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], Shoubiao Zhou [Corresp., 1,3], Jianwen Shao [Corresp. 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, Anhui Normal University, Wuhu, Anhui, China

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

Corresponding Authors: Shoubiao Zhou, Jianwen Shao
Email address: zhoushoubiao@vip.163.com, 545491044@qq.com

Background. Desiccation-tolerant (DT) plants can recover full metabolic competence upon rehydration after losing most of their cellular water (>95%) for extended periods of time. Functional genomic approaches such as transcriptome sequencing can help us understand how DT plants survive and respond to dehydration, which has great significance for plant biology and improving the drought tolerance of crops. *Boea clarkeana* Hemsl. (Gesneriaceae) is a DT dicot yledonous herb. Its genomic sequences characteristics remain unknown. Based on transcriptomic analyses, polymorphic EST-SSR (simple sequence repeats in expressed sequence tags) molecular primers can be designed, which will greatly facilitate further investigations of the population genetics and demographic histories of DT 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, the EST-SSR markers were developed, and the functional annotation of ESTs containing polymorphic SSRs were obtained through BLASTX. **Results.** A total of 91,449 unigenes were generated from the leaf cDNA library of *B. clarkeana*. 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 the 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 Cluster of Orthologous Groups (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 stable, highly polymorphic loci were selected for polymorphism screening. There was a total of 65 alleles, with 2–6 alleles at each locus. Primarily due to the unique biological characteristics of plants, the H_E (0–0.196), H_O (0.082–0.14) and PIC (0–0.155) per locus were very low. The functional annotation distribution centered on ESTs containing di- and tri-nucleotide SSRs, and the ESTs containing primers BC2, BC4 and BC12 were annotated to vegetative dehydration/desiccation pathways. **Discussion.** This work is the first genetic study of *B. clarkeana* as a new plant resource of DT genes. A substantial number of transcriptome sequences were generated in this study. These sequences are valuable resources for gene annotation and discovery as well as molecular marker development. These sequences could also provide a valuable basis for future molecular studies of *B*. *clarkeana*.

1   **Characterization of the transcriptome and EST-SSR development in *Boea clarkeana*, a**

2   **desiccation-tolerant plant endemic to China**

4   **Ying Wang[1,3], Kun Liu[1,3], De Bi[1], Shoubiao Zhou[1,2], Jianwen Shao[1,3]**

6   [1]College of Life Sciences, Anhui Normal University, Wuhu, Anhui, China

7   [2]College of Environmental Science and Engineering, Anhui Normal University, Wuhu, Anhui,

8   China

9   [3]Anhui Provincial Key Laboratory of the Conservation and Exploitation of Biological

10   Resources, Wuhu, Anhui, China

12   **Corresponding author:**

13   Shoubiao Zhou

14   Email address: zhoushoubiao@vip.163.com;

15   Jianwen Shao

16   Email address: 545491044@qq.com.

17   **Abstract**

18   **Background.** Desiccation-tolerant (DT) plants can recover full metabolic competence upon

19   rehydration after losing most of their cellular water (>95%) for extended periods of time.

20   Functional genomic approaches such as transcriptome sequencing can help us understand how

21   DT plants survive and respond to dehydration, which has great significance for plant biology and

22   improving the drought tolerance of crops. *Boea clarkeana* Hemsl. (Gesneriaceae) is a DT

23   dicotyledonous herb. Its genomic sequences characteristics remain unknown. Based on

24   transcriptomic analyses, polymorphic EST-SSR (simple sequence repeats in expressed sequence

25   tags) molecular primers can be designed, which will greatly facilitate further investigations of the

26   population genetics and demographic histories of DT plants.

27   **Methods**. In the present study, we used the platform Illumina HiSeq^{TM}2000 and *de novo*

28   assembly technology to obtain leaf transcriptomes of *B. clarkeana* and conducted a BLASTX

29   alignment of the sequencing data and protein databases for sequence classification and

30   annotation. Then, based on the sequence information, the EST-SSR markers were developed, and

31   the functional annotation of ESTs containing polymorphic SSRs were obtained through

32   BLASTX.

33   **Results.** A total of 91,449 unigenes were generated from the leaf cDNA library of *B. clarkeana*.

