Genetics and Molecular Research

Development of EST-SSR markers in the relict tree *Davidia involucrata* (Davidiaceae) using transcriptome sequencing

Z.C. Long[1,2], A.W. Gichira[1,2], J.M. Chen[1], Q.F. Wang[1] and K. Liao[1]

[1]Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, China
[2]Life Science College, University of Chinese Academy of Sciences, Beijing, China

Corresponding author: K. Liao
E-mail: liaokuo@wbgcas.cn

Genet. Mol. Res. 15 (4): gmr15048539
Received February 11, 2016
Accepted March 28, 2016
Published October 17, 2016
DOI http://dx.doi.org/10.4238/gmr15048539

ABSTRACT. *Davidia involucrata*, reputed to be a "living fossil" in the plant kingdom, is a relict tree endemic to China. Extant natural populations are diminishing due to anthropogenic disturbance. In order to understand its ability to survive in a range of climatic conditions and to design conservation strategies for this endangered species, we developed genic simple sequence repeats (SSRs) from mRNA transcripts. In total, 142,950 contigs were assembled. Of these, 30,411 genic SSR loci were discovered and 12,208 primer pairs were designed. Dinucleotides were the most common (77.31%) followed by trinucleotides (16.44%). Thirteen randomly selected primers were synthesized and validated using 24 individuals of *D. involucrata*. The markers displayed high polymorphism with the number of alleles per locus ranging from 3 to

12 and the observed and expected heterozygosities ranging from 0.083 to 1.0 and 0.102 to 0.69, respectively. This large expressed sequence tag dataset and the novel SSR markers will be key tools in comparative studies that may reveal the adaptive evolution, population structure, and resolve the genetic diversity in this endangered species.

Key words: Adaptive genetic variation; *Davidia involucrata*; EST-SSR; Living fossil; Transcriptome sequencing

INTRODUCTION

Davidia involucrata Baill. (Davidiaceae), the dove tree, is a deciduous tree well known for its characteristic inflorescences that appear as doves perched on its branches. It is a valuable species for ornamental purposes in China, Europe, and North America (Manchester, 2002). *D. involucrata* is the only living species in its genus and is native to South Central and South West China, inhabiting the altitudinal range of 600-3200 m above sea level, with a mean annual rainfall of 600-1200 mm (Tang and Ohsawa, 2002; You et al., 2014). This species, purported to be a "living fossil" (Manchester, 2002), was listed as a rare species in the China Plants Red Data Book (Fu and Jin, 1992). The natural populations have been decreasing continuously, mainly due to anthropogenic interference, including habitat disturbance throughout most of its natural distribution range, increased logging, and collection of wild seeds by locals for commercial reasons (Luo et al., 2011; You et al., 2014).

Genetic diversity and phylogeographic studies have previously been conducted on *D. involucrata*, using various molecular markers, including allozymes (Peng et al., 2003), random-amplified polymorphic DNA (Song and Bao, 2004), inter-simple sequence repeats (Luo et al., 2011), amplified fragment length polymorphisms (Li et al., 2012a), chloroplast DNA (cpDNA) non-coding regions (Chen et al., 2015), and cpDNA and nuclear simple sequence repeats (nSSR) (Ma et al., 2015). Researchers have thus broadened our understanding of the evolutionary history of this species.

However, the markers used are not linked to any genic function and, therefore, they are less effective in defining the adaptive genetics and in outlining the effects of natural selection. Nuclear microsatellites have previously been developed for *D. involucrata* (Du et al., 2012; Li et al., 2012b; Tao et al., 2012). In this study, the objectives were to generate an expressed sequence tag (EST) dataset and to develop novel EST-SSR markers, to facilitate further molecular studies and conservation of this endangered species.

