A peer-reviewed version of this preprint was published in PeerJ on 15 June 2017.

Wang Y, Liu K, Bi D, Zhou S, Shao J. 2017. Characterization of the transcriptome and EST-SSR development in *Boea clarkeana*, a desiccation-tolerant plant endemic to China. PeerJ 5:e3422 https://doi.org/10.7717/peerj.3422

Characterization of the transcriptome and EST-SSR development in *Boea clarkeana*, a desiccation-tolerant plant endemic to China

Ying Wang [1, 2], **Kun Liu** [1, 2], **De Bi** [1], **Biao Shou Zhou** Corresp., [1, 3], **Wen Jian Shao** Corresp. [1, 2]

[1] College of Life Sciences, Anhui Normal University, Wuhu, Anhui, 中国

[2] Anhui Provincial Key Laboratory of the Conservation and Exploitation of Biological Resources, Anhui Normal University, Wuhu, Anhui, 中国

[3] College of Environmental Science and Engineering, Anhui Normal University, Wuhu, Anhui, 中国

Corresponding Authors: Biao Shou Zhou, Wen Jian Shao
Email address: zhoushoubiao@vip.163.com, 545491044@qq.com

Background. Resurrection plants constitute a unique cadre within angiosperms. *Boea clarkeana* Hemsl. (*Boea*, Gesneriaceae) is a desiccation-tolerant dicotyledonous herb that is endemic to China. Although research on angiosperms with DT could be instructive for crops, genomic resources for *B. clarkeana* remain scarce. In addition, transcriptome sequencing could be an effective way to study desiccation-tolerant plants. **Methods**. In the present study, we used the platform Illumina HiSeq™ 2000 and *de novo* assembly technology to obtain leaf transcriptomes of *B. clarkeana* and conducted a BLASTX alignment of the sequencing data and protein databases for sequence classification and annotation. Then, based on the sequence information obtained, we developed EST-SSR markers by means of EST-SSR mining, primer design and polymorphism identification. **Results.** A total of 91,449 unigenes were generated from the leaf cDNA library of *B. clarkeana* in this study. Based on a sequence similarity search with a known protein database, 72,087 unigenes were annotated. Among the annotated unigenes, a total of 71,170 unigenes showed significant similarity to known proteins of 463 popular model species in the Nr database, and 59,962 unigenes and 32,336 unigenes were assigned to GO classifications and COG, respectively. In addition, 44,924 unigenes were mapped in 128 KEGG pathways. Furthermore, a total of 7,610 unigenes with 8,563 microsatellites were found. Seventy-four primer pairs were selected from 436 primer pairs designed for polymorphism validation. SSRs with higher polymorphism rates were concentrated on dinucleotides, pentanucleotides and hexanucleotides. Finally, 17 pairs with highly polymorphic and stable loci were selected for polymorphism screening. There were a total of 65 alleles, with 2–6 alleles at each locus. Mainly due to the unique biological characteristics of plants, the H_E, H_O and PIC per locus were very low, ranging from 0 to 0.196, 0.082 to 0.14 and 0 to 0.155, respectively. **Discussion.** A substantial fraction

transcriptome sequences of *B. clarkeana* were generated in this study, which is the first molecular-level analysis of this plant. These sequences are valuable resources for gene annotation and discovery and molecular marker development. These sequences could also provide a valuable basis for the future molecular study of *B*. *clarkeana*.

Title: Characterization of the transcriptome and EST-SSR development in *Boea clarkeana*, a desiccation-tolerant plant endemic to China

Ying Wang[1,2], Kun Liu[1,2], De Bi[1], Biao Shou Zhou[1,3], Wen Jian Shao[1,2]

[1]College of Life Sciences, Anhui Normal University, Wuhu, Anhui, China

[2] Anhui Provincial Key Laboratory of the Conservation and Exploitation of Biological Resources, Wuhu, Anhui, China

[3] College of Environmental Science and Engineering, Anhui Normal University, Wuhu, Anhui, China

Corresponding author:

Biao Shou Zhou[1,3]

Email address: zhoushoubiao@vip.163.com;

Wen Jian Shao[1,2]

Email address: 545491044@qq.com.

Abstract

Background. Resurrection plants constitute a unique cadre within angiosperms. *Boea clarkeana* Hemsl. (*Boea*, Gesneriaceae) is a desiccation-tolerant dicotyledonous herb that is endemic to China. Although research on angiosperms with DT could be instructive for crops, genomic resources for *B. clarkeana* remain scarce. In addition, transcriptome sequencing could be an effective way to study desiccation-tolerant plants.

Methods. In the present study, we used the platform Illumina HiSeq™ 2000 and *de novo* assembly technology to obtain leaf transcriptomes of *B. clarkeana* and conducted a BLASTX alignment of the sequencing data and protein databases for sequence classification and annotation. Then, based on the sequence information obtained, we developed EST-SSR markers by means of EST-SSR mining, primer design and polymorphism identification.

Results. A total of 91,449 unigenes were generated from the leaf cDNA library of *B. clarkeana* in this study. Based on a sequence similarity search with a known protein database, 72,087 unigenes were annotated. Among the annotated unigenes, a total of 71,170 unigenes showed significant similarity to known proteins of 463 popular model species in the Nr database, and 59,962 unigenes and 32,336 unigenes were assigned to GO classifications and COG, respectively. In addition, 44,924 unigenes were mapped in 128 KEGG pathways. Furthermore, a total of 7,610 unigenes with 8,563 microsatellites were found. Seventy-four primer pairs were selected from 436 primer pairs designed for polymorphism validation. SSRs with higher polymorphism rates were concentrated on dinucleotides, pentanucleotides and hexanucleotides. Finally, 17 pairs with highly polymorphic and stable loci were selected for polymorphism screening. There were a total

of 65 alleles, with 2–6 alleles at each locus. Mainly due to the unique biological characteristics of plants, the H_E, H_O and PIC per locus were very low, ranging from 0 to 0.196, 0.082 to 0.14 and 0 to 0.155, respectively.

Discussion. A substantial fraction transcriptome sequences of *B. clarkeana* were generated in this study, which is the first molecular-level analysis of this plant. These sequences are valuable resources for gene annotation and discovery and molecular marker development. These sequences could also provide a valuable basis for the future molecular study of *B. clarkeana*.

Introduction

Resurrection plants have desiccation tolerance (DT), which enables them to recover full metabolic competence upon rehydration after losing most of their cellular water (>95%) for extended periods of time (Farrant, Brandt & Lindsey, 2007). DT is commonly found in non-vascular plants and spores of tracheophytes (Rodriguez et al., 2010). It is rare in angiosperms (Porembski & Barthlott, 2000; Proctor & Pence, 2002) and in vegetative tissues of higher plants (Gaff, 1971). The mechanisms of DT are different between the extant lower orders and angiosperms (Farrant, Brandt & Lindsey, 2007). Understanding how plants with DT survive and respond to dehydration has great significance for plant biology and crop drought tolerance improvement, which could contribute to future water resource management decisions (Oliver et al., 2011a; Gechev et al., 2012; Xiao et al., 2015), and research on angiosperms with DT could be instructive for crops (Farrant, Brandt & Lindsey, 2007). In recent decades, efforts have focused on revealing the physiological and molecular mechanisms and their recovery processes in angiosperm plants with DT (Bianchi et al., 1993; Bernacchia, Salamini & Bartels, 1996; Sherwin & Farrant, 1998; Cooper & Farrant, 2002; Collett et al., 2003, 2004; Schneider et al., 2003; Alcazar et al., 2011; Oliver et al., 2011a, 2011b; Christ et al., 2014; Zhu et al., 2015). While a functional genomic approach, such as transcriptome sequencing, could be fruitful for exploring the mechanisms of DT (Xiao et al., 2015), transcriptomics could identify the metabolic processes involved in DT. Expressed sequence tag (EST) and EST-SSR (simple sequence repeat, a.k.a. microsatellite) markers could also be developed from transcriptome sequences (Dinakar & Bartels, 2013). EST-SSRs may regulate gene expression and function, making them valuable

resources for identifying associations with functional genes and phenotypes in future genetic studies (Zalapa et al., 2012). Therefore, transcriptomics would help to understand the mechanisms of DT. However, to our knowledge, only a few gene expression and EST sequencing studies have been performed in angiosperms with DT, including the dicot species *Craterostigma plantagineum* (Bockel, Salamini & Bartels, 1998), *Boea hygrometrica* (Xiao et al., 2015; Zhu et al., 2015), and *Haberlea rhodopensis* (Rodriguez et al., 2010; Gechev et al., 2013) and the monocot species *Sporobolus stapfianus* (Neale et al., 2000; Le et al., 2007), *Xerophyta viscosa* (Mundree et al., 2000; Mowla et al., 2002; Lehner et al., 2008), and *Xerophyta humilis* (Collett et al., 2004; Illing et al., 2005; Mulako et al., 2008).