34   Based on a sequence similarity search with a known protein database, 72,087 unigenes were

35   annotated. Among the annotated unigenes, a total of 71,170 unigenes showed significant

36   similarity to the known proteins of 463 popular model species in the Nr database, and 59,962

37   unigenes and 32,336 unigenes were assigned to GO classifications and Cluster of Orthologous

38   Groups (COG), respectively. In addition, 44,924 unigenes were mapped in 128 KEGG pathways.

39   Furthermore, a total of 7,610 unigenes with 8,563 microsatellites were found. Seventy-four

40   primer pairs were selected from 436 primer pairs designed for polymorphism validation. SSRs

41   with higher polymorphism rates were concentrated on dinucleotides, pentanucleotides and

42   hexanucleotides. Finally, 17 pairs with stable, highly polymorphic loci were selected for

43   polymorphism screening. There was a total of 65 alleles, with 2–6 alleles at each locus. Primarily

44   due to the unique biological characteristics of plants, the H_E (0–0.196), H_O (0.082–0.14) and PIC

45   (0–0.155) per locus were very low. The functional annotation distribution centered on ESTs

46   containing di- and tri-nucleotide SSRs, and the ESTs containing primers BC2, BC4 and BC12

47   were annotated to vegetative dehydration/desiccation pathways.

48   **Discussion.** This work is the first genetic study of *B. clarkeana* as a new plant resource of DT

49   genes. A substantial number of transcriptome sequences were generated in this study. These

50   sequences are valuable resources for gene annotation and discovery as well as molecular marker

51   development. These sequences could also provide a valuable basis for future molecular studies of

52   *B. clarkeana*.

Introduction

Resurrection plants are desiccation-tolerant (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). Though non-vascular plants and spores of

tracheophytes are commonly DT (Rodriguez et al., 2010), this feature is rare in angiosperms

(Gaff, 1971; Porembski & Barthlott, 2000; Proctor & Pence, 2002). The mechanisms of DT are

different between the extant lower orders and angiosperms (Farrant, Brandt & Lindsey, 2007).

Understanding how DT plants 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. Moreover, research on DT angiosperms could inform crop

cultivation (Farrant, Brandt & Lindsey, 2007; Oliver et al., 2011a; Gechev et al., 2012; Xiao et

al., 2015). In recent decades, research has focused on revealing the physiological and molecular

mechanisms of DT in angiosperm plants and their recovery processes (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. GO (Gene

Ontology, http://www.blast2go.com/b2ghome) and COG (Cluster of Orthologous Groups,

http://www.ncbi.nlm.nih.gov/COG/) analyses can also help us understand the distribution of

functional genes in plants at the macro level (Conesa et al., 2005; Ye et al., 2006). Moreover, the

75   gene products of metabolic processes and the functions of genes related to cellular processes can

76   be detected by BLASTX using the KEGG database (Kyoto Encyclopedia of Genes and Genomes,

77   http://www.kegg.jp/) (Kanehisa et al., 2008). These approaches can help us study gene behaviors

78   in biologically complex processes, such as vegetative dehydration/desiccation pathways, in DT

79   plants (Xiao et al., 2015). However, to the best of our knowledge, only a few gene expression

80   and EST sequencing studies have been performed in angiosperms with DT, including the dicot

81   species *Craterostigma plantagineum* (Bockel, Salamini & Bartels, 1998), *Boea hygrometrica*

82   (Xiao et al., 2015), and *Haberlea rhodopensis* (Rodriguez et al., 2010; Gechev et al., 2013) and

83   the monocot species *Sporobolus stapfianus* (Neale et al., 2000; Le et al., 2007), *Xerophyta*

84   *viscosa* (Mundree et al., 2000; Mowla et al., 2002; Lehner et al., 2008), and *Xerophyta humilis*

85   (Collett et al., 2004; Illing et al., 2005; Mulako et al., 2008).