MATERIAL AND METHODS

RNA extraction

Four accessions of *D. involucrata* were selected for RNA extraction. Total RNA was extracted from young leaf tissues and was immediately frozen in liquid nitrogen. For every 100 mg tissue, 1 mL TRIzol reagent (Invitrogen, USA) was added and treated with RNase-free DNase I (TaKaRa Bio, Shandong, China) for 1 h at 37°C. The extracted RNA was then diluted in RNase-free water (Ambion, USA). From each sample, 1 µL was used to check the quality and concentration by

both NanoDrop (Thermo Scientific, USA) and Agilent Bioanalyzer 2100 (Agilent, USA). The four samples were pooled by mixing the total RNA at equal volumes. This step was done to ensure the quality of the transcriptional units and to enhance the downstream processes.

cDNA synthesis and sequencing

The working concentration for the cDNA synthesis was reduced to be in the order of 50-100 ng/μL. mRNA was isolated from the pooled total RNA and purified using Micropoly (A) Purist™ mRNA purification kit (Ambion™), following the manufacturer instructions. This step was followed by cDNA synthesis employing a slightly modified protocol based on Ng et al. (2005). To synthesize the first-strand cDNA, 10 μg mRNA template was utilized, using *Gsu*I-oligo as reverse transcriptome primers and 1000 U Superscript II reverse transcriptase (Invitrogen). The mixture was incubated at 42°C for 1 h. The cDNA was treated with sodium periodate (Sigma, USA) and the mRNA 5' oxide caps were biotinylated using Dynal M280 beads™ (Invitrogen). The biotin-linked mRNA/cDNA was further treated with alkaline lysis to release the first-strand cDNA. The complementary strand cDNA was synthesized using *Ex Taq*™ polymerase (TaKaRa Bio Inc., Shiga, Japan) and *Gsu*I restriction enzyme was used to cut-off poly-A tails. The double-stranded cDNA was then purified using the QIAquick PCR extraction kit (Qiagen, Hilden, Germany), followed by ligation of sequencing adaptors onto the fragments. In order to maximize the quality and to enhance the accuracy of the sequencing process, uniformity of the fragments was ensured by selecting a range of 300-500 bp. These fragments were then purified using Ampure® beads (Agencourt Canada, USA). The final step of polymerase chain reaction (PCR) was performed to enrich the fragments and to construct a library of transcripts for sequencing. Finally, the library was sequenced using Illumina HiSeq™ 2000 platform (Illumina Inc., CA, USA).

De novo assembly and functional annotation of unigenes

A stringent filtering process was implemented and the clean reads were assembled using the Trinity software (Grabherr et al., 2011). The EMBOSS software (Rice et al., 2000) was used to predict and identify any putative open-reading frames as well as untranslated regions within the transcripts. The predicted protein-coding sequences (unigenes) were aligned against the Swiss-Prot and TrEMBL protein databases using BLASTp, with the E-value set at $<1e^{-5}$, for authentication and annotation.

We used GoPipe (Chen et al., 2005) to assign gene ontology (GO) annotations. The metabolic pathways were constructed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2010). The conditions for bi-directional BLAST were also set at E-value $<1e^{-5}$. The unigenes were assigned KEGG orthology (KO) numbers and their involvement in the metabolic pathways was defined.

SSR mining, primer designing, and marker validation

SSR motifs were detected using MiSa (http://pgrc.ipk-gatersleben.de/misa/), a Perl language-based program. Mononucleotides were excluded in our search. The search criteria were set at 6, 5, 3, 3, and 3 for di-, tri-, tetra-, penta-, and hexanucleotide repeats, respectively.

We used the Primer 3.0 software (Rozen and Skaletsky, 2000) to design SSR primers under the following conditions; primer length was set between 18-22 bp, 40-60% G-C composition, 45°-60°C annealing temperature, and an expected product size of 100-300 bp.

Firstly, we randomly selected 18 SSR primer pairs and tested them for amplification and specificity using the four accessions of *D. involucrata*. Genomic DNA was extracted using MagicMag Genomic DNA Micro Kit (Sangon Biotech Co., Shanghai, China) following the manufacturer protocol. The 13 best primers were selected and the sense sequences were labeled with 6-FAM fluorescent dye on the 5'-end. These were used to genotype 24 accessions of *D. involucrata*, collected from three distant natural populations in Sichuan and Shanxi Provinces in China. The total 20 µL PCR mixture was composed of 50-80 ng/µL genomic DNA, 2.5 µL 10X *Taq* Buffer (with Mg^{2+}), 0.8 µL 10 mM dNTPs, 0.8 µL 10 mM each primer, 0.2 U *Taq* Polymerase enzyme, and double-distilled water.

