Evolutionary Bioinformatics
Volume 14: 1–6
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176934318790263

Transcriptome Analysis and Microsatellite Markers Development of a Traditional Chinese Medicinal Herb *Halenia elliptica* D. Don (Gentianaceae)

Mingliu Yang[1], Nanyu Han[1], Heng Li[2] and Lihua Meng[1]

[1]Key Laboratory of Yunnan for Biomass Energy and Biotechnology of Environment, Key Laboratory of Ecological Adaptive Evolution and Conservation on Animals-Plants in Southwest Mountain Ecosystem of University in Yunnan Province, School of Life Sciences Yunnan Normal University, Kunming, P. R. China. [2]Economic and Management, Luoyang Institute of Science and Technology, Luoyang, P. R. China.

ABSTRACT: *Halenia elliptica* is a popular Chinese medicinal herb that is used to treat jaundice disease and virus hepatitis, and its wild populations have been reduced significantly due to overharvesting recently. However, effective conservation could not be implemented because of the lack of genomic information and genetic markers. In this study, a de novo transcriptome of *H elliptica* was sequenced using the NGS Illumina, and 132 695 unigenes with the length >200 bp (base pairs) were obtained. Among them, a total of 32 109 unigenes were scanned to develop simple sequence repeats (SSRs). Based on NCBI (National Center for Biotechnology Information) nonredundant database (Nr), these SSR sequences were annotated and assigned into gene ontology categories. In addition, we designed 126 pairs of SSR primers for polymerase chain reaction amplification, of which 12 pairs were identified to be polymorphic among 40 individuals from 8 populations. We then used the 12 polymorphic SSRs to construct a UPGMA dendrogram of the 40 individuals. In addition, a significant correlation between the genetic relationship and the geographic distance was found, suggesting a phylogeographic structure in *H elliptica*. Moreover, 2 of these SSRs were also successfully amplified in a related species *Veratrilla baillonii*, suggesting their cross-species transferability. Generally, the SSR markers with high polymorphisms identified in this study provide valuable genetic resources and represent an initial step for exploring the genetic diversity and population histories of *H elliptica* and its related species.

KEYWORDS: transcriptome, SSR, ***Halenia elliptica***, polymorphisms

RECEIVED: March 13, 2018. **ACCEPTED:** June 18, 2018.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the National Natural Science Foundation of China (31460096).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Lihua Meng, Key Laboratory of Yunnan for Biomass Energy and Biotechnology of Environment, Key Laboratory of Ecological Adaptive Evolution and Conservation on Animals-Plants in Southwest Mountain Ecosystem of University in Yunnan Province, School of Life Sciences Yunnan Normal University, Kunming 650500, Yunnan, P. R. China. Email: menglihua@mails.ucas.ac.cn

Introduction

Over the past decade, life sciences were greatly advanced based on the genome sequencing technologies, especially the next-generation sequencing (NGS) that provides a strategy of a low cost in sequencing and large quantities of genomic data. Based on NGS, a great number of sequenced genomes have been obtained in a short time, which enhanced our understanding on variations of genome sequence.[1] RNA-Seq of NGS is advantageous over chip technology on the digital region and can produce explicit transcriptome data for nonmodel species. The genome-scale transcriptome analysis is powerful in non-model species by revealing differential expressions of genes in time and spaces, determining the genetic basis of specific phenotypes, and outlining genomic diversity.[2,3] In addition, a lot of simple sequence repeat (SSR) markers can be rapidly developed based on the genome-scale transcriptome analysis, which would be of great help in analyzing population genetic structure.[4,5]

The SSRs consist of short tandem repeats of 1 to 6 bp (base pair) nucleotides and are abundant in protein-coding and non-coding regions. The SSRs are highly diverse, codominant, and stable and thus were extensively used in many research subjects, such as evolutionary biology, population genetics, and conservation genetics.[6] In the past, it takes long time and high cost to obtain SSR markers, whereas RNA-Seq makes it easy to develop a great deal of SSR markers in the present time,[3,7,8] which promoted research works in the genetic diversity and evolutionary biology.[9]

Halenia elliptica D. Don, a biennial herb in the Gentianaceae family, is a popular Chinese medicinal herb that is widely used to treat jaundice disease and virus hepatitis. This species is mainly distributed at elevations ranging from 700 to 4000 m in Yunnan, Sichuan, Qinghai, and Tibet.[10] Due to its effective therapeutic effects, *H elliptica* was overexploited, leading to a decrease in the population size and genetic diversity in recent years. It is difficult to propose effective conservation methods without genome information and genetic markers.