86   *Boea* (Gesneriaceae) is a rare group of resurrection plants within angiosperms (Liu, Hu &

87   Zhao, 2007; Xiao et al., 2015). *Boea clarkeana* Hemsl. is a desiccation-tolerant herb endemic to

88   China. The whole plant, detached leaf and leaf segment all retain the DT phenotype, and the

89   excellent drought tolerance of this plant has been of concern in the last few years (Chao et al.,

90   2013; Zhang et al., 2016). *B. clarkeana* is a small perennial dicotyledonous plant that is mainly

91   distributed in 8 provinces and 1 municipality along the middle-lower reaches of the Yangtze

92   River in China (Li, 1996; Li & Wang, 2005). It is found only on rock outcrops (such as

93   inselbergs) among some lithophytes, where dehydration occurs frequently (Jenks & Wood, 2007).

94   It is commonly used as a medicinal plant to treat traumatic hemorrhage and traumatic injury (Li

95   & Wang, 2005). However, genomic sequences of *B. clarkeana* are scarce, and only a few

96   nucleotide sequences are found in public databases (http://www.ncbi.nlm.nih.gov/). To fill this

97   critical gap and obtain the first genomic resources, we used the Illumina HiSeq™2000 platform

98   and *de novo* assembly to obtain leaf transcriptomes of *B. clarkeana* and conducted a BLASTX

99   (http://blast.ncbi.nlm.nih.gov/Blast.cgi) alignment of the sequencing data and protein databases

100 for sequence classification and annotation.

101   We also assessed the SSRs, or microsatellites, that are distributed universally in gene

102 coding and non-coding regions. As the major tool of genetic research, the neutral SSR markers

103 are widely used in botanic sciences and functional SSR markers might affect gene function by

104 influencing recombination and chromatin organization, regulating gene expression and activity,

105 etc.(Cavagnaro et al., 2010; Li et al., 2012a; Zalapa et al., 2012). In DT plants, identifying

106 functional genes that contain functional SSRs could help us to understand the evolution and

107 expression of these genes, making SSRs a valuable resource for genetic studies (Li et al., 2002;

108 Gupta et al., 2003). Therefore, based on the *B. clarkeana* transcriptome, 17 pairs of polymorphic

109 EST-SSR molecular primers were developed and characterized. The results of this study will

110 greatly facilitate further investigations of the genetics and demographic histories of populations

111 of this DT plant.

112

113 **Materials and Methods**

114 ***Plant materials and genomic DNA extraction***

115 The materials of 11 natural populations were sampled from 6 provinces and 1 municipality in

116 China that covered the vast majority of the natural habitats of *B. clarkeana* (Li & Wang, 2005).

117   Young leaves were collected, rapidly dried and preserved in silica gel.DNA extraction was

118   carried out using the QIAGEN® DNeasy® Plant Mini Kit (QIAGEN, Germany).

119

120   *RNA isolation and cDNA library construction*

121   The young leaves of three individual *B. clarkeana* plants from the population of Mt. Fenghuang

122   in Anhui Province (30°88′ N, 118°02′ E) were collected, mixed and frozen in liquid nitrogen;

123   then, the sampled tissues were stored at -80°C until RNA extraction. Total RNA was isolated

124   using a TRIzolKit (Life Technologies, USA) and DNase I (TaKaRa, Japan) followed the

125   manufacturer's protocols. After total RNA was obtained, mRNA + poly(A) were isolated using

126   beads with Oligo (dT), and fragmentation buffer was added to cut the mRNA into short

127   fragments. Then, first-strand cDNA was obtained from the RNA sequence fragments using

128   reverse transcriptase and random primers (Invitrogen, Carlsbad, CA), and second-strand cDNA

129   was synthesized using buffer, dNTPs, RNaseH and DNA polymerase I. Following the ligation of

130   adapters, a single 'A' base was added to the 3' end of these cDNA fragments to facilitate end

131   repair. Based on the amplification of these products, the cDNAs were separated on an agarose

132   gel, and the cDNA library was generated.

