted 25.14%.

3.3. Polymorphism of EST-SSR markers and population genetic diversity

A total of 6683 primer pairs were successfully designed, with amplicons ranging from 80 to 240. Subsequently, 40 primer pairs were randomly selected. In the initial screening of 24

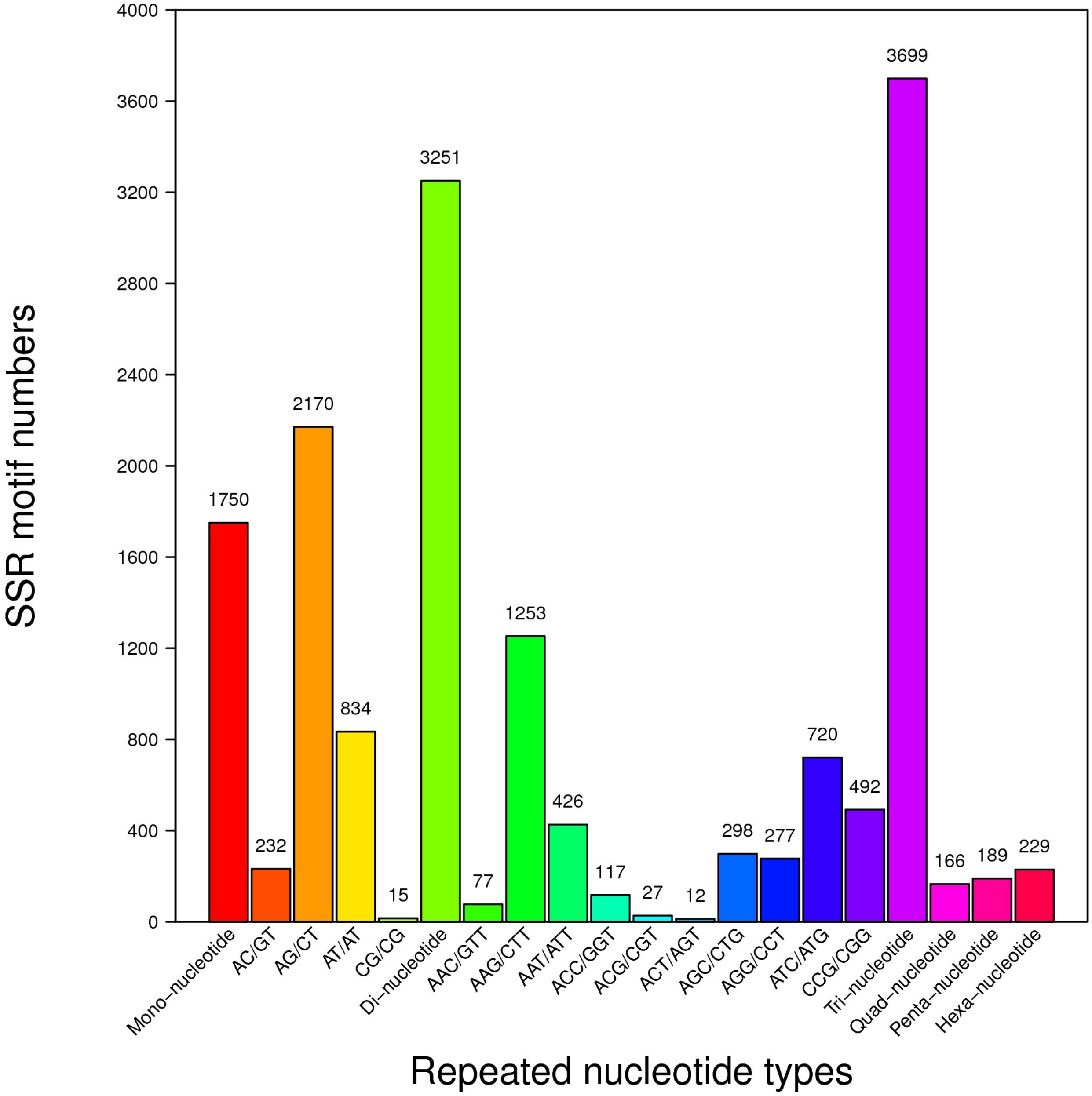

Fig 3. The SSRs repeated nucleotide types for *A. emeiensis*.

https://doi.org/10.1371/journal.pone.0278551.g003

Table 2. Ten EST-SSR markers developed in *A. emeiensis* and characteristics of each primer pairs.