In this study, we sequenced a de novo transcriptome for *H elliptica* on the Illumina platform and assembled the transcriptome sequences with software Trinity. As far as we know, this is the first exhibit of transcriptome results for *H elliptica*. In addition, we screened SSR markers in the transcriptome sequences and randomly selected markers to verify their amplification and polymorphism. The transcriptome sequence and polymorphic SSR markers developed in this work are believed to provide valuable genetic resources to study genetic diversity and

Table 1. Summary of assembly and annotation results for *Halenia elliptica* using Trinity.

RESULTS	NUMBER
Total no. of raw reads	19 668 659
Total no. of clean reads	19 426 614
Total no. of contigs	158 076
Total size of contigs, bp	117 982 598
Mean length of contigs	746
N50 value of contigs	1230
Length range of contigs	201-18 947
Total no. of unigenes	132 695
GC content	42.1%
Total no. of identified SSRs	32 109
SSR-containing sequences with BLASTX hit	17 973 (56.0%)
SSR-containing sequences with annotation	13 825 (43.1%)

population demographic history of *H elliptica* and its related species.

Materials and methods

Plant material

On July 2016, the fresh leaves of 10 *H elliptica* individuals were collected from Shangri-La in northwest Yunnan (28°31′0″N, 99°57′0″E, alt. 4514 m) and were kept immediately and separately in liquid nitrogen. In addition, a total of 40 individuals from 8 populations (Supplementary Table 1) were sampled and the leaves were stored in silica gel for polymorphic SSR markers validation.

RNA extraction and sequencing

We extracted the total RNA of each individual with a CTAB method[11] and measured the integrity of the RNA samples using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). To satisfy the criteria for RNA sequencing, equal amount of RNA from each replicate RNA was pooled together. We constructed the complementary DNA library using poly-A–enriched RNA method and fragmented the messenger RNA with fragmentation buffer based on Illumina protocols (San Diego, CA, USA). Random hexamer primers were used to synthesize the double strands. The short fragments were purified using the QIAquick PCR Purification Kit (Qiagen Inc., Courtaboeuf, France). Ultimately, the purified DNA libraries were first amplified by polymerase chain reaction (PCR) and then sequenced on Illumina HiSeq 2000 platform.

De novo assembly

The reads with many ambiguous bases (>8) and with more than 50% low-quality bases (quality score ⩽5) in raw reads were filtered out using *Perl* scripts. The transcriptome sequences of *H elliptica* were assembled using Trinity software with default parameters.[12]

SSR locus search, primer design, and validation

To detect the potential SSR loci, all the contigs were scanned by MicroSAtellite software (MISA, http://pgrc.ipk-gatersleben.de/misa).[13] In general, the SSR locus search minimum requirements were 5 repeats for the simple motifs and 3 repeats for the complex motifs. In this study, we set the minimum repeat unit as 10 for mononucleotides, 6 for dinucleotides, and 5 for trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides in MISA. The SSR primer pairs were designed with Primer 3.0.[14] The designed primer pairs in the study were excellent, and each target SSR was required to contain at least 5 repeats, with the length of PCR products ranging from 80 to 500 bp. Based on the above criteria, 126 primer pairs were randomly synthesized for validating SSR locus (Supplementary Table 2), of which 62 primer pairs were successful in the PCR products. Then, the successful primer pairs in PCR amplification were used to detect the polymorphism among 40 individuals from 8 populations (Supplementary Table 1). Polymerase chain reaction was performed in a 25-µL volume, and the PCR reaction program was set as the following conditions: (1) DNA initial denaturation was 4 minutes at 94°C, 35 cycles of 1 minute 30 seconds at 94°C; (2) the annealing temperature ranged from 50°C to 60°C for 50 seconds, following 72°C for 45 seconds; and (3) an extension was 7 minutes at 72°C. The PCR products were sequenced on the ABI 3730 genetic analyzer (Applied Biosystems, Foster City, CA). The statistics of polymorphic SSR loci were calculated using POPGEN v1.32.[15]