The PCR protocol was set as follows: an initial denaturation step of 3 min at 95°C, followed by 30 cycles of 30 s denaturation at 95°C, 30 s annealing at 50°-55°C, and 30 s extension at 72°C. A final extension step of 10 min at 72°C was added at the end of the program. Quality was checked on a 1.5% agarose gel. The PCR products were separated using an ABI 3730 XL automated sequencer (TsingKe Biotech, Beijing, China) and visualized using the GeneScan system (Applied Biosystems, Foster, CA, USA). The number of alleles (N_A), observed heterozygosity (H_O), expected heterozygosity (H_E), for each of the EST-SSR markers were calculated using GenAlex 6.5 (Peakall and Smouse, 2012).

RESULTS

Illumina *de novo* sequencing and functional annotations and classification

A total of 17.4 million paired-end raw reads (3.4 billion bp) were obtained for *D. involucrata*. The reads were assembled into 209,238 contigs that were reduced to 142,950 after removal of low-quality redundant contigs. The contig length ranged from 201 to 45,506 bp, with an average of 491 bp. In total, 142,432 total protein coding sequences were predicted. Of these, 25,220 (17.7%) genes had a clear biological function. Of these, 9123 (6.4%) transcripts significantly matched 3373 GO terms. These were organized into three major categories; biological processes (28,730), molecular function (25,267), and cellular component (22,870) (Figure 1). Of the predicted proteins, 2496 had significant matches in the KEGG database and could be assigned to 25 major pathways organized into five main classifications. These included metabolism, genetic information processing, environmental information processing, cellular processes, and organismal systems (Figure 2). In general, metabolism (2090) was highly represented, followed by genetic information processes (1061). A number of unigenes were found to be active in more than one biological process. The specific pathways involving the majority of the unigenes included signal transduction mechanisms (6634) and general function (5618). Cell motility (69) was the least represented pathway.

EST-SSR marker development, characterization, and validation

From 142,950 assembled contigs and singletons, a total of 68,325 SSRs were

identified in 52,630 sequences. Of these, 11,471 sequences contained more than one SSR motif. Among the SSRs, 25,806 (37.8%) were dinucleotide repeats, 5600 (8.2%) were trinucleotide repeats, the rest (0.8%) were tetra-, penta-, and hexanucleotides. The A/T (50.4%) was the most abundant repeat motif, followed by AG/CT (23.7%) and AAG/CTT (2.1%). The frequency analyses of the developed SSRs are shown in Table 1. A total of 12,208 primers were designed from the 30,411 identified SSRs. Of these, 7887 primers were designed from dinucleotide repeats, 3472 were designed from trinucleotides, whereas the rest were designed from tetra-, penta-, and hexanucleotides. The N_A per locus ranged from 3 (Locus1) to 12 (Locus13), whereas H_O and H_E varied from 0.083 to 1.0 and 0.102 to 0.69, respectively (Table 2).

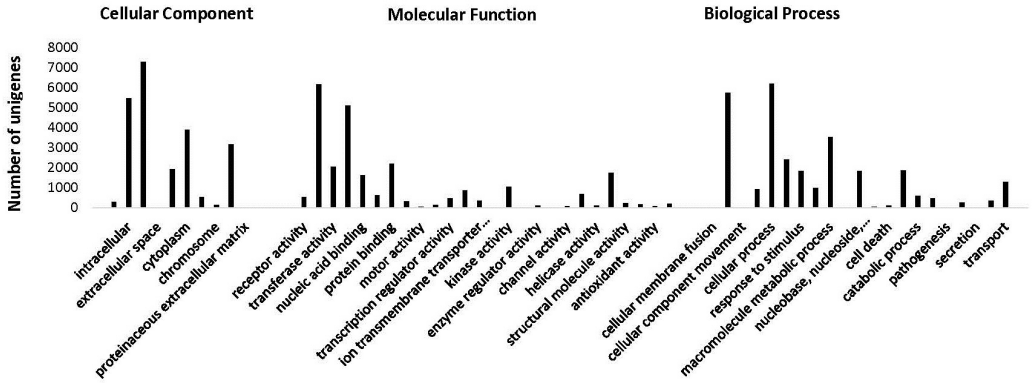

Figure 1. Classifications of assembled unigenes on gene ontology (GO). Of the 9123 predicted proteins, 3373 significantly matched GO annotations.