Locus	Primer pair (5′-3′)	Repeat motif	Size range(bp)	N_A	H_O	H_E	PIC	Tm (˚C)
U1017	F:CACCAAAAGAGACTCATCTCACC R:TTGCCATTTATTGTCAAATTGCT	$(CAT)_8$	156–177	8	0.446	0.799	0.767	60
U42760	F:CATCTTATCCTTGCTCTCCTCAA R:TGTTATCAGTGGGCTGAAACTG	$(CGC)_7$	157–169	5	0.269	0.344	0.314	56
U32259	F:CTCTTCATCGATGTCCAAAGACT R:CTTCGTCTGGAACTTATTGGGTT	$(TTC)_6$	123–135	5	0.505	0.729	0.678	58
CL1528	F:TCTACACCCATTCAAAGAAGCAT R:CAATCGGCATCTATTTTTACTCG	$(CTGC)_5$	154–178	7	0.481	0.597	0.542	60
U28269	F:AGTTCCCTGCTTTATTTTGTTCC R:AAAAGGAGTTGATGCCTAGTGTG	$(CT)_7$	156–166	5	0.359	0.695	0.636	58
U1140	F:AGCAAGGTAATAGTTGTGCCAAA R:TCCCTTCCCTCAAATTCTTATTC	$(GA)_6$	157–165	5	0.337	0.518	0.470	58
CL1364	F:ACCTTTTTCCATTAATCTGGGAA R:CAACTCTCTCCGTCAAGTCAAAT	$(GT)_6$	149–158	3	0.328	0.387	0.353	60
U7123	F:ATCTTGAAGTTCGATGCAGAAAG R:AACCCCTTTATGTATTCTTTGCAG	$(AAG)_7$	113–128	5	0.237	0.238	0.217	56
U30086	F:CCCCATAAGTAATTGTTTTGCAT R:GACAGAGTTTGTGTGAGAGAGGG	$(TG)_6$	155–158	4	0.481	0.678	0.629	56
CL3430	F:ATAACTTGGGGAAGCTGTGAAGT R:TGATTTGAAGTAAATGGCACAGA	$(GA)_6$	153–168	6	0.457	0.612	0.531	56

N_A = number of alleles per locus; H_O = observed heterozygosity; H_E = expected heterozygosity; PIC = polymorphism information content; Tm = annealing temperature.

https://doi.org/10.1371/journal.pone.0278551.t002

individuals from the six populations, 10 primer pairs were polymorphic. These were used to test the polymorphism in 186 individuals from the six populations. A total of 53 alleles were detected at ten loci. The number of alleles (N_A) ranged from 3 to 8, with an average of 5.3. The observed heterozygosity (H_O) varied from 0.237 to 0.505 with a mean value of 0.39. The expected heterozygosity (H_E) ranged from 0.238 to 0.799 with an average of 0.560. The PIC values fluctuated from 0.217 to 0.767, with a mean of 0.514 (Table 2).

In each population, the mean number of alleles (N_A) across the 10 surveyed SSR markers varied from 2.6 to 4.8, and the LDC population was the highest. The average allelic richness (R_S) values ranged from 2.6 to 3.153, and the LDC population was also the highest. The mean expected heterozygosity (H_E) was 0.0211–0.7134. The mean F_{IS} fluctuated from −0.058 to 0.332, and the value in the LDH population was the highest (Table 3).

4. Discussion

4.1. Characterization of the *A. emeiensis* transcriptome

Previously, no ESTs of *A. emeiensis* were available in public databases. In this study, 78,381 unigenes were obtained from *A. emeiensis* leaves with a mean length of 1002 bp, which were longer than those of previous transcriptome studies of *A. formosanus* (mean length 679 bp), *Dendrobium officinale* (mean length 742 bp), *Panax ginseng* (mean length 469 bp), and *Siberian ginseng* (mean length 785 bp), but lower than those of *Panax vietnamensis* (mean length 1304 bp) [18–21, 29]. The number of unigenes assembled in *A. emeiensis* was lower than that reported for *A. formosanus* (173,513), *Panax ginseng* (178,145), and *Panax vietnamensis* (126,758), but higher than that in *Dendrobium officinale* (36,407) and *Dysosma versipellis* (44,855) [7]. Approximately 60.65% (47,541) unigenes were successfully annotated in public databases (Nr, Nt, Swiss-Prot, COG, GO, and KEGG). Compared with previous studies, the number of annotated unigenes in *A. emeiensis* was greater than that in *Dendrobium officinale*

Table 3. Characteristics of the 10 EST-SSRs in each population of *A. emeiensis*.