Functional annotation for contigs containing SSRs

All the SSR-containing contigs were used to search objective sequences in the NCBI's NR protein database using BLASTx with the E-value threshold setting as 1e−6. The contigs were assigned with gene names according to best BLASTx hits. Functional annotation of contigs was conducted by the program Blast2GO.[16] Functional categories were classified with the program WEGO.[17]

Genetic analyses in populations

We conducted phylogenetic analysis of the 40 individuals from 8 populations with the 12 primer pairs and used software MEGA6[18] to construct the dendrogram tree using the UPGMA method.

Results and Discussion

De novo assembly of H elliptica

A total of 19 668 659 raw reads data were generated by the Illumina HiSeq sequencer. After all adaptor sequences were

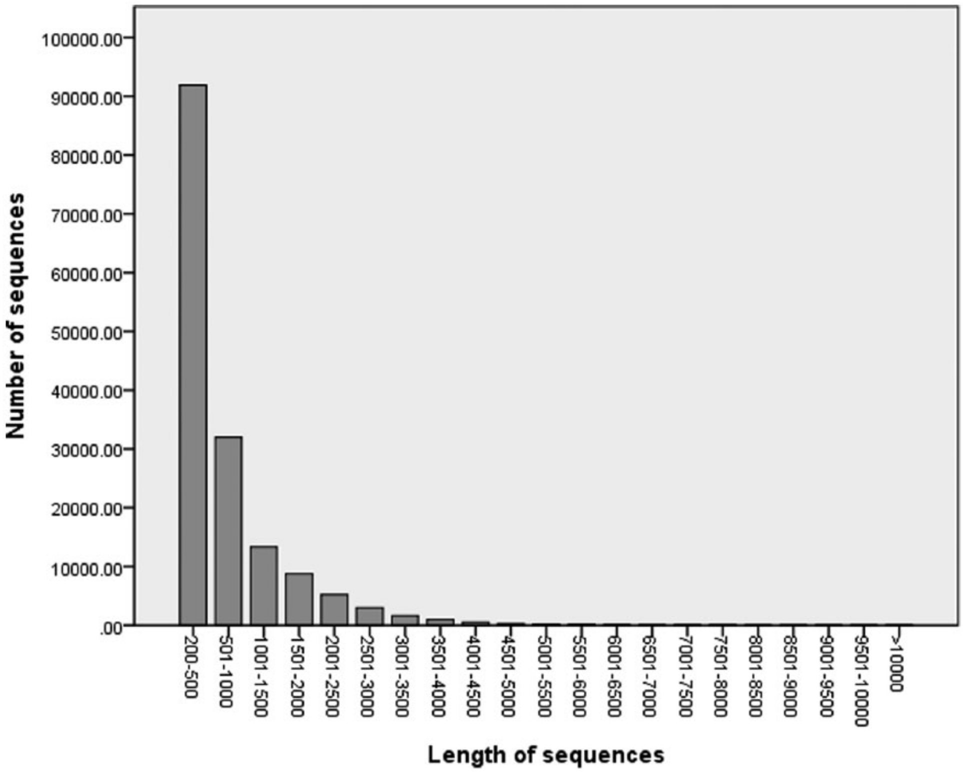

Figure 1. Number of sequences for all 158 076 transcriptome contigs for *Halenia elliptica*.

Table 2. Frequency of mono- to hexanucleotide repeat motifs in *Halenia elliptica*.