Boea (Gesneriaceae) is a rare group of resurrection plants within angiosperms (Liu, Hu & Zhao, 2007; Xiao et al., 2015). *Boea clarkeana* Hemsl. is a desiccation-tolerant herb endemic to China. The whole plant, detached leaf and leaf segment all retain the DT phenotype, and this excellent drought-tolerant plant has been of concern in the last few years (Chao et al., 2013; Zhang et al., 2016). *B. clarkeana* is a small perennial dicotyledonous plant that is mainly distributed in 8 provinces and 1 municipality (Li, 1996; Li & Wang, 2005) along the middle-lower reaches of the Yangtze River in China. It is only found on rock outcrops (such as inselbergs) among some lithophytes, similar to mosses, ferns and ferns allies (Jenks & Wood, 2007). It is commonly used as a medicinal plant to treat traumatic hemorrhage and traumatic injury (Li & Wang, 2005). Genomic sequences of *B. clarkeana*, however, are scarce, with only a few nucleotide sequences found in public databases (http://www.ncbi.nlm.nih.gov/). To fill this critical gap and obtain the first genomic resources, we used the platform Illumina HiSeq™ 2000

and *de novo* assembly technology to obtain leaf transcriptomes of *B. clarkeana* and conducted a BLASTX (http://blast.ncbi.nlm.nih.gov/Blast.cgi) alignment of the sequencing data and protein databases for sequence classification and annotation. We developed polymorphic EST-SSR molecular markers based on the obtained sequence information. The preliminary accumulation of molecular data for *B. clarkeana* will help to understand transcript gene functions and classifications. Furthermore, molecular marker development can contribute to subsequent molecular studies of this plant with DT.

Materials and Methods

Plant materials and genomic DNA extraction

The materials of 11 natural populations were sampled from 6 provinces and 1 municipality that covered the vast majority of the natural habitats of *B. clarkeana* in China (Li & Wang, 2005). Young leaves were collected, rapidly dried and preserved in silica gel. DNA extraction was carried out with the QIAGEN® DNeasy® Plant Mini Kit (QIAGEN, Germany).

RNA isolation and cDNA library construction

The young leaves of three individual *B. clarkeana* plants from the population of Fenghuangshan in Anhui Province (30°88′ N, 118°02′ E) were collected, mixed and frozen in liquid nitrogen; then, the sampled tissues were stored at -80°C until used for RNA extraction. Total RNA isolation using a TRIzol kit (Life Technologies, USA) and DNase I (TaKaRa, Japan) followed the manufacturer's protocols. After total RNA was obtained, mRNA + poly(A) were isolated

using beads with Oligo (dT), and fragmentation buffer was added to cut mRNA into short fragments. Then, the transcription of RNA sequence fragments constituted first-strand cDNA using reverse transcriptase and random primers (Invitrogen, Carlsbad, CA), and the second-strand cDNA was synthesized using buffer, dNTPs, RNaseH and DNA polymerase I. Followed by the ligation of adapters, a single 'A' base was added to the 3' end of these cDNA fragments for end repair. Based on the amplification of these products, the cDNA library was generated and was separated on an agarose gel.

Sequencing and *de novo* assembly

The raw reads were produced from a cDNA library with an Illumina HiSeq™ 2000 genomic sequencer at the Beijing Genomics Institute (BGI, Shenzhen, China, http://www.genomics.cn/index). The subsequent analysis was based on clean reads that were generated by filtering raw reads. We therefore used the filter_fq program (BGI, Shenzhen, China) to remove reads with more than 5% unknown nucleotides 'N' and low-quality sequences with more than 20% low-quality bases (quality value ≤10) and adaptors to obtain clean reads. Then, we used the short read assembly program Trinity (Release-2013-02-25, http://trinityrnaseq.sourceforge.net/) for transcriptome *de novo* assembly (Grabherr et al., 2011) by combining clean reads to contigs with a sequence fragment length range of 200 bp (±25 bp), and two contigs were connected into a single scaffold. We called the resulting sequences unigenes. These unigenes were removed to prevent redundancy with TGICL (version 2.1) and further spliced to generate non-redundant unigenes that were as long as possible (Pertea et al.,

2003). The raw sequencing data with accession number SRX1600046 were deposited in the Sequence Read Archive (SRA) of National Center for Biotechnology Information (NCBI), which will be released in March 2018.

Functional annotation and classification of unigenes

BLASTX (http://blast.ncbi.nlm.nih.gov/Blast.cgi) alignment (*E*-value < 10^{-5}) between the unigenes and protein databases, such as NCBI non-redundant protein (Nr), Gene Ontology (GO, http://www.blast2go.com/b2ghome), and Cluster of Orthologous Groups (COG, http://www.ncbi.nlm.nih.gov/COG/), was performed to annotate and classify the transcriptome. Based on the Nr database annotation, we used Blast2GO v2.5.0 (Conesa et al., 2005) to obtain GO terms with an *E*-value threshold of 10^{-5}. With the Web Gene Ontology Annotation Plot (WEGO) (Ye et al., 2006), the distributions of GO terms were plotted to describe the categories, and the unigenes were also aligned to the COG database for possible functional prediction and classification. The unigenes containing SSRs were also aligned to euKaryotic Orthologous Groups (KOGs) through BLASTX. Finally, we annotated the unigenes to each level 3 pathway graph by mapping using the KEGG database to obtain pathway annotation for the unigenes (Kanehisa et al., 2008).

Detection and filtering of SNPs

Using SOAPsnp (Release 1.03) with All - Unigene as a reference to find the SNP for each sample, we analyzed the commonalities and differences of all SNP sites among the samples (Li

et al., 2009).

EST-SSR mining, primer design and polymorphism identification

SSRs from unigenes were detected and located using MIcroSAtellite (MISA, http://pgrc.ipk-gatersleben.de/misa/misa.html) (Zalapa et al., 2012). Compound SSRs (two or more SSRs in which the interval was no more than 100 bp) were excluded, and only SSRs with flanking sequences longer than 150 bp and containing 2 to 6 repeat motifs were considered. The mono-, di-, tri-, tetra-, penta- and hexa-nucleotide motif SSRs with a minimum of 12, 6, 5, 5, 4 and 4 repeats, respectively, were detected. We designed primer pairs with the online Primer3.0 (http://www.onlinedown.net/soft/51549.htm) using the following criteria: (1) a product sequences length of 100–300 bp and no secondary structure; (2) a primer length of 18–28 bp with an optimum of 23 bp; (3) a Tm of 55–65°C with an optimum of 60°C (with a difference between the Tm values of the forward and reverse primers no greater than 4°C); and (4) a GC content of 40–60% with 50% as the optimum. For other parameters, the default settings were used.

Seventy-four primer pairs divided into two groups were selected for DNA amplification. The first group of 50 primer pairs was randomly selected for amplification, and the motifs that had more polymorphic alleles in the first group would increase the selected ratio in the second 24 primer pairs. The mixed DNA from 3 individuals of *B. clarkeana* from different populations was used to verify amplification products, and the primers that amplified successfully were chosen for primary polymorphism identification, for which amplification was conducted using 12

172 individuals from 11 natural populations. Then, 128 individuals from 11 populations were
173 amplified using primer pairs that had more polymorphic loci for further polymorphism
174 identification.

175 The reverse primer with fluorescent (6-FAM, HEX, TAMRA or ROX) M13 forward
176 primer (M13F: 5'-GTAAAACGACGGCCAG-3') tails was used to accurately screen the
177 variation among individuals. PCR was performed in a 15-μL reaction containing 2.5 mM MgCl2
178 and dNTP (TaKaRa, Dalian, China), 0.5 U of *Taq* polymerase (TaKaRa, Dalian, China), 1× PCR
179 buffer, and 50 ng of genomic DNA. The primers included 0.04 μM forward primer, 0.04 μM
180 reverse primer with fluorescent M13 tails, and 0.01 μM M13 reverse primer (M13R: 5'-
181 CAGGAAAC AGCTATGAC-3'). The annealing temperature was different for each locus. We
182 used 54°C as the unified annealing temperature for PCR, and the amplification conditions were
183 as follows: initial denaturation at 94°C for 5 min; 35 cycles of 30 s at 94°C, 40 s annealing at
184 54°C, and 45 s elongation at 72°C; and a final extension at 72°C for 10 min. After screening on a
185 1.0% agarose gel, the sequence typing of successful products was carried out with an ABI 3730
186 DNA Analyzer (Applied Biosystems, Foster City, California, USA). Then, we manually scored
187 alleles using GeneMarker software (version 2.2.0).

188 Deviations from Hardy-Weinberg equilibrium (HWE) were calculated using GENEPOP
189 on the Web (http://www.genepop.curtin.edu.au/) with Bonferroni's correction. The number of
190 alleles (N_A) was calculated using MicroChecker (version 2.2.3). The expected (H_E) and observed
191 heterozygosity (H_O) of each locus were detected by GenALEx 6 (Peakall & Smouse, 2006), and
192 the polymorphism information content (PIC) was calculated using program PowerMarker

(version 3.25) (Liu & Muse, 2005). Then, neutral markers were detected using LOSITAN (Beaumont & Nichols, 1996; Antao et al., 2008).

Results

Illumina sequencing and *de novo* assembly

A total of 9,361,934,460 nt bases were generated in this study. After cleaning and quality checks, we obtained 104,021,494 clean reads with Q20 bases (sequences with sequencing error rates <1%) at 97.55%, and the N (ambiguous bases) and GC contents were 0 and 45.43%, respectively. *De novo* assembly was carried out with the program Trinity; a total of 94,546 contigs were generated with an average length of 487 nt and an N50 value of 1,075 nt. Finally, a total of 91,449 unigenes with a total length of 148,176,175 nt were detected; the average length and N50 were 1,620 nt and 2,389 nt, respectively. A summary of the sequence assembly after Illumina sequencing is shown in Table 1. The sequence-length distribution of the unigenes is shown in Fig. 1.