Locus	Longdongcun (LDC, $N = 83$)				Daping (DP, $N = 45$)				Longdonghu (LDH, $N = 27$)			
	N_A	R_S	H_E	F_{IS}	N_A	R_S	H_E	F_{IS}	N_A	R_S	H_E	F_{IS}
U1017	8	4.654	0	0.455	7	4.787	0.1312ns	0.201	6	4.207	0	0.472
U42760	5	2.548	0.2463ns	0.138	4	2.296	0.1382ns	0.275	3	2.082	0.0026	0.566
U32259	4	3.587	0.0004	0.290	5	3.713	0.0022	0.268	4	3.808	0.0003	0.417
CL1528	5	3.230	0.0706	0.159	6	3.131	0.0094	0.102	5	3.302	0.0055	0.445
U28269	5	3.626	0	0.467	4	3.443	0	0.436	4	2.885	0.0001	0.652
U1140	5	2.732	0	0.536	5	3.121	0.0213	0.284	4	3.159	0.4490	0.160
CL1364	3	2.436	0.0666	0.161	3	2.192	0.0258	0.266	3	2.486	1	-0.112
U7123	4	1.975	0.0524	0.172	4	1.848	1	-0.090	2	1.466	1	-0.040
U30086	4	3.496	0	0.291	4	3.456	0.0001	0.238	4	3.561	0.0002	0.533
CL3430	5	3.243	0.0003	0.313	4	2.903	0.5137	0.052	3	2.335	0.0192	0.231
Mean	4.8	3.153	0.0211	0.298	4.6	3.089	0.1966	0.203	3.8	2.929	0.2477	0.332
Locus	Jinzhulin (JZL, $N = 19$)				Chuanzhusi (CZS, $N = 7$)				Wuhe (WH, $N = 5$)			
	N_A	R_S	H_E	F_{IS}	N_A	R_S	H_E	F_{IS}	N_A	R_S	H_E	F_{IS}
U1017	5	4.353	0	0.596	4	3.867	0.0077	0.636	3	3.000	0.0159	1.000
U42760	2	1.462	1.0000ns	-0.029	2	1.934	1.0000ns	-0.091	2	2.000	1.0000ns	-0.143
U32259	5	3.959	0.0148ns	0.371	4	3.648	0.3606ns	0.000	3	3.000	0.619ns	0.360
CL1528	4	3.292	0.2470	0.225	4	3.647	0.2727	-0.500	2	2.000	-	0.000
U28269	4	3.405	0.1281	0.253	1	1.000	-	-	3	3.000	0.3333	0.238
U1140	4	2.918	0.2648	0.289	2	1.934	1	-0.091	3	3.000	1	-0.200
CL1364	3	2.185	0.0071	0.636	3	2.868	1	-0.200	2	2.000	1	-0.333
U7123	2	1.936	0.5384	-0.241	2	1.989	1	-0.200	2	2.000	1	-0.333
U30086	4	3.484	0.4428	0.038	3	2.923	1	-0.021	3	3.000	0.1111	0.143
CL3430	3	2.235	1	-0.087	1	1.000	-	-	3	3.000	0.3651	0.407
Mean	3.6	2.923	0.3285	0.205	2.6	2.481	0.7134	-0.058	2.6	2.6	0.5465	0.114

N_A = number of detected alleles per locus; R_S = allelic richness; H_E = expected heterozygosity; F_{IS} = inbreeding coefficient; n.s = not significant.

https://doi.org/10.1371/journal.pone.0278551.t003

(25,473) but lower than those in *A. formosanus* (128,054), *Panax ginseng* (94,535), *Panax vietnamensis* (85,214), and *Siberian ginseng* (48,300). The major chemical constituents of *A. emeiensis* are polysaccharides, flavonoids, glycosides, steroids, and triterpenes et al. [1]. Unigenes sorted to these constituent biosynthesis pathways had been annotated as carbon metabolism, O-glycan and N-glycan biosynthesis, flavonoid, flavone and flavonol biosynthesis, terpenoid backbone biosynthesis, and mono-, di-, sesqui-, and triterpenoid biosynthesis.