REPEATS	COUNTS
Mononucleotide	21 013
A/T	20 344
C/G	669
Dinucleotide	4718
AC/GT	678
AG/CT	900
AT/AT	3090
CG/CG	50
Trinucleotide	5569
AAC/GTT	365
AAG/CTT	1005
AAT/ATT	940
ACC/GGT	643
ACG/CGT	210
ACT/AGT	87
AGC/CTG	700
AGG/CCT	621
ATC/ATG	608
CCG/CGG	390
Tetranucleotide	339
Pentanucleotide	235
Hexanucleotide	235

removed, the ambiguous and low-quality sequences were filtered, a total of 19 426 614 RNA-Seq clean reads remained for further analysis, which had been deposited into the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/) with accession number SRP126366. These clean reads were used to assemble the contigs with Trinity software, which generated 158 076 contigs with the length ranging from 201 to 18 947 bp. The median length and the N50 value of the contigs are 450 and 1230 bp, respectively. The high N50 value suggested the high quality of the assembly. The GC content of our contigs is 42.1% (Table 1). With the increase in contig length, the frequency of the contigs decreased, suggesting a power law–like distribution. The contigs with lengths ranging from 200 to 500 bp were dominant, making up 58.14% of the total contigs (Figure 1), as detected in *Veratrilla baillonii*[4] and *Gentiana straminea*[19] in the same family of Gentianaceae.

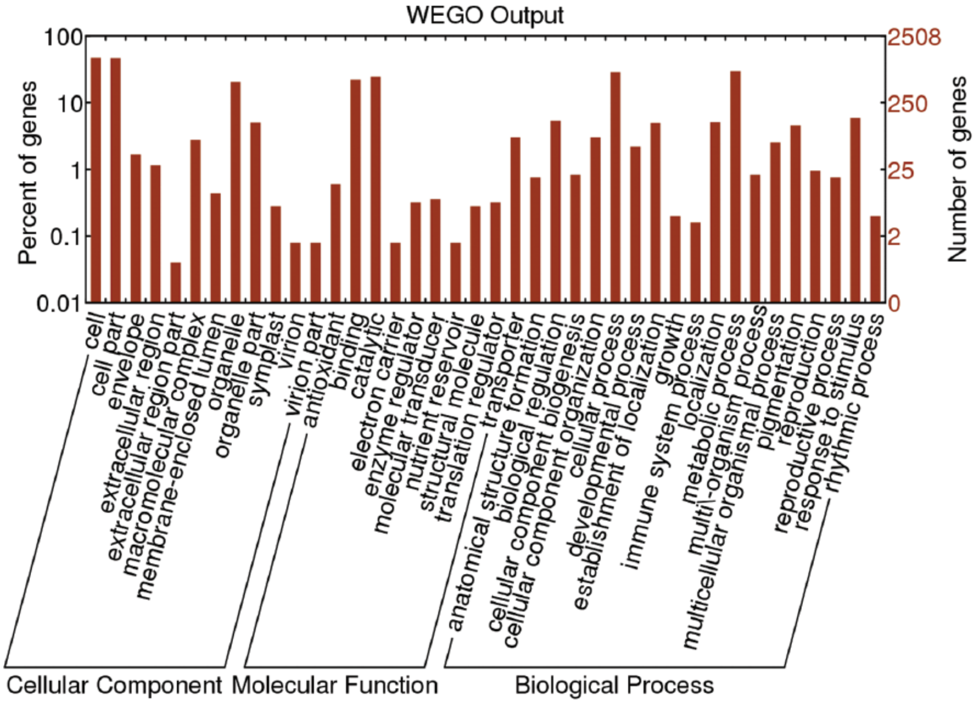

Figure 2. GO classification of SSRs in coding regions. GO indicates gene ontology; SSRs, simple sequence repeats.

Validation and distribution of SSRs

Using MISA software, we analyzed the contigs, and 32 109 SSRs of total 158 076 contigs were validated. The number of SSRs identified in *H elliptica* was higher than that in *G straminea* (14 561), but lower than that in *V baillonii* (40 885). The density of SSRs for *H elliptica* was 244.3 per MB, similar to *V baillonii* (243.3).