133

134   *Sequencing and* de novo *assembly*

135   The raw reads were produced from the cDNA library using an Illumina HiSeq™2000 genomic

136   sequencer at the Beijing Genomics Institute (BGI, Shenzhen, China,

137   http://www.genomics.cn/index). The subsequent analysis was based on the clean reads generated

138   by filtering the raw reads. We used the filter_fq program (BGI, Shenzhen, China) to remove

139   reads with more than 5% unknown nucleotides (N) and low-quality sequences with more than 20%

140   low-quality bases (quality value ≤10) and adaptors to obtain clean reads. Then, we used the short

141   read assembly program Trinity (Release-2013-02-25, http://trinityrnaseq.sourceforge.net/) for *de*

142   *novo* transcriptome assembly by combining the clean reads into contigs with a sequence

143   fragment length of 200 bp (±25 bp) (Grabherr et al., 2011).Two contigs were then connected into

144   a single scaffold, and we called the resulting sequences unigenes. These unigenes were removed

145   to prevent redundancy with TGICL (version 2.1) and further spliced to generate non-redundant

146   unigenes that were as long as possible (Pertea et al., 2003). The raw sequencing data with

147   accession number SRX1600046 were deposited in the Sequence Read Archive (SRA) of

148   National Center for Biotechnology Information (NCBI), which will be released upon publication.

149

150   ***Functional annotation and classification of unigenes***

151   BLASTX alignment (E-value $< 10^{-5}$) between the unigenes and protein databases, such as NCBI

152   non-redundant protein (Nr), GO, COG, and KEGG, was performed to annotate and classify the

153   transcriptome. Based on the Nr database annotation, we used Blast2GO v2.5.0 to obtain GO

154   terms with an E-value threshold of 10^{-5} (Conesa et al., 2005). With the Web Gene Ontology

155   Annotation Plot (WEGO) (Ye et al., 2006), the distributions of GO terms were plotted to

156   describe the categories, and the unigenes were aligned to the COG database for possible

157   functional prediction and classification. The unigenes containing SSRs were also aligned to

158   euKaryotic Orthologous Groups (KOGs) via BLASTX. Finally, we mapped the unigenes to each

159   level 3 pathway graph using the KEGG database to obtain pathway annotations for the unigenes.

160

161   ***EST-SSR mining, primer design and polymorphism identification***

162   SSRs from unigenes were detected and located using MIcroSAtellite (MISA, http://pgrc.ipk-

163   gatersleben.de/misa/misa.html) (Zalapa et al., 2012). Compound SSRs (two or more SSRs in

164   which the interval was no more than 100 bp) were excluded, and only SSRs with flanking

165   sequences longer than 150 bp and containing 2 to 6 repeat motifs were considered. The mono-,

166   di-, tri-, tetra-, penta- and hexa-nucleotide motif SSRs with a minimum of 12, 6, 5, 5, 4 and 4

167   repeats, respectively, were detected. We designed primer pairs using the online program

168   Primer3.0 (http://www.onlinedown.net/soft/51549.htm) with the following criteria: (1) a product

169   sequences length of 100–300 bp and no secondary structure; (2) a primer length of 18–28 bp

170   with an optimum length of 23 bp; (3) a Tm of 55–65°C with an optimum Tm of 60°C and a

171   difference between the Tm values of the forward and reverse primers of no greater than 4°C; and

172   (4) a guanine-cytosine (GC) content of 40–60%, with 50% as the optimum. For other parameters,

173   the default settings were used.

174   Seventy-four primer pairs divided into two groups were selected for DNA amplification.

175   The first group of 50 primer pairs was randomly selected for amplification, and the motifs that

176   had more polymorphic alleles in the first group were used to increase the selected ratio in the

177   second 24 primer pairs. The mixed DNA from 3 individuals from different populations of *B.*

178   *clarkeana* was used to verify the amplification products, and the primers that amplified

179   successfully were chosen for primary polymorphism identification. Using these primers,

180 amplification was conducted using 12 individuals from 11 natural populations. Then, the DNA

181 of 128 individuals from 11 populations were amplified using primer pairs that had more

182 polymorphic loci for further identification of polymorphisms. The ESTs containing SSRs were

183 aligned to GO, COG, and KEGG databases through BLASTX to help us understand the

184 functional annotations of the sequences.