In planting, the quality of *Anoectochilus* is mostly evaluated by the content of flavonoids, such as rutin, quercetin, isorhamnetin, and kaempferol [30, 31]. Flavonoids are important pharmacodynamic components in *Anoectochilus*. Flavonoid biosynthesis is very complex, and there are many intersection points in the synthetic pathway. According to the KEGG annotation and the species of flavonoids separated and purified from *Anoectochilus*, the key enzymes in the flavonoid synthesis pathway include chalcone synthase, chalcone isomerase, naringenin 3-dioxygenase, flavonol synthase, flavonoid 3'-hydroxylase, flavone O-methyltransferase, flavonol 3-O-glucosyltransferase, and flavonol-3-O-glucoside L-rhamnosyltransferase. These enzymes play key roles in the metabolic pathways from cinnamoyl-CoA, naringenin chalcone, naringenin, dihydrokaempferol, and kaempferol, to quercetin, isorhamnetin, isoquercetin, rutin, and narcissoside [32]. Many components are both intermediates and important pharmacodynamic substances such as kaempferol, quercetin, isoquercetin, and rutin. *A. emeiensis*, together with other species of *Anoectochilus*, have been simultaneously studied for novel gene discovery and exploration of their medicinal mechanism.

Furthermore, 4.47% unigenes were annotated to the plant-pathogen interaction pathway, most of which belonged to environmental adaptation groups. *Anoectochilus* plants are symbiotic with fungi and gain nutrition from the environment. These genes will help reveal how tiny exalbuminous seeds can survive under extreme conditions.

4.2. Frequency and distribution of EST-SSRs

A total of 9284 EST-SSRs were mined and distributed among 7742 unigenes. The SSR frequency in the *A*. *emeiensis* transcriptome (9.88%) was higher than that in maize (1.5%) and rice (4.7%) but lower than that in *A*. *formosanus* (12.21%) and *Neolitsea sericea* (16.25%), and it was the same as that in many dicot species that fall in this range of values (2.65–16.82%) [7, 29, 33]. The variable results may be due to different SSR mining software and criteria [33].

In *A*. *emeiensis*, the SSR motif types were not evenly distributed. Trinucleotide repeats (39.84%) were the most frequent type, followed by dinucleotides (35.02%) and mononucleotides (18.85%). The hexa-, penta-, and tetranucleotide repeat types were much rarer (2.47, 2.04, and 1.79%, respectively). The present result was congruent with earlier suggestions that trinucleotide repeats are the main EST-SSR type in both mono- and dicots [7, 34–37]. This is in contrast to many conclusions that dinucleotide repeats are the dominant repeats [33]. Also, this result is different from *A*. *formosanus*, in which mononucleotides have been determined to be the most common type (61.93%) [31]. The predominant trimeric SSRs may be attributed to the suppression of non-trimeric SSRs in coding regions [32]. In *A*. *emeiensis*, AAG/CTT (13.50%) and AG/CT (23.37%) motifs were the most common tri- and dinucleotide repeats, and CCG/CGG (5.30%) was also frequent. These results are congruent with those of previous studies on *Arabidopsis*, soybean, cucumber, sesame, cocoa, and caster bean, which also suggested that the trinucleotide AAG motif may be prominent in dicots [33, 38].

4.3. Polymorphism of EST-SSR markers and population genetic diversity and structure

In total, 53 alleles were detected across 186 *A*. *emeiensis* individuals from six populations with 10 polymorphic EST-SSR markers (Table 2). The N_A values across these markers ranged from 3 to 8 with a mean of 5.3, which were higher than those of *Dysosma versipellis* and *D*. *pleiantha*, but lower than that of *Neolitsea sericea* [7, 14]. The average PIC value was 0.514 for the ten developed loci. The PIC is used to assess the genetic information level. The results of this study showed that the PIC value fell at a medium level and showed high polymorphism.