Based on different sizes, SSR loci were classified into 2 categories of genetic markers. Class I was hypervariable markers with the length of SSRs more than 20 bp, and Class II was potentially variable markers with the length of SSRs ranging from 12 to 20 bp. Because of the large sizes and long repeats, SSRs of Class I generally have much information and high polymorphism, which are beneficial in developing SSR markers. Owing to the small sizes, SSRs of Class II are less variable and thus are difficult to find polymorphism with SSR markers. In *H elliptica*, 23.9% of the SSRs were categorized as Class I and 76.1% as Class II. The proportion of Class I in *H elliptica* was higher than that in *V baillonii* (6.8%), but the fraction of SSRs was still deficient for further development of efficient and polymorphic SSR markers.

The detailed information of SSRs with different repeat styles is showed in Table 2. The result suggested the dominance of the SSRs with mononucleotide motifs, which accounted for 65.4% of the total. Without regard to the mononucleotide, trinucleotide, and dinucleotide motifs were predominant in quantity, accounting for 50.2% and 42.5% among the rest of these motifs, respectively. The total number of tetranucleotide, pentanucleotide, and hexanucleotide motifs was 2.8% of the total SSRs, close to *V baillonii* (2.7%). Moreover, the mode of SSRs distribution in *H elliptica* is similar to that of *V baillonii*. In mononucleotide motifs, the frequency of $(A/T)_n$ type was 96.8%, a situation found in many plant species.[4,20] In dinucleotide repeat motifs, the frequencies of AT/AT, AG/CT, and AC/GT were 65.5%, 19.1%, and 14.4%, respectively. In contrast, the content of CG/CG was lower, contributing to 1% only in *H elliptica*. These results are similar to that of Gentianaceae family and also in accordance with other dicot genomes in which A/T-rich repeats in trinucleotide motifs were frequent.[21] In general, AAC/GTT, AAG/CTT, and AAT/ATT were extensive in dicot genomes,[3,8] and our results are consistent with this conclusion. In *H elliptica*, the repeats AAC/GTT, AAG/CTT, and AAT/AAT were dominant with the frequencies of 6.6%, 18.0%, and 16.9%, respectively, and the total frequency of the trinucleotide motifs amounted to 41.5%.

In general, the feature of SSRs distribution in *H elliptica* was similar to that of *V baillonii*, indicating the close phylogenetic relationship and high similarity in transcriptome level between the 2 species. However, the similarity between the 2 species may also suggest their similar evolutionary histories because *H elliptica* originated from the East Asia[22] and *V baillonii* are restricted in the Hengduan Mountains.[10]

Functional annotation based on SSR-containing coding sequences of H elliptica

Using BLASTx, a lot of the SSR-containing contigs (17 973) were detected to have no less than one hit in the NCBI's NR

Table 3. Results of primer screening through 40 diversified accessions in *Halenia elliptica*.

LOCUS	REPEAT	FORWARD PRIMER (5′ -3′)	REVERSE PRIMER (5′ -3′)	TA, °C	SIZE, BP	NA	NE	HO	HE	PIC
FR11	(AAAGAA)11	TCCAGTTTGTTTTCTTGGGC	AATTGAAGCGTGGAAATTGG	56	496–538	6	1.837	0.100	0.383	0.7263
FR166	(CTCTTC)8	ATGAAGGTTGAGCTTGGTGG	AGGTGTGGTTGGACTTGGAC	54	219–261	6	1.821	0.200	0.378	0.7852
FR197	(ATT)14	TTGCCTCATTCCTCTCTCGT	GGGTGTTCTCCCTTCTTTTT	49.3	203–221	3	1.027	0.025	0.023	0.2019
FR200	(ATCC)6	TACTTCCCGAAATACCC	ACCTCCATTCTTGATAG	43.5	179–191	3	1.233	0.000	0.140	0.1736
FR202	(CATA)6	CCTTCTTTTTTTTCTTC	ATCCTCTGGAGCGTTAT	47	411–459	10	2.231	0.175	0.473	0.7099
FR248	(TGC)10	TTGGTTGATGACTCG	CAATGACTGGGGCTA	49.3	375–387	5	1.208	0.025	0.133	0.4980
FR265	(GGAA)6	AAAGTGTCCATCAAATCA	CCGTTCAGTTCACAATCC	48.5	328–336	3	1.086	0.025	0.063	0.0714
FR268	(AAAT)6	AGAAACAGAGAGACGAGG	ATAAGATGGGTAAGAGGC	57.8	368–404	6	2.554	0.400	0.530	0.7398
FR283	(TC)12(TA)10	CCCAAATGCCATAGTG	AGGAAGGGAAAACAGA	52	125–131	4	1.686	0.400	0.343	0.5638
FR288	(CGG)7	TAACGAATGAAGACACG	ATCAGGAAGACTATGCT	52	184–190	3	1.208	0.025	0.133	0.3764
FR295	(ACC)9	ACATTCTCGTAAGTAT	CAACCAGTATTCGGCT	52	285–297	5	1.204	0.025	0.143	0.4234
FR299	(AT)13	GTAACAAGAAGAGAAGG	GATAAATGGGAAGTAGA	52	175–191	5	1.567	0.150	0.330	0.5498