185   Using fluorescently-labeled (6-FAM, HEX, TAMRA or ROX) M13-tailed (5'-

186 TGTAAAACGACGGCCAGT-3') primers to accurately screen the variation among individuals.

187 PCR was performed in a 15-μL reaction containing 2.5 mM MgCl$_2$ and dNTP (TaKaRa, Dalian,

188 China), 0.5 U of *Taq* polymerase (TaKaRa, Dalian, China), 1× PCR buffer, and 50 ng of

189 genomic DNA. The primers included 0.04 μM forward primers, 0.01 μM M13-tailed reverse

190 primers, and 0.04 μM M13 primers with fluorescent tails. The annealing temperature was

191 different for each locus. We used 54°C as the unified annealing temperature for PCR, and the

192 amplification conditions were as follows: initial denaturation at 94°C for 5 min; 35 cycles of 30 s

193 at 94°C, 40 s annealing at 54°C, and 45 s elongation at 72°C; and a final extension at 72°C for 10

194 min. After screening on a 1.0% agarose gel, the sequence typing of successfully amplified

195 products was performed using an ABI 3730 DNA Analyzer (Applied Biosystems, Foster City,

196 California, USA). Then, we manually scored alleles using GeneMarker software (version 2.2.0).

197   Deviations from Hardy-Weinberg equilibrium (HWE) were calculated using the online

198 tool GENEPOP (http://www.genepop.curtin.edu.au/) with Bonferroni's correction. The number

199 of alleles (N_A) was calculated using MicroChecker (version 2.2.3). The expected (H_E) and

200 observed heterozygosity (H_O) of each locus were detected by GenALEx 6 (Peakall &Smouse,

201   2006), and the polymorphism information content (PIC) was calculated using the PowerMarker

202   program (version 3.25) (Liu & Muse, 2005). Then, neutral markers were detected using

203   LOSITAN (Beaumont & Nichols, 1996; Antao et al., 2008).

204

205   **Results**

206   ***Illumina sequencing and* de novo *assembly***

207   Sequencing success was determined by the length of the reads, as longer reads would increase

208   the probability of SSRs being discovered (Zalapa et al., 2012). A total of 9,361,934,460 nt bases

209   were generated in this study. After cleaning and quality checks, we obtained 104,021,494 clean

210   reads with Q20 bases (sequences with sequencing error rates <1%) at 97.55%, and the N

211   (ambiguous bases) and GC contents were 0 and 45.43%, respectively. *De novo* assembly was

212   performed using the Trinity program; a total of 94,546 contigs were generated with an average

213   length of 487 nt and an N50 value of 1,075 nt. Finally, a total of 91,449 unigenes with a total

214   length of 148,176,175 nt were detected, and the average length and N50 were 1,620 nt and 2,389

215   nt, respectively. The final assembled transcripts of *B. clarkeana* were longer than those of its

216   sibling species, i.e., the *Primulina* species and *B. hygrometrica*, which were assembled using

217   Illumina (Ai et al., 2015) and the 454 pyrosequencing platform (Zhu et al., 2015), respectively.

218   As 454 pyrosequencing produces longer reads than Illumina, the sequencing results were ideal in

219   this study (Zalapa et al., 2012). A summary of the sequence assembly after Illumina sequencing

220   is shown in Table 1. The sequence-length distribution of the unigenes is shown in Figure 1.

221

222 ***Functional annotation and classification of unigenes***

223 For function annotation analysis, we obtained 71,170, 59,962, 32,336 and 44,929 unigenes

224 annotated to the Nr, GO, COG and KEGG databases, respectively. In total, 72,078 unigenes

225 (78.82% of all assembled unigenes) were successfully annotated in the present study. This

226 number of successful annotations was more than those reported for other DT plants, including *B.*

227 *hygrometrica* (66.6% (Zhu et al., 2015) and 47.09% (Xiao et al. 2015)) and *Syntrichia caninervis*

228 (58.7%) (Gao et al., 2014), which indicates that the functions of genes in *B. clarkeana* are better

229 conserved in this study.