After years of investigation, I found that there were two reproductive modes in the wild *A*. *emeiensis* community: asexual reproduction and insect pollination. Perennial rhizomes with induced *A*. *emeiensis* buds can grow to more than 20 cm. The new shoots grow completely from the succulent stem nodes. Although wild populations lack pollinating insects, *A*. *emeiensis* can occasionally be pollinated by insects. These factors contribute to the unique genetic structure of *A*. *emeiensis*.

Typical habitat fragmentation occurs in wild populations of *A*. *emeiensis*. Owing to the construction of roads, parking lots, and hotels, the distribution area has been isolated into several groups. Wild *A*. *emeiensis* was concentrated in the LDC, DP, LDH, and JZL populations. Allelic richness (*Rs*) was positively correlated with population size (Table 3). Almost all of the detected alleles (N_A) declined in the small populations (Table 3). For example, in the LDC population, the N_A of U1017 was 8 but declined to 4 in CZS and 3 in WH. Thus, four and five alleles of U1017 were lost in the CZS and WH populations, respectively. The lowest allele frequency of U1017 was 174. However, this allele was lost in the JZL, CZS, and WH populations (S1 Table). The results showed that when the population becomes smaller owing to habitat fragmentation, the number of alleles in the remaining population mainly depends on the size of the population and the allele frequency distribution [39, 40]. Alleles with lower frequencies will be lost.

In each population, inbreeding was prominent among the individuals (Table 3): 83.33% of the mean inbreeding coefficient (F_{IS}) was positive, providing evidence for this. Inbreeding increases homozygosity and reduces heterozygosity [41]. In the continuous inbreeding population, the frequency of heterozygotes tended to be zero. This was proven by the fact that H_E of many alleles in four larger populations (LDC, DP, LDH, and JZL) was 0. Lacy and Lindenmayer (1995) found that when the population is divided into different small parts, heterozygosity and allelic diversity decrease rapidly [42]. The results of this study support this conclusion.

Overall, the EST-SSRs derived from this study are of effective quality in *A*. *emeiensis*. These markers will aid in investigating the genetic structure of wild populations, and the generated database will help us understand the molecular synthesis mechanisms of medicinal components in *A*. *emeiensis*.

5. Conclusions

In this study, I performed the first transcriptome analysis of *A*. *emeiensis* leaves using the Illumina HiSeq 2000 sequencing system. A total of 78,381 unigenes were assembled from 64.2 million reads, and 47,541 (60.65%) unigenes were matched to known proteins in public databases. Then, 9284 EST-SSRs were identified, and the frequency of SSRs in the *A*. *emeiensis* transcriptome was 9.88%. Trinucleotide repeats (3699, 39.84%) were the most common type, followed by dinucleotide (3251, 35.02%) and mononucleotide (1750, 18.85%) repeats. Based on the SSR sequence, 6683 primer pairs were successfully designed, 40 primer pairs were randomly selected, and 10 primer pairs were identified as polymorphic loci from 186 individuals of *A*. *emeiensis*. The EST-SSR markers examined in this study will be informative for future population genetic studies of *A*. *emeiensis*.

Supporting information

S1 Table. Raw allele scoring table based on the 10 EST-SSRs in each population of *A*. *emeiensis*.
(XLSX)

Acknowledgments

The author is grateful to the anonymous reviewer for insightful comments on this manuscript. The author would like to thank Editage (www.editage.cn) for English language editing.

Author Contributions

Writing – review & editing: Song Lu.

References

1. Ye SY, Shao QS, Ye AL. *Anoectochilus roxburghii*: A review of its phytochemistry, pharmacology, and clinical applications. J Ethnopharmacol. 2017; 209: 184–202. https://doi.org/10.1016/j.jep.2017.07.032 PMID: 28755972

2. Zeng YJ, Xiao HSH, Huang DQ. Chromosome karyotype analysis of *Anoectochilus roxburghii*. J Fujian Normal Univ (Natural Science). 2001; 17(4): 118–120.

3. Lin H, Wang JG, Huang HX, He BZH, Lan SR, Ma XK. Estimation of genome size of *Anoectochilus roxburghii* by flow cytometry analysis. J. Fujian Agric For Univ (Natural Science). 2020; 49(6):766–771.