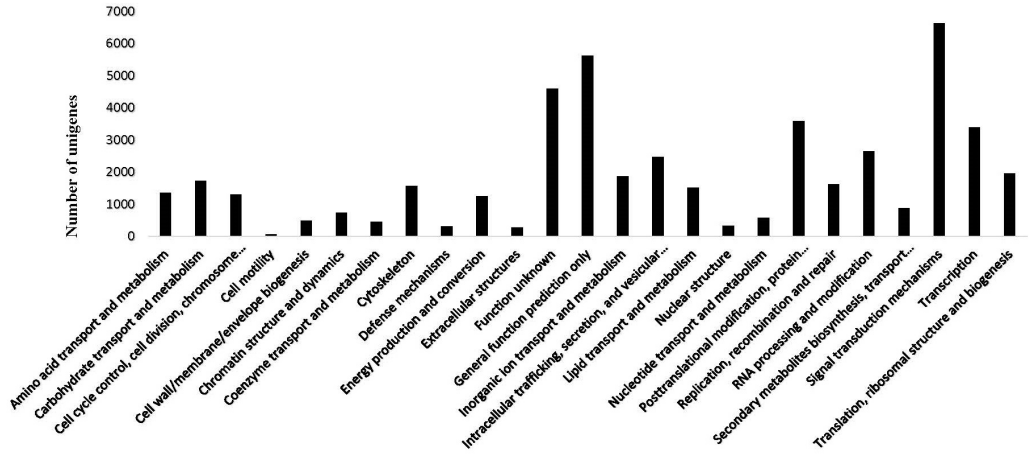

Figure 2. Kyoto Encyclopedia of Genes and Genome (KEGG) pathway assignment of 2496 unigenes to metabolic pathways, based on cellular process, genetic information processing, environmental information processing, metabolism, and organismal system.

Table 1. Type and number of SSR motifs in *Davidia involucrata*.

Type	Repeats	5	6	7	8	9	10	10	Total
Mono-repeats	A/T	-	-	-	-	-	12,652	21,763	34,415
	C/G	-	-	-	-	-	445	1498	1943
Di-repeats	AC/GT	-	1133	709	664	717	547	135	3905
	AG/CT	-	3232	2486	3521	4853	1908	225	16,225
	AT/AT	-	1481	1181	1094	1201	553	97	5607
	CG/CG	-	37	21	8	2	1	-	69
Tri-repeats	ATC/ATG	348	210	117	6	-	-	-	681
	AAG/CTT	852	373	197	3	-	-	1	1426
	AAT/ATT	356	231	70	3	-	-	-	660
	ACC/GGT	549	320	177	7	-	-	6	1059
	Others	1023	421	272	49	2	6	1	1774
Tetra-repeats	AAAT/ATTT	66	5	-	-	-	-	-	71
	ACAT/ATGT	68	1	-	-	-	-	-	69
	AAAG/CTTT	51	8	1	-	-	-	-	60
	Others	118	36	2	-	-	-	-	156
Penta-repeats	Penta-nucleotides	33	1	12	-	-	-	-	46
Hexa-repeats	Hexa-nucleotides	99	41	15	3	1	-	-	159
	Total	3563	7530	5260	5358	6776	16,112	23,726	68,325

Table 2. Characterization of 13 polymorphic EST-SSR markers in *Davidia involucrata*.