Abbreviations: He, expected heterozygosity; Ho, observed heterozygosity; Na, number of alleles; Ne, effective number of alleles; PIC, polymorphism information content; size, size of cloned allele; Ta, annealing temperature.

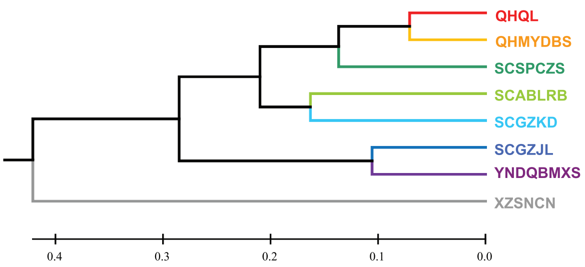

Figure 3. UPGMA dendrogram constructed among 40 individuals from 8 populations based on 12 SSR markers developed in this study. SSR indicates simple sequence repeats.

protein database, but the percentage (56.0%) was lower than that of *V baillonii* (70.7%). A further analysis of the contigs was implemented by the program Blast2GO. The contigs were assigned with gene names based on best BLASTx hits, and 13 825 SSR sequences were annotated. We conducted WEGO[17] to achieve functional categories. The contigs were classified into 3 categories of gene ontology terms, respectively, as cellular component, molecular function, and biological processes (Figure 2). Within the cellular component category, the cell and cell part were the most abundant types. Within the molecular function category, catalytic activity was the most dominant group, followed by binding. For the biological processes category, cellular process and metabolic process were the most common.

Polymorphism of SSR markers and phylogenetic analysis

To obtain polymorphic SSR markers, we designed 126 pairs of SSR primers for PCR amplification in 8 populations, and the PCR amplification of the 62 SSRs pairs was successful. The validated primers were used to screen genetic polymorphism of *H elliptica* with 40 individuals (Supplementary Table 1). And 12 pairs of these SSRs with polymorphism were found. The number of alleles per locus varied from 3 to 10, and the expected heterozygosity ranged from 0.023 to 0.530, whereas the observed heterozygosity varied between 0.000 and 0.400 (Table 3). Polymorphism information content values of the SSR markers varied between 0.0714 and 0.7852. Then, we used UPGMA to construct a phylogenetic tree of the 40 individuals. There was a significant correlation between the genetic relationship and the geographic distance in the phylogenetic tree of *H elliptica* (Figure 3), suggesting a phylogeographic structure in *H elliptica*. Further study with more populations is required to reveal the dynamic evolutionary history of *H elliptica*.

The 12 SSRs validated in this study exhibited high-quality and high genetic polymorphism, which allows us to analyze the genetic diversity and dynamic evolutionary history of *H elliptica*. The further analysis will provide us conservation strategies for this traditional medicinal plant. In addition, 2 pairs of these SSRs primers are also available in *V* *baillonii*, which suggests their transferability in the related species of *H elliptica*.