Functional annotation and classification of unigenes

For function annotation analysis, we obtained 71,170, 59,962, 32,336 and 44,929 unigenes annotated to the Nr, GO, COG and KEGG databases, respectively. The total number of annotated unigenes was 72,087 (78.82% of all unigenes).

Nr annotation

Using BLAST, 71,170 unigenes were annotated from 463 popular model species with databases of Nr. The species distribution of Nr annotations (Fig. 2) comprised *Lycopersicon esculentum* (35.1%), *Vitis vinifera* (27.8%), *Amygdalus persica* (6.7%), castor bean (*Ricinus communis*; 6.1%), black cottonwood (*Populus trichocarpa*; 5.2%), *Fragaria vesca* subsp. *vesca* (3.2%) and *Glycine max* (2.8%). Only a small fraction of all transcripts showed similarities to genes in other species. The most common species found in terms of this similarity were those of Solanaceae; only 25 species had genes similar (≥100) to those of *B. clarkeana* (not shown in the figure), and there were 6 species with genes similar to those of Solanaceae (26,585, 37.35%).

Gene ontology (GO) classification

Based on Nr annotations, 59,962 unigenes (65.57% of all unigenes) were assigned to three ontologies and subdivided into 55 subcategories with 501,897 functional GO terms of GO classifications (Fig. 3). Among these GO terms, the proportions of the Biological process, Cellular component and Molecular function ontologies were 49.45%, 37.11% and 13.43%, respectively. In the Biological process ontology, a high percentage of genes was classified under 'Cellular process' (39,131, 65.26% of Nr unigenes), 'Metabolic process' (36,670, 61.16%) and 'Single-organism process' (28,177, 46.99%), while only a few genes were classified under the terms 'Locomotion' (58, 0.10%), 'Rhythmic process' (441, 0.74%) and 'Biological adhesion' (549, 0.92%). 'Cell' and 'Cell part' were the same (47,457, 79.15%) in the Cellular component category, followed by 'Organelle' (38,055, 63.47%). Regarding Molecular function, the most represented category was 'Catalytic activity' (30,599, 51.03%), followed by 'Binding' (27,383,

45.67%).

COG and KOG classification of unigenes with SSRs

In total, 56,493 functionally annotated unigenes from 32,336 (35.36% of all unigenes) COG unigenes were assigned to 25 possible functional categories in COG annotations (Fig. 4-A). Among the categories, the largest group was the cluster for 'General function prediction only' (10,438, 32.28%), followed by 'Replication, recombination and repair' (5,561, 17.20%) and 'Transcription' (5,322, 13.46%). The smallest groups were 'Cell motility' (228, 0.71%), 'Extracellular structures' (17, 0.05%) and 'Nuclear structure' (14, 0.04%). After SSR detection using the software MIcroSAtellite (MISA) with unigenes as references, 7,610 unigenes carrying 8,563 SSRs were found. Then, 3,267 unigenes with SSRs had hits in 24 categories of the KOG database without 'Nuclear structure' (Fig. 4-B). Among 24 categories, the largest group was 'General function prediction' (1,166, 35.69% of unigenes with SSRs in KOG), followed by 'Transcription' (797, 24.40%), 'Replication, recombination and repair' (737, 22.56%) and 'Signal transduction mechanisms' (684, 20.94%).

Functional classification using KEGG

Based on sequence homology searches against the KEGG database, 44,924 unigenes (49.12% of all unigenes) were mapped in 128 pathways. Among these pathways, 'Metabolic pathway' (9,232, 20.55% of KEGG unigenes) and 'Metabolic biosynthesis of secondary metabolites' (3,764, 8.38%) were the largest categories of Metabolism. However, the second category was

also the greatest highlight of the KEGG pathway, with *B. clarkeana* as an environment-related pathway, in addition to 'Plant hormone signal transduction' (1,783, 3.97%), 'Plant-pathogen interaction' (1,769, 3.94%), 'Phosphatidylinositol signaling system' (535, 1.19%), 'ABC transporters' (499, 1.11%) and 'Circadian rhythm-plant' (377, 0.84%).

SNP detection

SNPs with at least 150 bp of flanking sequence on both sides were selected for further analysis. After quality filtering, a total of 11,330 high-quality SNPs were identified from all of the unigenes. The predicted SNPs included 6,903 transitions (C-T, 3,446 and A-G, 3,457) and 4,427 transversions (A-T, 1,293; A-C, 1,189; G-T, 1,203; and C-G, 742).

Frequency and distribution of SSRs

All 91,449 unigenes assembled were used to mine potential SSRs in this study, and a total of 7,610 unigenes containing 8,563 SSRs were identified. Among those unigenes with SSRs, 338 SSRs presented a compound formation, and 812 unigenes contained more than one SSR. On average, one SSR was found every 17.30 kb. Among SSRs, dinucleotide motifs were the most abundant (3,991, 46.61% of all SSRs), followed by mono- (2,163, 25.26%), tri- (1,957, 22.85%), hexa- (267, 3.12%), tetra- (198, 2.3%), and penta- (36, 0.42%) nucleotide motifs. The distribution and frequency of different motifs are shown in Fig. 5.

Among all SSR loci, 109 different motifs were identified. A/T (2,093, 24.44% of all SSRs) comprised the main part of the mononucleotide, and there were only 70 C/G in total. Of

277 the dinucleotides, AT/TA (1,564, 18.26%) and AG/CT (1,391, 16.24%) were roughly equivalent,
278 followed by AC/GT (1,035, 12.09%). Of the trinucleotides, AAG/CTT (441, 5.15%) was the
279 most common, followed by AAT/ATT (389, 4.54%), AGC/GCT (341, 3.98%), AGG/CCT (284,
280 3.32%) and ATC/GAT (232, 2.71%). The ACAT/ATGT (18, 0.21%) motif comprised the most
281 common tetranucleotides, and the most common pentanucleotides and hexanucleotides were
282 AAAAG/CTTTT (42, 0.49%) and AAGAGC/GCTCTT (68, 0.79%, Fig. 6), respectively.
283 The repeat numbers of most SSRs ranged from 4 to 12, and the most frequent repeat
284 number was 6 (2,066, 24.13%), followed by 5 (1,233, 14.40%) and 7 (1,113, 13.00%).
285 Furthermore, the length of SSRs ranged from 12 to 25 bp. The most common length was 12 bp
286 (2,442, 28.52%), followed by 15 bp (1,421, 16.60%) and 14 bp (1,111, 12.97%) (Fig. 7). Among
287 dinucleotides and trinucleotides, the most common lengths were 12 bp and 15 bp, respectively.
288 The longest length of di-, tri- and tetranucleotides was 24 bp, while the longest length of
289 pentanucleotides was 25 bp; all hexanucleotides were 24 bp.
290
291 **Development and validation of polymorphic SSR markers**
292 As a result, a total of 436 (only 5.73% of all sequences with SSRs) eligible primer pairs
293 (mononucleotide, 1; di-, 191; tri-, 205; tetra-, 5; penta-, 12; hexa-, 22) were designed using
294 Primer 3.0. The other 7,174 sequences were not successful in primer design mainly due to too-
295 long sequence lengths, insufficient flank lengths, and abundant sequences with mononucleotides.
296 Then, 74 primer pairs (dinucleotide, 20; tri-, 38; penta-, 3; hexa-, 13) were selected to validate
297 amplification across a composite sample of 3 individuals. A total of 60 primer pairs (81.08% of

298 74 primer pairs) showed stable and clear amplification. Meanwhile, the 14 remaining pairs with
299 failed PCR produced multiple bands or amplified unstably. Twenty-three primer pairs were
300 found to be monomorphic and 37 were found to be polymorphic after polymorphism screening
301 across 12 individuals. Among 37 polymorphic primer pairs, 17 pairs of highly polymorphic and
302 stable loci were selected for further screening across 128 individuals from 11 populations. For
303 the 17 polymorphic loci, there were 2–6 alleles at each locus, with a total of 65 alleles. The H_E,
304 H_O and PIC per locus ranged from 0 to 0.196, 0 to 0.14 and 0.155 to 0.664, respectively. For the
305 PIC values of the 17 polymorphic loci, 8 pairs having highly informative scores (PIC>0.50) and
306 5 pairs having weakly informative scores (0<PIC<0.25). Two primers (BC6 and BC11) could not
307 be calculated, and BC14 significantly deviated from HWE. The other 14 primers had no
308 significant departures from HWE after Bonferroni's correction (Table 2). The neutrality test by
309 LOSITAN showed that all 17 primer pairs agreed with the neutral theory (Fig. 8).
310
311 **Discussion**
312 **Assembly and functional annotation of unigenes**
313 *Unigenes*
314 Sequencing success was determined by the length of the reads, as longer reads would increase
315 the probability of SSRs being discovered (Zalapa et al., 2012). The final assembled transcripts
316 (average length was 1,620 nt; N50 was 2,389 nt) were longer than the sibling species, i.e., the
317 *Primulina* species with Illumina (Ai et al., 2015) and *B. hygrometrica* using the 454
318 pyrosequencing platform (Zhu et al., 2015), which produced longer reads than did Illumina

(Zalapa et al., 2012). Therefore, the sequencing results were ideal in this study.