Locus	Primer sequence	Repeat motif	Tm (°C)	Allele range	N_A	H_O	H_E	GenBank accession No.
GTSSR2	F:ACATTACCGAGCCAAAGTGG R:AAAGGCAATAACAAGCCTGG	$(AAT)_5(TAA)_6$	55	246-258	1.7	0.250	0.229	KU525184
GTSSR3	F:TCATTGAGGCCACCCTTTAC R:ATGACTTGCCACCTTATGGC	$(TA)_7(TA)_6$	55	226-252	3.7	0.875	0.578	KU525185
GTSSR4	F:ACTGGACATGCCCTATCTGC R:TGTGATCGTCACAAGGAAGC	$(TC)_6(GC)_7$	55	208-224	2.3	0.250	0.359	KU525186
GTSSR5	F:AGGCCTTGCTTGAAACTAACA R:ATTGACATTTGGGTGATGGG	$(AC)_6(AC)_6$	53	139-168	2.0	0.292	0.258	KU525187
GTSSR6	F:GGCTTGACATCAGCACTTCA R:TGAGACTTGGGGACCTTTTG	$(AC)_6(AT)_7$	53	207-242	2.7	0.083	0.375	KU525188
GTSSR7	F:GGGGAGGGTACGGTAACAAT R:CAATTTCTTCCTCATCCCGA	$(ATG)_5(ACG)_5$	55	207-213	1.3	0.125	0.102	KU525189
GTSSR12	F:GCAATTGAGGCTGGAAACAT R:CCTTTGCGTCTCTCTTGGTC	$(CAC)_5(CAT)_5$	55	188-196	2.7	1.000	0.536	KU525190
GTSSR13	F:TCCTTGCAAACACACCATGT R:CTCGGAGTCCACTTTACTATCAA	$(GA)_8(GA)_7$	55	237-258	3.7	0.458	0.435	KU525191
GTSSR15	F:CCAAAAGGCCAACAACAACT R:CAGCAATTCCTCCAACACCT	$(GAA)_5(ATG)_5$	55	210-271	2.3	0.333	0.273	KU525192
GTSSR18	F:TGTTGGAGGAGGGGTAGAGA R:AAGAGGGAAAATTTGGGAGC	$(AAG)_6(GAT)_5$	55	163-179	4.0	0.500	0.641	KU525193
GTSSR23	F:GACCGTTAGGTTGATTGCGT R:ATTAGGGCCCCGAACATTAC	$(GT)_6(CT)_6$	55	234-240	1.7	0.208	0.174	KU525194
GTSSR24	F:GGCTGCATGAACACTGGATA R:TTAAATTGTGCATCTTAGTTGTGAA	$(AT)_6$	53	210-271	4.0	0.500	0.523	KU525195
GTSSR26	F:ATGAGTGCAAACCCTTTGGT R:ACATGCTTCAAAGATTGGGG	$(TTA)_6$	53	213-259	5.7	0.833	0.690	KU525196

DISCUSSION

High-throughput sequencing technology has emerged as a powerful approach. Among its numerous applications, it has significantly advanced the development of transcriptome-derived SSR markers in a number of plant and animal species, including tree peony and yellow catfish (Wu et al., 2014; Zhang et al., 2014). In this study, using the Illumina Hiseq-2000 platform, we were able to analyze transcriptome ESTs, generate 142,950 contigs, and design a higher number of potential SSR primers (12,208) compared to previous studies of *D. involucrata* (Du et al., 2012; Li et al., 2012b; Tao et al., 2012). The newly developed SSR

markers are located within the coding regions and are functionally linked to genes. EST-based SSR markers are best suited for use in gene targeting, QTL mapping, and in adaptive evolution studies. Dinucleotide repeat types were the most abundant (77.31%), followed by trinucleotides (16.44%), and tetranucleotides (0.98%), whereas penta- and hexanucleotides together accounted for 0.64%. The longest detected repeat motif was CTT with 17 reiterations that would not allow for primer design. The 13 randomly selected primers were successfully genotyped in 24 individuals of *D. involucrata* obtained from three distant populations. Expected product sizes of 100-300 bp of the individual primers were obtained.

In order to gain insight into the evolutionary history of *D. involucrata*, it is crucial to fully analyze its genetic diversity and reveal its phylogeographic patterns. The large number of polymorphic EST-SSR primers developed in this study will enhance molecular research on widely distributed natural populations of *D. involucrata*. The resulting information may reveal how *D. involucrata* survived through the Pleistocene glaciations and provide the foundation on which appropriate conservation measures may be taken.

Conflicts of interest

The authors declare no conflict of interest

ACKNOWLEDGMENTS

We thank Zhi-Yuan Du, Le-Na Li, and Hua Liu for their help in the laboratory. Research supported by grants from the National Natural Science Foundation of China (#31200170 and #31570220).