230

231 ***Nr annotation***

232 In total, 71,170 unigenes were annotated from 463 popular model species in the Nr databases.

233 The species distribution of Nr annotations (Figure 2) mainly comprised *Lycopersicon esculentum*

234 (35.1%), *Vitis vinifera* (27.8%), *Amygdalus persica* (6.7%), castor bean (*Ricinus communis*;

235 6.1%), black cottonwood (*Populus trichocarpa*; 5.2%), *Fragaria vesca* subsp. *vesca* (3.2%) and

236 *Glycine max* (2.8%). The most common species found in terms of this similarity were those of

237 Solanaceae, with which 6 species had similar genes (26,585, 37.35%). Only a small fraction of

238 all transcripts showed similarities to genes in other species. The structural features of the protein-

239 coding genes were similar to those of their homologs in other previously studied DT plants,

240 including *C. plantagineum* (Rodriguez et al., 2010), *B. hygrometrica* (Zhu et al., 2015) and *H.*

241 *rhodopensis* (Gechev et al., 2013). The species distribution of Nr annotations primarily consisted

242 of *V. vinifera*, *R. communis* and *P. trichocarpa*, which showed significant homology; however, *B.*

243 *clarkeana* showed some differences in our study, which indicated that *B. clarkeana* shares a

244 common origin with *L. esculentum* and *V. vinifera*.

245

246 ***GO and KEGG classification***

247 Based on the Nr annotations, 59,962 unigenes (65.57% of all unigenes) were assigned to three

248 ontologies and subdivided into 55 functional GO terms. The annotation scale in *B. clarkeana* was

249 much greater than that in the related species *B. hygrometrica* (28.71% (Xiao et al., 2015); and

250 43.7% (Zhu et al., 2015)). Similarly to previous studies, the 'Biological process' (49.45%) was

251 the main ontology, followed by 'Cellular component' and 'Molecular function' ontologies (37.11%

252 and 13.43%). A high percentage of genes were classified under the GO terms 'Cellular process',

253 'Metabolic process', 'Cell', 'Cell part', 'Organelle', 'Catalytic activity' and 'Binding' (Gupta et

254 al., 2003; Durand et al., 2010; Blanca et al., 2011; Li et al., 2012a; Xiao et al., 2015; Zhu et al.,

255 2015). The assignment of GO terms in *B. clarkeana* in this study focused on 'Single-organism

256 process', 'Physiological response to stimulus', 'Biological regulation', 'Localization',

257 'Macromolecular complex', 'Symplast' and 'Transporter activity', which reflected the functional

258 gene expression characteristics of *B. clarkeana* during normal growth. Compared with the related

259 species *B. hygrometrica* under different DT treatments, there were more functional GO terms,

260 more dispersed gene distributions, and different sets of GO terms, especially in the 'Molecular

261 function' ontology (Figure 3) (Xiao et al., 2015; Zhu et al., 2015). This result was mainly due to

262 selective gene expression caused by the adaptation of cells to various physiological states and

263 environmental changes.

264   Based on sequence homology searches against the KEGG database, 44,924 unigenes

265   (49.12% of all unigenes) were mapped in 128 pathways. The enrichment of the KEGG

266   annotation in this study was much greater than that of *B. hygrometrica* (24.43% (Xiao et al.,

267   2015); 15.1% (Zhu et al., 2015)). Among these pathways, 'Metabolic pathway' (9,232, 20.55%

268   of KEGG unigenes) and 'Metabolic biosynthesis of secondary metabolites' (3,764, 8.38%) were

269   the largest categories of 'Metabolism'. The greatest highlight of the KEGG analysis in our study

270   was the enrichment of the following vegetative dehydration/desiccation pathways: 'Plant-

271   pathogen interaction' (1,769 unigenes, 3.94% of KEGG unigenes) in the pathogen defense

272   system; 'Glycerophospholipid metabolism' (803, 1.79%) in vesicular trafficking for protein

273   receptor interactions; 'Plant hormone signal transduction' (1,783, 3.97%) for abiotic stress

274   responses; the mRNA surveillance (1,027, 2.29%) pathway for damaged transcript removal; and

275   'Photosynthesis' (154, 0.34%) and 'Nitrogen metabolism' (154, 0.34%) for the depletion of

276   transcripts during dehydration. In addition, some other environment-related pathways, including

277   'Phosphatidylinositol signaling system' (535, 1.19%), 'ABC transporters' (499, 1.11%) and

278   'Circadian rhythm-plant' (377, 0.84%) were also enriched. These results indicate that in normal

279   metabolic processes, *B. clarkeana* maintains its abundant vegetative dehydration/desiccation

280   pathways. The results of our study are consistent with those of other studies, which identified the

281   plant genes and gene products with central roles in DT (Gechev et al., 2012; Xiao et al., 2015).