Annotation

The predicted genes were functionally annotated using Nr, GO, KEGG and COG. In total, 72,078 unigenes (78.82% of all assembled unigenes) were successfully annotated in the present study, which was more than in the previous desiccation-tolerant plants reported for *B. hygrometrica* (66.6% (Zhu et al., 2015), 47.09% (Xiao et al. 2015)) and *Syntrichia caninervis* (58.7%) (Gao et al., 2014), which indicates that the functions of genes in *B. clarkeana* are better conserved. The structural features of the protein-coding gene complements (the species distribution of Nr annotation) for desiccation-tolerant plants in a previous report for *C. plantagineum* (Rodriguez et al., 2010), *B. hygrometrica* (Zhu et al., 2015) and *H. rhodopensis* (Gechev et al., 2013) were similar. Mainly, *V. vinifera*, *R. communis* and *P. trichocarpa* showed significant homology, but *B. clarkeana* in our study was obviously different, mainly due to *L. esculentum* (35.1%), *V. vinifera* (27.8%) and *A. persica* (6.7%). These species reflect a common origin with Solanales and Rhamnales different from *B. hygrometrica* (Xiao et al., 2015).

The enrichment of the GO (65.57% of all unigenes) and KEGG (49.12%) annotation in this study was much greater for *B. hygrometrica* (GO, 28.71%, KEGG, 24.43%; GO, 43.7%, KEGG, 15.1%) (Xiao et al., 2015; Zhu et al., 2015). The KEGG annotation in our study was enriched in the following vegetative dehydration/desiccation pathways: 'Plant-pathogen interaction' (1,769 unigenes, 3.94% of KEGG unigenes) in the pathogen defense system; 'Glycerophospholipid metabolism' (803, 1.79%) for protein receptor interactions in vesicular

trafficking; 'Plant hormone signal transduction' (1,783, 3.97%) of abiotic stress responses; the mRNA surveillance (1,027, 2.29%) pathway for damaged transcript removal; and Photosynthesis (154, 0.34%) and nitrogen metabolism (154, 0.34%) for the depletion of transcripts in dehydration (Xiao et al., 2015). The results of our study are consistent with genes and gene products whose central core is associated with DT in plants.

The cluster for 'General function prediction only' among all COG categories was the largest group in our study. This pattern is similar for some angiosperms, including *Camelina sativa* (Liang et al., 2013), *Apium graveolens* (Fu, Wang & Shen, 2013) and *Chrysanthemum nankingense* (Wang et al., 2013). The 'Replication, recombination and repair' (17.20%) category of *B. clarkeana* was much larger and showed more repaired genes in the plant.

Characteristics of EST-SSRs

In the present study, a total of 7,610 unigenes with 8,563 EST-SSRs were identified from the transcriptome of *B. clarkeana*. Compared with other reports for EST-SSRs identified using NGS (Next-Generation Sequencing, all of the approximately 2000 EST-SSRs) (Liu et al., 2013; Wang et al., 2013; Xiang et al., 2015), the quantity of EST-SSRs in our study was significantly larger, probably due to longer reads (Zalapa et al., 2012). In total, 3,267 unigenes with SSRs had hits in 24 categories of the KOG database compared with other studies of EST-SSRs (Li et al., 2012a; Liang et al., 2013; Liu et al., 2013). 'Replication, recombination and repair' and 'Signal transduction mechanisms' (684, 20.94%) were highlights for *B. clarkeana*.

Among the SSR repeats in our study, dinucleotide motifs were the most abundant (3,991,

46.61% of all SSRs), followed by mono- (2,163, 25.26%) and trinucleotide motifs (1,957, 22.85%). This result, is similar to reports on *A. graveolens* (Fu, Wang & Shen, 2013) and *Hevea brasiliensis* (Li et al., 2012a). Due to the long sequence length and short flack length, the vast majority of unigenes were not fit to design primers, and only 436 eligible primer pairs (mononucleotide, 1; di-, 191; tri-, 205; tetra-, 5; penta-, 12; hexa-, 22) were obtained. Additionally, 37 pairs (dinucleotide, 13; tri-, 13; penta-, 2; hexa-, 9) of 74 primer pairs (dinucleotide, 20; tri-, 38; penta-, 3; hexa-, 13) that were selected to validate amplification were polymorphic. The polymorphic percentage in dinucleotides was 65% (13 of the 20 selected were polymorphic), 34.21% (13 of 38) in trinucleotides, 66.67% (2 of 3) in pentanucleotides and 69.23% (9 of 13) in hexanucleotides.

Intrinsic features (such as repeat number, motif size, and length) could influence the rate and probability of slippage. These features were the strongest predictors of microsatellite mutability (Kelkar et al., 2008). The increased probability of slippage and mutation rates may be due to, for example, a greater number of repeats (Ellegren, 2004; Kelkar et al., 2008), a greater length irrespective of the repeat numbers (Webster, Smith & Ellegren, 2002), and lengths that were with inversely proportional to their motif sizes (Chakraborty et al., 1997). Additionally, the mutation rates might vary among SSRs with different motif compositions due to the dissimilarities of secondary DNA structure (Baldi & Baisnee, 2000). As a result, in our study, SSRs with higher polymorphism rates were concentrated on shorter motifs with a higher number of repeats (dinucleotides, 65%) and longer motifs with fewer repeats (hexanucleotides, 69.23%; pentanucleotides, 66.67%). Our analysis confirmed that mutability might increase with both

increased repeat number and greater length, as reported by Baldi and Baisnee (2000).

To increase accuracy in the polymorphism identification of primers and genetic variability comparison in the population, we chose 128 individuals from 11 populations that covered the majority of habitats of these plants for polymorphism screening. Nevertheless, the observed number of polymorphic primers was actually higher, but compared with other SSR and EST- SSR reports (Choudhary et al., 2009; Li et al., 2012a, 2012b; Yuan et al., 2012; Fu, Wang & Shen, 2013), the polymorphism level of the markers and the H_O, H_E, HWE and PIC of the population of *B. clarkeana* were still much lower in our study and were similar to those of *B. hygrometrica* (Xiao et al., 2015). These results could be attributed to two main reasons: first, the number of SSRs and polymorphisms of the DNA protein-coding sequence was expected to be lower than that in noncoding sequences, and the mutation rate within these regions was lower than that in other DNA sequences (Blanca et al., 2011; Zalapa et al., 2012). Second, due to the unique biological characteristics of *B. clarkeana*, the short stature of these plants could only be found on the north side of rock outcrops (mostly limestone) under the shade of trees and shrubs because these plants need scattered light (Chao et al., 2013), which might significantly reduce the potential for the long-distance dispersal of the wind-borne seeds. Furthermore, the occurrence of biparental inbreeding could be universal in the plants with high self-compatibility (Li & Wang, 2005), which would cause lower genetic variability within populations of *B. clarkeana*.

Conclusions

In this study, 91,449 unigenes were detected by NGS transcriptomics. A total of 8,563 SSRs

were identified from 7,610 unigenes, 72,087 unigenes were successfully annotated to protein databases, and polymorphic primer pairs of EST-SSRs were also developed. These results indicate that transcriptome sequencing is a highly efficient method of EST-SSR identification in non-model species that lack a reference genome and associations with functional genes. Therefore, by characterizing phenotypic features, these species can be identified (Li et al., 2002). These data will accelerate our assessment of functional gene identification and genetic variation in plants with DT, such as *B. clarkeana*. In addition, polymorphic primer pairs can continue to be developed from the remaining primers of EST-SSRs. The large-scale transcriptome dataset is a powerful resource for functional gene marker-assisted selection and DT exploration in *Boea*.

Acknowledgments

We are grateful to Cunhai Li and Fei Tan (JiangXi Guanshan National Nature Reserve) for assistance with sampling.

References

Ai B, Gao Y, Zhang X, Tao J, Kang M, Huang H. 2015. Comparative transcriptome resources of eleven Primulina species, a group of 'stone plants' from a biodiversity hot spot. *Molecular Ecology Resources* 15:619-632

Alcazar R, Bitrian M, Bartels D, Koncz C, Altabella T, Tiburcio AF. 2011. Polyamine metabolic canalization in response to drought stress in Arabidopsis and the resurrection plant *Craterostigma plantagineum*. *Plant Signaling & Behavior* 6:243-250

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G. 2008. LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics* 9:323. DOI: 10.1186/1471-2105-9-323.