REFERENCES

Chen JM, Zhao SY, Liao YY, Gichira AW, et al. (2015). Chloroplast DNA phylogeographic analysis reveals significant spatial genetic structure of the relictual tree *Davidia involucrata* (Davidiaceae). *Conserv. Genet.* 16: 583-593. http://dx.doi.org/10.1007/s10592-014-0683-z

Chen Z, Xue C, Zhu S, Zho F, et al. (2005). GoPipe: streamlined gene ontology annotation for batch anonymous sequences with statistics. *Prog. Biochem. Biophys.* 32: 187-191.

Du YJ, Dai QY, Zhang LY, Qiu YX, et al. (2012). Development of microsatellite markers for the dove tree, *Davidia involucrata* (Nyssaceae), a rare endemic from China. *Am. J. Bot.* 99: e206-e209. http://dx.doi.org/10.3732/ajb.1100507

Fu LK and Jin JM (1992). China plant red data book - rare and endangered plants. Vol. 1. Science Press, Beijing.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29: 644-652. http://dx.doi.org/10.1038/nbt.1883

Kanehisa M, Goto S, Furumichi M, Tanabe M, et al. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38: D355-D360. http://dx.doi.org/10.1093/nar/gkp896

Li XP, Li ZL, He CL, Zhu WY, et al. (2012a). Genetic diversity of the endangered *Davidia involucrata* by AFLP analysis. *Acta Hort. Sin.* 39: 992-998.

Li Z, Wang C, Liu Y and Li J (2012b). Microsatellite primers in the Chinese dove tree, *Davidia involucrata* (Cornaceae), a relic species of the Tertiary. *Am. J. Bot.* 99: e78-e80. http://dx.doi.org/10.3732/ajb.1100365

Luo S, He Y, Ning G, Zhang J, et al. (2011). Genetic diversity and genetic structure of different populations of the endangered species *Davidia involucrata* in China detected by inter-simple sequence repeat analysis. *Trees* 25: 1063-1071. http://dx.doi.org/10.1007/s00468-011-0581-7

Ma Q, Du YJ, Chen N, Zhang LY, et al. (2015). Phylogeography of *Davidia involucrata* (Davidiaceae) inferred from cpDNA haplotypes and nSSR data. *Syst. Bot.* 40: 796-810. http://dx.doi.org/10.1600/036364415X689267

Manchester SR (2002). Leaves and fruits of *Davidia* (Cornales) from the Paleocene of North America. *Syst. Bot.* 27: 368-382.

Ng P, Wei CL, Sung WK, Chiu KP, et al. (2005). Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2: 105-111. http://dx.doi.org/10.1038/nmeth733

Peakall R and Smouse PE (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* 28: 2537-2539.

Peng YL, Hu YQ and Sun H (2003). Allozyme analysis of *Davidia involucrata* var. *vilmoriniana* and its biogeography significance *Acta Bot. Yunnanica* 25: 55-62.

Rice P, Longden I and Bleasby A (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16: 276-277. http://dx.doi.org/10.1016/S0168-9525(00)02024-2

Rozen S and Skaletsky H (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* 132: 365-386.

Song CW and Bao MZ (2004). Study on genetic diversity of RAPD mark for natural *Davidia involucrata* population. *Sci. Silvae Sin.* 40: 75-79.

Tang CQ and Ohsawa M (2002). Tertiary relic deciduous forests on a humid subtropical mountain, Mt. Emei, Sichuan, China. *Folia Geobot.* 37: 93-106. http://dx.doi.org/10.1007/BF02803193

Tao C, Yang Z, Lu R, Liu T, et al. (2012). Microsatellite markers for the relictual dove tree, *Davidia involucrata* (Cornaceae). *Am. J. Bot.* 99: e108-e110. http://dx.doi.org/10.3732/ajb.1100414

Wu J, Cai C, Cheng F, Cui H, et al. (2014). Characterisation and development of EST-SSR markers in tree peony using transcriptome sequences. *Mol. Breed.* 34: 1853-1866. http://dx.doi.org/10.1007/s11032-014-0144-x

You H, Fujiwara K and Liu Y (2014). A preliminary vegetation-ecological study of *Davidia involucrata* forest. *Nat. Sci.* 6: 1012-1029.

Zhang J, Ma W, Song X, Lin Q, et al. (2014). Characterization and development of EST-SSR markers derived from transcriptome of yellow catfish. *Molecules* 19: 16402-16415. http://dx.doi.org/10.3390/molecules191016402