282

283   ***COG and KOG classification of unigenes with SSRs***

284   In total, 56,493 functionally annotated unigenes from 32,336 (35.36% of all unigenes) COG

285   unigenes were assigned to 25 possible functional categories in COG annotations (Figure 4).

286   Among the categories, the largest group was the cluster for 'General function prediction only'

287   (10,438, 32.28%), followed by 'Replication, recombination and repair' (5,561, 17.20%) and

288   'Transcription' (5,322, 13.46%). The smallest groups were 'Cell motility' (228, 0.71%),

289   'Extracellular structures' (17, 0.05%) and 'Nuclear structure' (14, 0.04%). This pattern is similar

290   for some angiosperms, including *Camelina sativa* (Liang et al., 2013), *Apium graveolens* (Fu,

291   Wang & Shen, 2013) and *Chrysanthemum nankingense* (Wang et al., 2013). The 'Replication,

292   recombination and repair' (17.20%) category has abundant genes in *B. clarkeana*, and this plant

293   showed more repaired genes.

294   After SSR detection using the MIcroSAtellite (MISA) software with unigenes as

295   references, 7,610 unigenes carrying 8,563 SSRs were found. Then, 3,267 unigenes with SSRs

296   had hits in 24 categories of the KOG database without 'Nuclear structure'. Among 24 categories,

297   the largest group was 'General function prediction' (1,166, 35.69% of unigenes with SSRs in

298   KOG), followed by 'Transcription' (797, 24.40%), 'Replication, recombination and repair' (737,

299   22.56%) and 'Signal transduction mechanisms' (684, 20.94%). Compared with other studies of

300   EST-SSRs (Li et al., 2012a; Liang et al., 2013; Liu et al., 2013). 'Replication, recombination and

301   repair' and 'Signal transduction mechanisms' (684, 20.94%) were highlighted in *B. clarkeana*.

302   These 3,267 ESTs will provide a valuable repository of abundant information for future

303   functional SSR studies.

304

305   ***Frequency and distribution of SSRs***

306   All 91,449 assembled unigenes were used to mine potential SSRs in this study, and a total of

307   7,610 unigenes containing 8,563 SSRs were identified. Other reports have identified

308   approximately 2000 EST-SSRs using NGS (Next-Generation Sequencing) (Liu et al., 2013;

309   Wang et al., 2013; Xiang et al., 2015); the quantity of EST-SSRs in our study was significantly

310   larger, which was probably due to the use of longer reads and the expression characteristics of

311   the species (Zalapa et al., 2012).

312   Among those unigenes containing SSRs, 338 SSRs presented a compound formation, and

313   812 unigenes contained more than one SSR. On average, one SSR was found every 17.30 kb.

314   Among the identified SSRs, dinucleotide motifs were the most abundant (3,991, 46.61% of all

315   SSRs), followed by mono- (2,163, 25.26%), tri- (1,957, 22.85%), hexa- (267, 3.12%), tetra- (198,

316   2.3%), and penta- (36, 0.42%) nucleotide motifs. This result was similar to the findings reported

317   for *A. graveolens* (Fu, Wang & Shen, 2013) and *Hevea brasiliensis* (Li et al., 2012a). The

318   distributions and frequencies of different motifs are shown in Figure 5.