Conclusions

In this study, the de novo transcriptome for *H elliptica* was determined with RNA-Seq. A number of microsatellite markers were identified and 126 pairs of SSR primers were designed for PCR amplification. We found 12 SSR markers to be polymorphic, which can be used for future studies in *H elliptica*. The SSR markers with polymorphism identified in this study provide valuable genetic resources and represent an initial step for exploring the genetic diversity and population history of *H elliptica* and its related species.

Acknowledgements

The authors are grateful to Dr Yuanwen Duan for the field sampling and Dr Dongrui Jia for English editing.

Author Contributions

MY conducted the sample collections, the laboratory experiments, and statistical analyses. NH and HL assisted with bioinformatics tools. LM designed the study, conducted statistical analyses, and drafted the manuscript. All authors read and reviewed the final manuscript.

REFERENCES

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next generation sequencing technologies. *Nat Rev Genet*. 2016;17:333–351.
2. Colburn BC, Mehlenbacher SA, Sathuvalli VR. Development and mapping of microsatellite markers from transcriptome sequences of European hazelnut (*Corylus avellana* L.) and use for germplasm characterization. *Molec Breeding*. 2017;37:16.
3. Xu M, Liu X, Wang JW, et al. Transcriptome sequencing and development of novel genic SSR markers for *Dendrobium officinale*. *Molec Breeding*. 2017;37:18.
4. Wang L, Wang ZK, Chen JB, et al. De novo transcriptome assembly and development of novel microsatellite markers for the traditional Chinese medicinal herb, *Veratrilla baillonii* Franch (Gentianaceae). *Evol Bioinform*. 2015;11:39–45.
5. Rahemi A, Fatahi R, Ebadi A, et al. Genetic diversity of some wild almonds and related Prunus species revealed by SSR and EST-SSR molecular markers. *Plant Syst Evol*. 2012;298:173–192.
6. Lawson MJ, Zhang L. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol*. 2006;7:R14.
7. Zhang L, Yan HF, Wu W, et al. Comparative transcriptome analysis and marker development of two closely related Primrose species (*Primula poissonii* and *Primula wilsonii*). *BMC Genomics*. 2013;14:329.
8. Silva PIT, Martins AM, Gouvea EG, et al. Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. *BMC Genomics*. 2013;14:17.
9. Wee AKS, Takayama K, Chua JL, et al. Genetic differentiation and phylogeography of partially sympatric species complex *Rhizophora mucronata* Lam. and *R. stylosa* Griff. using SSR markers. *BMC Evol Biol*. 2015;15:57.
10. Ho TN, Pringle JS. Gentianaceae. In: Wu ZY, Raven PH, eds. *Floral of China* (vol. 16). Beijing, China and St. Louis, MI: Science Press and Missouri Botanical Garden; 1995:1–139.
11. Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol Biol Rep*. 1993;11:113–116.
12. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29:644–652.
13. Sharma PC, Grover A, Kahl G. Mining microsatellites in eukaryotic genomes. *Trends Biotechnol*. 2007;25:490–498.
14. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Bioinform Methods Protoc*. 2000;132:365–386.
15. Yeh FC, Yang R, Boyle TJ. *Popgene Version 1.32 Microsoft Windows-Based Freeware for Populations Genetic Analysis*. Edmonton, AB, Canada: University of Alberta; 1999.
16. Conesa A, Gotz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics*. 2008;2008:619832.
17. Ye J, Fang L, Zheng H, et al. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res*. 2006;34:W293–W297.
18. Tamura K, Stecher G, Peterson D, et al. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol*. 2013;30:2725–2729.
19. Zhou DW, Gao S, Wang H, et al. De novo sequencing transcriptome of endemic *Gentiana straminea* (Gentianaceae) to identify genes involved in the biosynthesis of active ingredients. *Gene*. 2016;575:160–170.
20. Gao Z, Wu J, Liu Z, et al. Rapid microsatellite development for tree peony and its implications. *BMC Genomics*. 2013;14:886.
21. Sonah H, Deshmukh RK, Sharma A, et al. Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. *PLoS ONE*. 2011;6:e21298.
22. Von Hagen KB, Kadereit JW. The diversification of *Halenia* (Gentianaceae): ecological opportunity versus key innovation. *Evolution*. 2003;57:2507–2518.