Baldi P, Baisnee PF. 2000. Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* 16:865-889

Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B, Biological Sciences* 263:1619-1626

Bernacchia G, Salamini F, Bartels D. 1996. Molecular characterization of the rehydration process in the resurrection plant Craterostigma plantagineum. *Plant Physiology* 111:1043-1050

Bianchi G, Gamba A, Limiroli R, Pozzi N, Elster R, Salamini F, Bartels D. 1993. The unusual sugar composition in leaves of the resurrection plant *Myrothamnus flabellifolia*. *Physiologia Plantarum* 87:223-226

438 Blanca J, Canizares J, Roig C, Ziarsolo P, Nuez F, Pico B. 2011. Transcriptome characterization
439 and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). *BMC*
440 *Genomics* 12:104. DOI: 10.1186/1471-2164-12-104.
441 Bockel C, Salamini F, Bartels D. 1998. Isolation and characterization of genes expressed during
442 early events of the dehydration process in the resurrection plant *Craterostigma*
443 *plantagineum*. *Journal of Plant Physiology* 152:158-166
444 Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997. Relative mutation rates at
445 di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of*
446 *Sciences of the United States of America* 94:1041-1046
447 Chao T, Zhou S, Chang L, Chen Y, Xu H, Zhou Q. 2013. Effects of light intensity on the leaf
448 morphology and physiological parameters of *Boea clarkeana*. *Chinese Journal of*
449 *Ecology* 32:1161-1167
450 Choudhary S, Sethy NK, Shokeen B, Bhatia S. 2009. Development of chickpea EST-SSR
451 markers and analysis of allelic variation across related species. *Theoretical and Applied*
452 *Genetics* 118:591-608. DOI: 10.1007/s00122-008-0923-z.
453 Christ B, Egert A, Suessenbacher I, Kraeutler B, Bartels D, Peters S, Hoertensteiner S. 2014.
454 Water deficit induces chlorophyll degradation via the 'PAO/phyllobilin'pathway in
455 leaves of homoio-(*Craterostigma pumilum*) and poikilochlorophyllous (*Xerophyta*
456 *viscosa*) resurrection plants. *Plant, Cell & Environment* 37:2521-2531
457 Clarke K, Gorley R. 2001. *PRIMER v5: user manual/tutorial*. Plymouth: Primer-E Ltd.
458 Collett H, Butowt R, Smith J, Farrant J, Illing N. 2003. Photosynthetic genes are differentially

459 transcribed during the dehydration-rehydration cycle in the resurrection plant, *Xerophyta*
460 *humilis*. *Journal of Experimental Botany* 54:2593-2595. DOI: 10.1093/jxb/erg285.
461 Collett H, Shen A, Gardner M, Farrant JM, Denby KJ, Illing N. 2004. Towards transcript
462 profiling of desiccation tolerance in *Xerophyta humilis*: construction of a normalized 11 k
463 *X. humilis* cDNA set and microarray expression analysis of 424 cDNAs in response to
464 dehydration. *Physiologia Plantarum* 122:39-53
465 Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal
466 tool for annotation, visualization and analysis in functional genomics research.
467 *Bioinformatics* 21:3674-3676. DOI: 10.1093/bioinformatics/bti610.
468 Cooper K, Farrant JM. 2002. Recovery of the resurrection plant *Craterostigma wilmsii* from
469 desiccation: protection versus repair. *Journal of Experimental Botany* 53:1805-1813
470 Dinakar C, Bartels D. 2013. Desiccation tolerance in resurrection plants: new insights from
471 transcriptome, proteome and metabolome analysis. *Frontiers in Plant Science* 4:482. DOI:
472 10.3389/fpls.2013.00482.
473 Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews*
474 *Genetics* 5:435-445. DOI: 10.1038/nrg1348.
475 Farrant JM, Brandt W, Lindsey GG. 2007. An overview of mechanisms of desiccation tolerance
476 in selected angiosperm resurrection plants. *Plant Stress* 1:72-84
477 Fu N, Wang Q, Shen H. 2013. *De novo* assembly, gene annotation and marker development
478 using Illumina paired-end transcriptome sequences in celery (*Apium graveolens* L.). *PloS*
479 *One* 8:e57686

480 Gaff DF. 1971. Desiccation-tolerant flowering plants in southern Africa. *Science* 174:1033-1034.
481 DOI: 10.1126/science.174.4013.1033.
482 Gao B, Zhang D, Li X, Yang H, Wood AJ. 2014. *De novo* assembly and characterization of the
483 transcriptome in the desiccation-tolerant moss *Syntrichia caninervis*. *BMC Research*
484 *Notes* 7:490. DOI: 10.1186/1756-0500-7-490.
485 Gechev TS, Benina M, Obata T, Tohge T, Sujeeth N, Minkov I, Hille J, Temanni MR, Marriott
486 AS, Bergstrom E, Thomas-Oates J, Antonio C, Mueller-Roeber B, Schippers JH, Fernie
487 AR, Toneva V. 2013. Molecular mechanisms of desiccation tolerance in the resurrection
488 glacial relic *Haberlea rhodopensis*. *Cellular and Molecular Life Sciences* 70:689-709.
489 DOI: 10.1007/s00018-012-1155-6.
490 Gechev TS, Dinakar C, Benina M, Toneva V, Bartels D. 2012. Molecular mechanisms of
491 desiccation tolerance in resurrection plants. *Cellular and Molecular Life Sciences*
492 69:3175-3186. DOI: 10.1007/s00018-012-1088-0.
493 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
494 Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq
495 data without a reference genome. *Nature Biotechnology* 29:644-652
496 Illing N, Denby KJ, Collett H, Shen A, Farrant JM. 2005. The signature of seeds in resurrection
497 plants: a molecular and physiological comparison of desiccation tolerance in seeds and
498 vegetative tissues. *Integrative and Comparative Biology* 45:771-787. DOI:
499 10.1093/icb/45.5.771.
500 Jenks MA, Wood AJ, eds. 2007. *Plant desiccation tolerance*. Oxford, UK: Blackwell Publishing

Ltd.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* 36:D480-484. DOI: 10.1093/nar/gkm882.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research* 18:30-38. DOI: 10.1101/gr.7113408.

Le TN, Blomstedt CK, Kuang J, Tenlen J, Gaff DF, Hamill JD, Neale AD. 2007. Desiccation-tolerance specific gene expression in leaf tissue of the resurrection plant *Sporobolus stapfianus*. *Functional Plant Biology* 34:589-600

Lehner A, Chopera DR, Peters SW, Keller F, Mundree SG, Thomson JA, Farrant JM. 2008. Protection mechanisms in the resurrection plant *Xerophyta viscosa*: cloning, expression, characterisation and role of XvINO1, a gene coding for a myo-inositol 1-phosphate synthase. *Functional Plant Biology* 35:26-39

Li DJ, Deng Z, Qin B, Liu XH, Men ZH. 2012a. *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 Genomics* 13:192

Li M, Zhu L, Zhou CY, Lin L, Fan YJ, Zhuang ZM. 2012b. Development and characterization of EST-SSR markers from Scapharca broughtonii and their transferability in *Scapharca subcrenata* and *Tegillarca granosa*. *Molecules* 17:10716-10723

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved

522 ultrafast tool for short read alignment. *Bioinformatics* 25:1966-1967. DOI:
523 10.1093/bioinformatics/btp336.
524 Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution,
525 putative functions and mutational mechanisms: a review. *Molecular Ecology* 11:2453-
526 2465
527 Li Z, Wang Y. 2005. *Plants of Gesneriaceae in China*. Henan, China: Henan Science and
528 Technology Publishing House.
529 Li ZY. 1996. The georaphical distribution of the subfamily Cyrtandroideae Endl. emend. Burtt
530 (Gesneriaceae). *Acta Phytotax Sin* 34:341-360
531 Liang C, Liu X, Yiu SM, Lim BL. 2013. *De novo* assembly and characterization of *Camelina*
532 *sativa* transcriptome by paired-end sequencing. *BMC Genomics* 14:146. DOI:
533 10.1186/1471-2164-14-146.
534 Liu G, Hu Y, Zhao F. 2007. Molecular cloning of BcWRKY1 transcriptional factor gene from
535 *Boea crassifolia* Hemsl and its preliminary functional analysis. *Acta Scicentiarum*
536 *Naturalum Universitis Pekinesis* 43:446-452
537 Liu K, Muse SV. 2005. PowerMarker: an integrated analysis environment for genetic marker
538 analysis. *Bioinformatics* 21:2128-2129. DOI: 10.1093/bioinformatics/bti282.
539 Liu ZP, Chen TL, Ma LC, Zhao ZG, Zhao PX, Nan ZB, Wang YR. 2013. Global transcriptome
540 sequencing using the Illumina platform and the development of EST-SSR markers in
541 *Autotetraploid alfalfa*. *PLoS One* 8:e83549
542 Mowla SB, Thomson JA, Farrant JM, Mundree SG. 2002. A novel stress-inducible antioxidant

543 enzyme identified from the resurrection plant *Xerophyta viscosa* Baker. *Planta* 215:716-
544 726. DOI: 10.1007/s00425-002-0819-0.
545 Mulako I, Farrant J, Collett H, Illing N. 2008. Expression of Xhdsi-1VOC, a novel member of
546 the vicinal oxygen chelate (VOC) metalloenzyme superfamily, is up-regulated in leaves
547 and roots during desiccation in the resurrection plant *Xerophyta humilis* (Bak) Dur and
548 Schinz. *Journal of Experimental Botany* 59:3885-3901
549 Mundree SG, Whittaker A, Thomson JA, Farrant JM. 2000. An aldose reductase homolog from
550 the resurrection plant *Xerophyta viscosa* Baker. *Planta* 211:693-700. DOI:
551 10.1007/s004250000331.
552 Neale AD, Blomstedt CK, Bronson P, Le TN, Guthridge K, Evans J, Gaff DF, Hamill JD. 2000.
553 The isolation of genes from the resurrection grass *Sporobolus stapfianus* which are
554 induced during severe drought stress. *Plant, Cell & Environment* 23:265-277
555 Oliver MJ, Guo LN, Alexander DC, Ryals JA, Wone BWM, Cushman JC. 2011a. A sister group
556 contrast using untargeted global metabolomic analysis delineates the biochemical
557 regulation underlying desiccation tolerance in *Sporobolus stapfianus*. *The Plant Cell*
558 23:1231-1248
559 Oliver MJ, Jain R, Balbuena TS, Agrawal G, Gasulla F, Thelen JJ. 2011b. Proteome analysis of
560 leaves of the desiccation-tolerant grass, *Sporobolus stapfianus*, in response to dehydration.
561 *Phytochemistry* 72:1273-1284. DOI: 10.1016/j.phytochem.2010.10.020.
562 Peakall R, Smouse PE. 2006. GENALEX 6: genetic analysis in Excel. Population genetic
563 software for teaching and research. *Moleculary Ecology Notes* 6:288-295