4. Wang S, Xie Y. China species red list. Higher Education Press, Beijing, 2004.

5. Marsden CD, Woodroffe R, Mills MG, McNutt JW, Creel S, Groom R, et al. Spatial and temporal patterns of neutral and adaptive genetic variation in the endangered African wild dog (*Lycaon pictus*). Mol. Ecol. 2012; 21; 1379–1393. https://doi.org/10.1111/j.1365-294X.2012.05477.x PMID: 22320891

6. Guichoux E, Lagache L, Wagner S, Chaumel P, Ger PLE, Lepais O, et al. Current trends in microsatellite genotyping. Mol Ecol Res. 2011; 11:591–611. https://doi.org/10.1111/j.1755-0998.2011.03014.x PMID: 21565126

7. Guo R, Mao YR, Cai JR, Wang JY, Wu J, Qiu YX. Characterization and cross-species transferability of EST–SSR markers developed from the transcriptome of *Dysosma versipellis* (Berberidaceae) and their application to population genetic studies. Mol Breed. 2014; 34: 1733–1746. https://doi.org/10.1007/s11032-014-0134-z

8. Li DJ, Deng Z, Qin B, Liu XH, Men ZH. *De novo* assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). BMC Genom. 2012; 13:192. https://doi.org/10.1186/1471-2164-13-192 PMID: 22607098

9. Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, et al. Using Next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot. 2012; 99: 193–208. https://doi.org/10.3732/ajb.1100394 PMID: 22186186

10. Liu ZP, Chen TL, Ma LC, Zhao ZG, Zhao PX, Nan ZB, et al. Global transcriptome sequencing using the Illumina platform and the development of EST-SSR markers in autotetraploid alfalfa. PLoS One. 2013b; 8: e83549. https://doi.org/10.1371/journal.pone.0083549 PMID: 24349529

11. Zheng XF, Pan C, Diao Y, You YN, Yang CZ, Hu ZL. Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). BMC Genom. 2013; 14:490. https://doi.org/10.1186/1471-2164-14-490 PMID: 23870214

12. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29: 644–652. https://doi.org/10.1038/nbt.1883 PMID: 21572440

13. Wu J, Cai C, Cheng F, Cui H, Zhou H. Characterization and development of EST-SSR markers in tree peony using transcriptome sequences. Mol Breed. 2014; 34: 1853–1866. https://doi.org/10.1007/s11032-014-0144-x

14. Chen LY., Cao YN., Yuan N. et al. 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, 187. https://doi.org/10.1007/s11032-015-0379-1

15. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch SR. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 2001; 11:1441–1452. https://doi.org/10.1101/gr.184001 PMID: 11483586

16. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 2003; 106: 411–422. https://doi.org/10.1007/s00122-002-1031-0 PMID: 12589540

17. Yu JK, Dake TM, Singh S, Benscher D, Li W, Gill B, et al. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. Genome. 2004; 47:805–818. https://doi.org/10.1139/g04-057 PMID: 15499395

18. Guo X, Li Y, Li CF, Luo HM, Wang LZ, Qian J, et al. Analysis of the *Dendrobium officinale* transcriptome reveals putative alkaloid biosynthetic genes and genetic markers. Gene. 2013; 527: 131–138. https://doi.org/10.1016/j.gene.2013.05.073 PMID: 23756193

19. Li CF, Zhu YJ, Guo X, Sun CH, Luo HM, Song JY, Li Y, Wang LZ, Qian J, Chen SL. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* C. A. Meyer. BMC Genom.2013; 14(245): 1–11. https://doi.org/10.1186/1471-2164-14-245 PMID: 23577925

20. Hwang HS, Lee H, Choi YE. Transcriptomic analysis of Siberian ginseng (*Eleutherococcus senticosus*) to discover genes involved in saponin biosynthesis. BMC Genom. 2015; 16(180): 1–12. https://doi.org/10.1186/s12864-015-1357-z PMID: 25888223

21. Zhang GH, Ma CH, Zhang JJ, Chen JW, Tang QY, He MH, et al. Transcriptome analysis of *Panax vietnamensis* var. *fuscidicus* discovers putative ocotillol-type ginsenosides biosynthesis genes and genetic

markers. BMC Genom. 2015; 16:159:1–19. https://doi.org/10.1186/s12864-015-1332-8 PMID: 25765814

22. Liu TM, Zhu SY, Tang QM, Chen P, Yu YT, Tang SW. *De novo* assembly and characterization of transcriptome using Illumina paired-end sequencing and identification of CesA gene in ramie (Boehmeria nivea L. Gaud). BMC Genom. 2013a; 14:125. https://doi.org/10.1186/1471-2164-14-125 PMID: 23442184

23. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 2003; 19:651–652. https://doi.org/10.1093/bioinformatics/btg034 PMID: 12651724

24. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005; 21(18): 3674–3676. https://doi.org/10.1093/bioinformatics/bti610 PMID: 16081474

25. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 1999; 132:365–386. https://doi.org/10.1385/1-59259-192-2:365

26. Steffens DL, Sutter SL, Roerner SC. An alternate universal forward primer for improved automated DNA sequencing of M13. Biotechniques. 1993; 15: 580–581. PMID: 8251153

27. Kalinowski ST, Taper ML, Marshall TC. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol. 2007; 16: 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x PMID: 17305863

28. Goudet J. Fstat, a program to estimate and test gene diversities and fixation indices. Version 2.9.3.2001. http://www2.unil.ch/popgen/softwares/fstat.htm

29. Jiang FX, Wei PW, Kou YP, Jia Y, Chen QB. Study on the Characterization of the Transcriptome of *Anoectochilus formosanus* Taiwan. Mol Plant Breed. 2015; 13(12): 2743–2753.

30. Shen YM, Huang CQ, Liu ZY. Study on quality standard of imitation ecological planting *Anoectochilus roxburghii* under forest covering. China Tradit Herb Drugs. 2018; 49 (02): 450–454.

31. Zhang FS, Lv YL, Zhao Y, Guo SX. Promoting role of an endophyte on the growth and contents of kinsenosides and flavonoids of *Anoectochilus formosanus* Hayata, a rare and threatened medicinal Orchidaceae plant. J Zhejiang Univ Sci, B. 2013; 9: 785–792. https://doi.org/10.1631/jzus.B1300056 PMID: 24009198

32. Zou FX, Xu W, Huang ZH, Zhang X, Chen SH, Lin Y, et al. Analysis of transcriptome sequencing and related genes of flavonoid biosynthesis from *Anoectochilus roxburghi*. J China Pharm Univ, 2019, 50 (1): 66–74.

33. Kumpatla SP, Mukhopadhyay S. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome. 2006; 48(6):985–998. https://doi.org/10.1139/g05-060 PMID: 16391668

34. Koilkonda P, Sato S, Tabata S, Shirasawa K, Hirakawa H, Sakai H, et al. Large-scale development of expressed sequence tag-derived simple sequence repeat markers and diversity analysis in *Arachis* spp. Mol Breed. 2012; 30:125–138. https://doi.org/10.1007/s11032-011-9604-8 PMID: 22707912

35. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 2005; 23:48–55. https://doi.org/10.1016/j.tibtech.2004.11.005 PMID: 15629858

36. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 2004; 21(6):991–1007. https://doi.org/10.1093/molbev/msh073 PMID: 14963101

37. Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 2000; 10:72–80. https://doi.org/10.1101/gr.10.1.72 PMID: 10645952

38. Xia W, Xiao Y, Liu Z, Luo Y, Mason AS, Fan HK, et al. Development of gene-based simple sequence repeat markers for association analysis in *Cocos nucifera*. Mol Breed. 2014; 34:525–535. https://doi.org/10.1007/s11032-014-0055-x

39. Templeton AR, Shaw K, Routman EJ, Davis SK. The genetic consequences of habitat fragmentation. Ann MO Bot Gard, 1990; 77: 13–27. https://doi.org/10.2307/2399621

40. Young A, Boyle T, Brown T. The population genetic consequences of habitat fragmentation for plants. Trends Ecol Evol. 1996; 11(10): 413–418. https://doi.org/10.1016/0169-5347(96)10045-8 PMID: 21237900

41. Charlesworth B, Charlesworth D. Inbreeding depression and its evolutionary consequences. Ann Rev Ecol Syst. 1987; 18: 237–268. https://doi.org/10.1146/annurev.ecolsys.18.1.237

42. Lacy RC, Lindenmayer D. A simulation study of the impacts of population subdivision on the mountain brushtail possum *Trichosurus caninus* Ogilby (Phalangeridae: Marsupialia), in South-Eastern Australia. II Loss of genetic variation within and between subpopulations. Biol Conserv. 1995; 73: 131–142. https://doi.org/10.1016/0006-3207(95)90037