564 Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung
565 F, Parvizi B, Tsai J, Quackenbush J. 2003. TIGR Gene Indices clustering tools (TGICL):
566 a software system for fast clustering of large EST datasets. *Bioinformatics* 19:651-652
567 Porembski S, Barthlott W. 2000. Granitic and gneissic outcrops (inselbergs) as centers of
568 diversity for desiccation-tolerant vascular plants. *Plant Ecology* 151:19-28
569 Proctor MCF, Pence VC. 2002. Vegetative tissues: bryophytes, vascular resurrection plants and
570 vegetative propagules. In: Black M, and Pritchard HW, eds. *Desiccation and survival in*
571 *plants: drying without dying*. New York: CABI Publishing, 207-237.
572 Rodriguez MC, Edsgard D, Hussain SS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH,
573 Bartels D, Mundy J. 2010. Transcriptomes of the desiccation-tolerant resurrection plant
574 *Craterostigma plantagineum*. *The Plant Journal* 63:212-228. DOI: 10.1111/j.1365-
575 313X.2010.04243.x.
576 Schneider H, Manz B, Westhoff M, Mimietz S, Szimtenings M, Neuberger T, Faber C, Krohne
577 G, Haase A, Volke F. 2003. The impact of lipid distribution, composition and mobility on
578 xylem water refilling of the resurrection plant *Myrothamnus flabellifolia*. *New*
579 *Phytologist* 159:487-505
580 Sherwin HW, Farrant JM. 1998. Protection mechanisms against excess light in the resurrection
581 plants *Craterostigma wilmsii* and *Xerophyta viscosa*. *Plant Growth Regulation* 24:203-
582 210
583 Wang H, Jiang J, Chen S, Qi X, Peng H, Li P, Song A, Guan Z, Fang W, Liao Y, Chen F. 2013.
584 Next-generation sequencing of the *Chrysanthemum nankingense* (Asteraceae)

585 transcriptome permits large-scale unigene assembly and SSR marker discovery. *PLoS*
586 *One* 8:e62293. DOI: 10.1371/journal.pone.0062293.
587 Webster MT, Smith NG, Ellegren H. 2002. Microsatellite evolution inferred from human-
588 chimpanzee genomic sequence alignments. *Proceedings of the National Academy of*
589 *Sciences of the United States of America* 99:8748-8753. DOI: 10.1073/pnas.122067599.
590 Xiang XY, Zhang ZX, Wang ZG, Zhang XP, Wu GL. 2015. Transcriptome sequencing and
591 development of EST-SSR markers in *Pinus dabeshanensis*, an endangered conifer
592 endemic to China. *Molecular Breeding* 35:1-10
593 Xiao LH, Yang G, Zhang LC, Yang XH, Zhao S, Ji ZZ, Zhou Q, Hu M, Wang Y, Chen M. 2015.
594 The resurrection genome of *Boea hygrometrica*: a blueprint for survival of dehydration.
595 *Proceedings of the National Academy of Sciences of the United States of America*
596 112:5833-5837
597 Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, Wang J. 2006.
598 WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research* 34:W293-297.
599 DOI: 10.1093/nar/gkl031.
600 Yuan N, Sun Y, Nakamura K, Qiu YX. 2012. Development of microsatellite markers in
601 heterostylous *Hedyotis chrysotricha* (Rubiaceae). *American Journal of Botany* 99:e43-45.
602 DOI: 10.3732/ajb.1100304.
603 Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P.
604 2012. Using next-generation sequencing approaches to isolate simple sequence repeat
605 (SSR) loci in the plant sciences. *American Journal of Botany* 99:193-208

606 Zhang D, Zhou S, Zhou H, Liu F, Yang S, Ma Z. 2016. Physiological response of *Boea*
607 *clarkeana* to dehydration and rehydration. *Chinese Journal of Ecology* 35:72-78
608 Zhu Y, Wang B, Phillips J, Zhang ZN, Du H, Xu T, Huang LC, Zhang XF, Xu GH, Li WL. 2015.
609 Global transcriptome analysis reveals acclimation-primed processes involved in the
610 acquisition of desiccation tolerance in *Boea hygrometrica*. *Plant and Cell Physiology*
611 56:1429-1441

613 **Tables**

614 **Table 1 Summary of sequence assembly using Illumina sequencing**

Sequence	Items	Value
Reads	Total Raw Reads	110,834,050
	Total Clean Reads	104,021,494
	Total Clean Nucleotides (nt)	9,361,934,460
	Q20 percentage (%)	97.55
	N percentage (%)	0
	GC percentage (%)	45.43
Contig	Total number	94,546
	Total length (nt)	46,012,409
	Mean length (nt)	1,075
	Contig N50 (nt)	487
Unigene	Total number	91,449
	Total length (nt)	148,176,175
	Mean length (nt)	1,620
	Unigene N50 (nt)	2,389
	Distinct Clusters	55,888
	Distinct Singletons	35,561,561

615

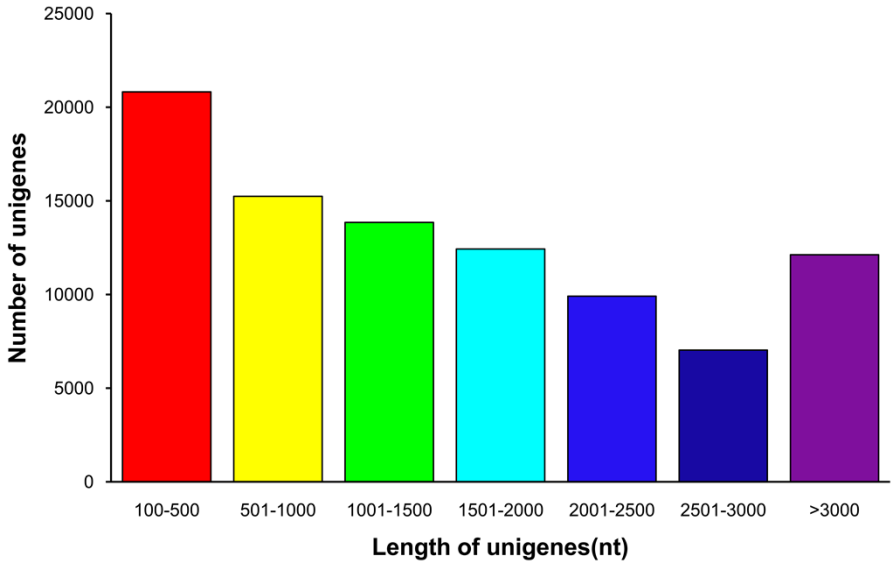
617 **Table 2 Characteristics of 17 polymorphic EST-SSR markers**

Locus	Primer sequence 5'–3'	Repeat motif	N_A	Size range (bp)	H_E	H_O	HWE[a]	PIC	GenBank Accession No.
BC1	F:GCAGTTCTGTGCAGTACCATACAT	$(TA)_6$	4	172-182	0.065	0.038	0.036*	0.193	Pr032805680
	R:TGGCTTCTGATCAGGTTTCTGAAT								
BC2	F:GAGATCCCAGATCCAGATCTTCT	$(TC)_6$	3	160-164	0.038	0.023	0.192 n.s	0.423	Pr032805681
	R:AACATTAATGGAAACACGTCGTC								
BC3	F:ATTCGCTCTCTTGGTATGACTGT	$(TA)_6$	5	170-184	0.054	0.045	0.380 n.s	0.664	Pr032805682
	R:CCCAATTTGAAGTGTTGCTTTAC								
BC4	F:TATCAGCGTGTGTGAATAGTTGC	$(TA)_7$	4	157-163	0.097	0.045	0.004**	0.491	Pr032805683
	R:TAACCTAAATTCGAATCCATCCA								
BC5	F:CAAACTTGGCTTAATACCATTCG	$(TG)_9$	3	119-125	0.079	0.083	0.713 n.s	0.469	Pr032805684
	R:CCATGATCATCTCTATTTCAGGC								
BC6	F:CCTTAAGGAGATGCATTGTGAAT	$(TC)_9$	3	159-169	0.000	0.000	- n.c.	0.299	Pr032805685
	R:GTATGAAGGGCATCAACAATAGG								
BC7	F:GCTGAAAGTTGGTGATTGCTAGT	$(AT)_9$	4	166-178	0.120	0.125	0.087 n.s	0.526	Pr032805686
	R:AGTTATGTCTTCGCTTGCTTCAG								
BC8	F:AACGTGAGAGTGCTAGTTCGGTA	$(TGA)_5$	3	167-173	0.014	0.000	0.041*	0.17	Pr032805687
	R:TCTTCCTCACTTTATCATCCACG								
BC9	F:AGAAGAGGTACGACAGTTTGCTG	$(GCG)_5$	2	156-159	0.059	0.064	1.000 n.s	0.195	Pr032805688
	R:TTCACGTCCGAATTCTTAGTCTC								
BC10	F:CACTGCACATAGAAGGAGGAGTT	$(GCG)_6$	5	108-129	0.081	0.076	0.146 n.s	0.581	Pr032805689
	R:GTAATCGCCTACATGATTCATCC								
BC11	F:CAGCAGTATGTCGGGATTATTTC	$(TTTCT)_4$	2	123-133	0.000	0.000	-n.c.	0.155	Pr032805690
	R:TCCTCTGGTCATATTGCTGTTACC								
BC12	F:AACAAGAGGGTCAGCTACAACAG	$(CAGCAA)_4$	4	160-178	0.104	0.095	0.184 n.s	0.549	Pr032805691
	R:CAGCAATGGTATTAGCAGAGGAC								
BC13	F:ACCTTGACGATCCTTCATCTTCT	$(GGTGCG)_4$	6	132-174	0.124	0.095	0.161 n.s	0.701	Pr032805692
	R:TTATGTTCTCCATATCCGTCAGC								
BC14	F:GGCAGCAATATAGCTCAAATACG	$(GACAAG)_4$	4	170-188	0.196	0.083	0.000***	0.516	Pr032805693
	R:ACCTGATCGTTCACAACTTCATC								
BC15	F:TCTTATTCAACACAACAGCCTGA	$(ATGATA)_4$	5	151-175	0.157	0.140	0.228 n.s	0.528	Pr032805694
	R:TGCTGCAGTTGATAATGAGAAGGA								
BC16	F:ACCAATGGTCTATATTTCAACGG	$(ATTACT)_4$	6	149-179	0.132	0.125	0.174 n.s	0.643	Pr032805695
	R:TGTGCCCCACATAGCTTCTATCTA								
BC17	F:TGACGAGGCTTCTACAGAATGAG	$(CATCCT)_4$	2	137-143	0.034	0.045	1.000 n.s	0.186	Pr032805696
	R:TACAAACAACAAGATGGGAATCAT								

618 *Note:* N_A = number of alleles per locus across all populations; H_E = expected heterozygosity (mean value); H_O
619 = observed heterozygosity (mean value); PIC, polymorphic information content; HWE = Hardy-Weinberg
620 equilibrium.[a] After Bonferroni correction, the significant departures from Hardy-Weinberg equilibrium: *

621 $P<0.05$, ** $P<0.01$, *** $P<0.001$. n.s. = not significant, n.c. = not calculated (Clarke & Gorley, 2001).
622

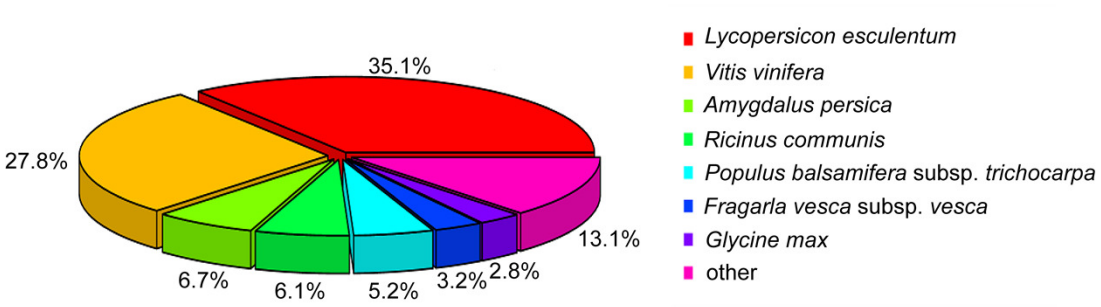
623 **Figure captions**



624

625 **Fig. 1 The length distribution of the unigenes**
626

627



628

629 **Fig. 2 The species distribution of Nr annotation**

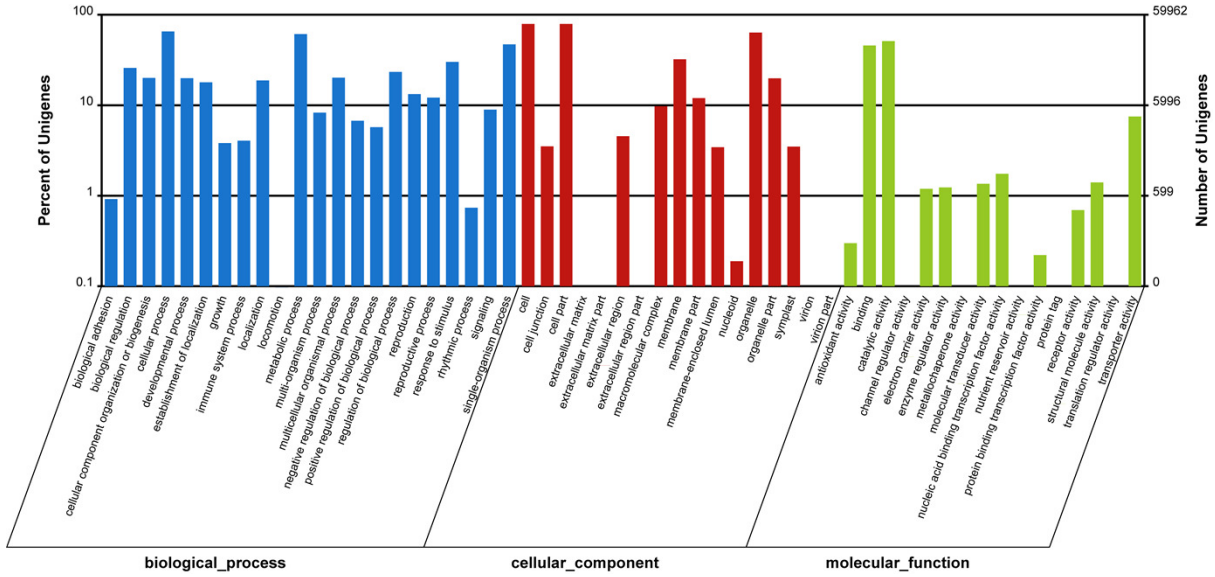


Fig. 3 Gene ontology classification of unigenes

GO functions are shown in the X-axis. The right Y-axis shows the number of genes with the GO function, and the left Y-axis shows the percentage.

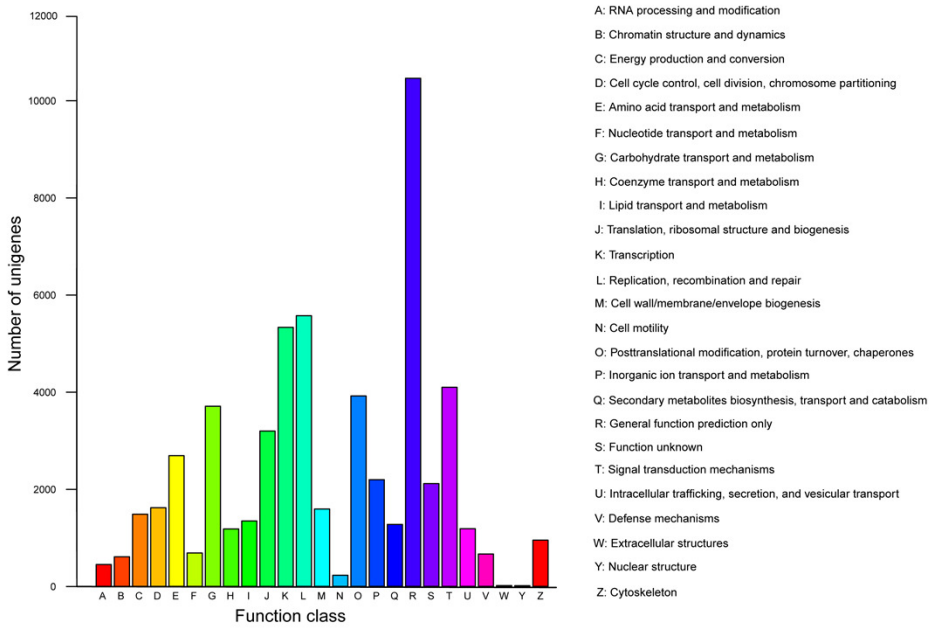
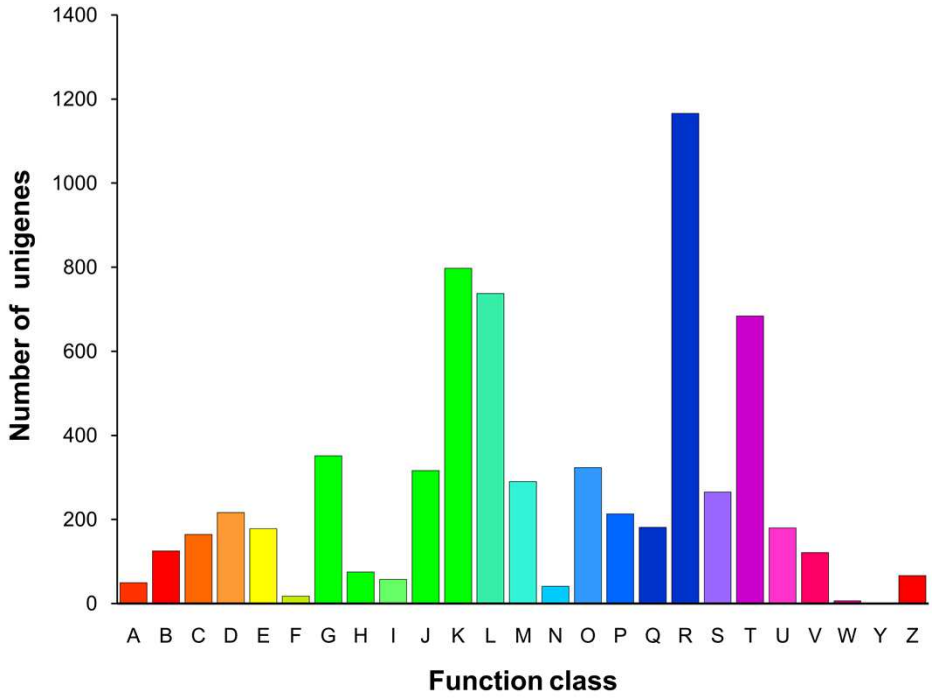
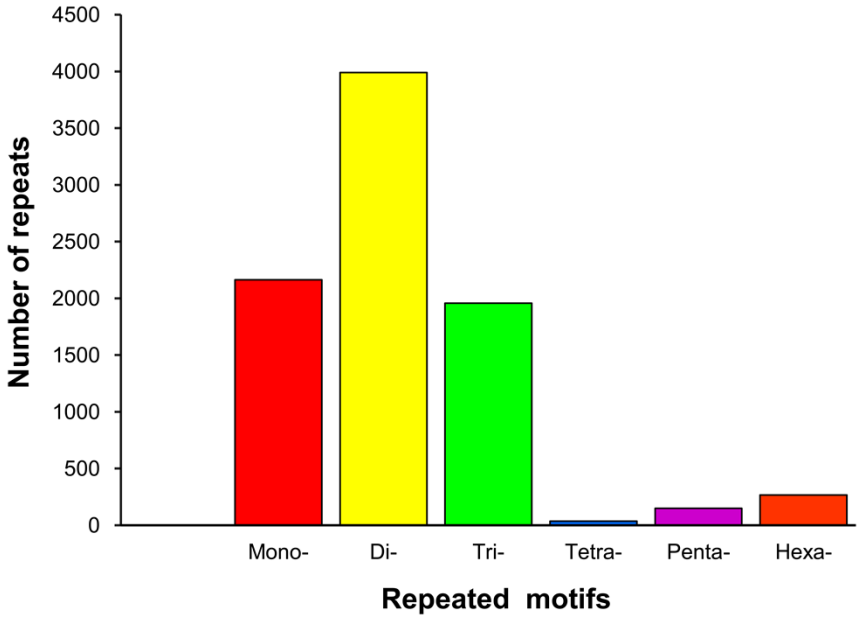


Fig. 4-A The COG functional classification of unigenes

In Fig. 4-A and Fig. 4-B, the horizontal coordinates are functional classes of COG and KOG, and the vertical coordinates are numbers of unigenes in one class. The notation on the right in Fig. 4-A is the full name of the functions on the X-axis.

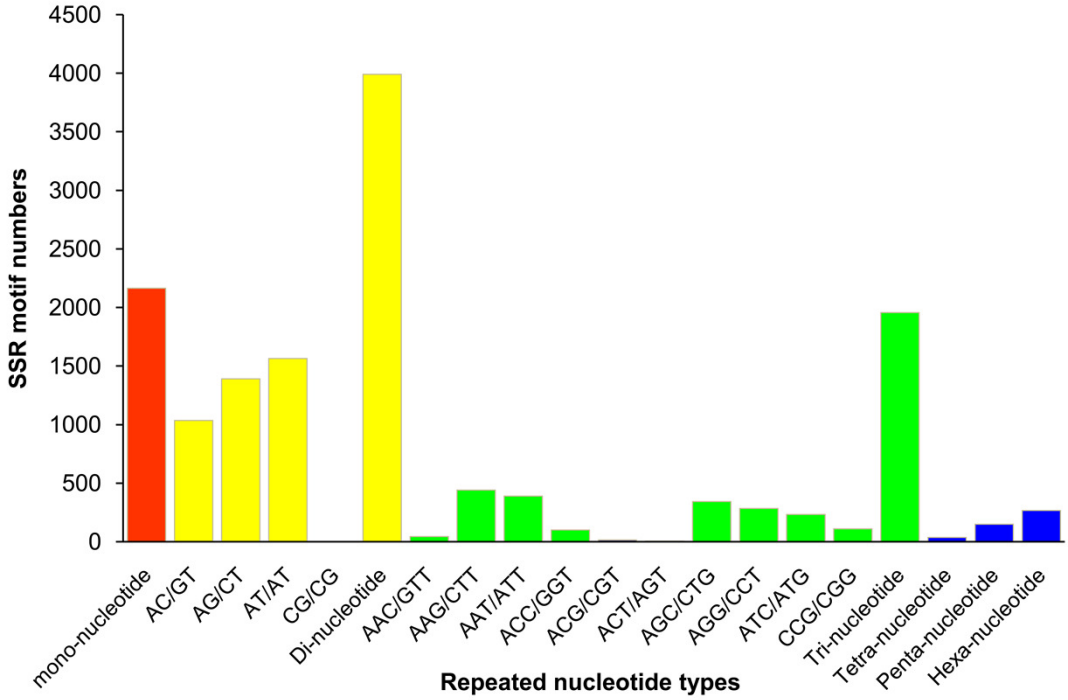


642
643 **Fig. 4-B The KOG functional classification of unigenes with SSRs**



644

645 **Fig. 5 The distribution and frequency of different motifs**



646

647 **Fig. 6 The distribution of mainly repeated nucleotide types**

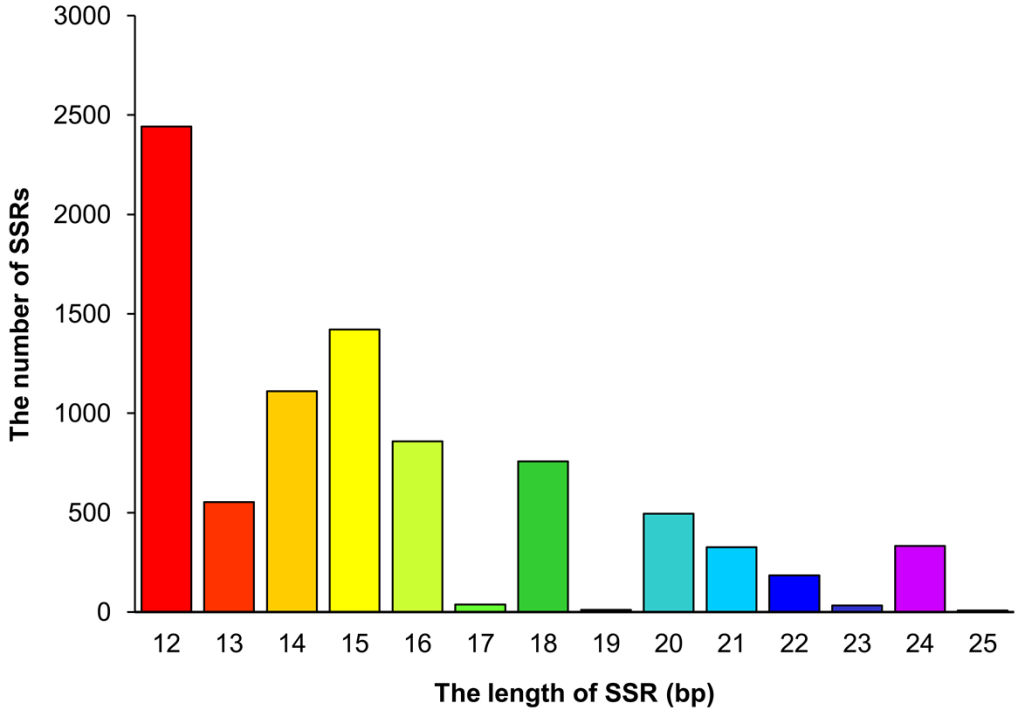
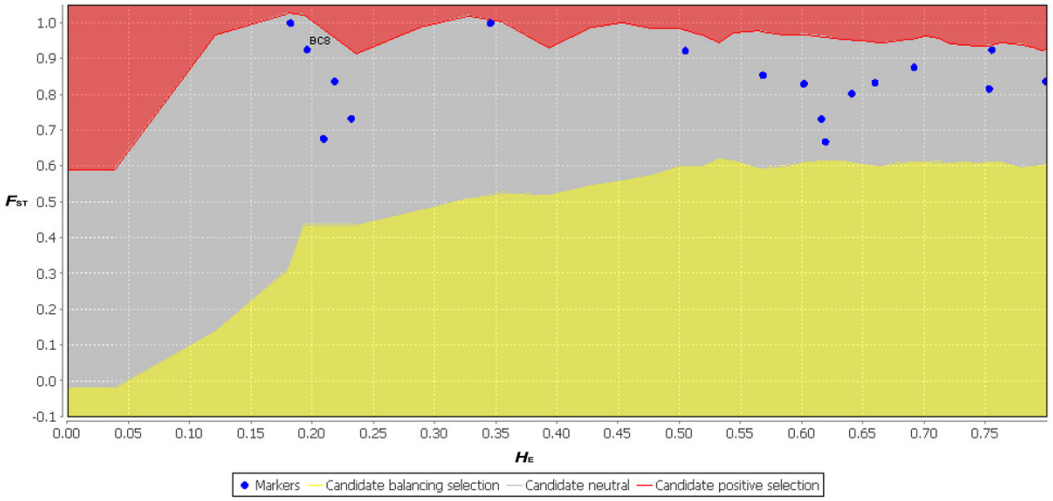


Fig. 7 The distribution of SSRs of different lengths



651

652 **Fig. 8 The neutral test results of 17 primer pairs using F_{ST} and H_E from 11 populations by**

653 **LOSITAN**

654