319   Among all SSR loci, 109 different motifs were identified. The largest subset of

320   mononucleotides were A/T (2,093, 24.44% of all SSRs), and there were only 70 C/G nucleotides

321   in total. Of the dinucleotides, AT/AT (1,564, 18.26%) and AG/CT (1,391, 16.24%) were roughly

322   equivalent, followed by AC/GT (1,035, 12.09%). Of the trinucleotides, AAG/CTT (441, 5.15%)

323   was the most common, followed by AAT/ATT (389, 4.54%) and AGC/GCT (341, 3.98%). The

324   ACAT/ATGT (18, 0.21%) motif comprised the most common tetranucleotide, and the most

325   common pentanucleotides and hexanucleotides were AAAAG/CTTTT (42, 0.49%) and

326   AAGAGC/GCTCTT (68, 0.79%), respectively. The repeat numbers of most SSRs ranged from 4

327 to 12, and the most frequent repeat number was 6 (2,066, 24.13%), followed by 5 (1,233, 14.40%)

328 and 7 (1,113, 13.00%). Furthermore, the length of SSRs ranged from 12 to 25 bp (Figure 6).

329 Among the di- and trinucleotides, the most common lengths were 12 bp and 15 bp, respectively.

330 The longest di-, tri- or tetranucleotide was 24 bp, whereas the longest pentanucleotide was 25 bp

331 in length; all hexanucleotides were 24 bp.

332

333 ***Development and validation of polymorphic SSR markers***

334 As a result, a total of 436 (only 5.73% of SSR-containing ESTs) eligible primer pairs

335 (mononucleotide, 1; di-, 191; tri-, 205; tetra-, 5; penta-, 12; hexa-, 22) were designed. Primers

336 could not be successfully designed for the other 7,174 sequences, primarily due to their overly

337 long sequence lengths and insufficient flank lengths as well as the abundance of sequences

338 containing mononucleotides. Then, 74 primer pairs (dinucleotide, 20; tri-, 38; penta-, 3; hexa-,

339 13) were selected to validate the amplification across a composite sample of 3 individuals. A

340 total of 60 primer pairs (81.08% of 74 primer pairs) showed stable and clear amplification.

341 Meanwhile, the 14 remaining pairs with failed PCR produced multiple bands or showed unstable

342 amplification. After polymorphism screening across 12 individuals, twenty-three primer pairs

343 were found to be monomorphic and 37 were found to be polymorphic. Among 37 polymorphic

344 primer pairs, 17 pairs of highly polymorphic and stable loci across 128 individuals from 11

345 populations, which covered the majority of habitats of these plants, were selected for further

346 polymorphism screening. For the 17 polymorphic loci, there were 2–6 alleles at each locus, with

347 a total of 65 alleles. The H_E, H_O and PIC per locus ranged from 0 to 0.196, 0 to 0.14 and 0.155 to

348   0.664, respectively. For the PIC values of the 17 polymorphic loci, 8 pairs had highly

349   informative scores (PIC>0.50) and 5 pairs had weakly informative scores (0<PIC<0.25) (Table

350   2).

351

352   ***Functional annotation of SSR-containing ESTs***

353   The functional annotation distribution of SSR-containing ESTs centered on ESTs containing di-

354   and tri-nucleotide SSRs (BC1–BC10). With the exception of one EST (BC12), the ESTs

355   containing penta- and hexa- nucleotide SSRs (BC11 to BC17) had almost no functional

356   annotations. However, the ESTs containing BC2, BC4 and BC12 were annotated by the KEGG

357   analysis to vegetative dehydration/desiccation pathways (Table 3 and Table S1). Thus, these

358   ESTs may contain the SSRs involved in regulating the function of DT-related genes. Although

359   the SSR variation of the functional markers did not agree with the neutral theory, the neutrality

360   test conducted using LOSITAN showed that all 17 primer pairs agreed with the neutral theory

361   (Figure 7). These 17 primer pairs may have contained exactly neutral markers, or perhaps the

362   sample size and randomness of sampling in this study was deficient. Therefore, increasing the

363   sample size in future studies will provide us with more accurate results. It should be noted that

364   compared with previous reports that identified EST-SSRs using NGS, the hexa- nucleotide SSR-

365   containing ESTs, which comprised a relatively larger number of sequences with more

366   polymorphic markers and fewer annotations, were unique to *B. clarkeana* (Liu et al., 2013;

367   Wang et al., 2013; Xiang et al., 2015).

368

369 **Discussion**

370 ***Gene expression characteristics and comparison of* B. clarkeana *with* B. hygrometrica**

371 This work is the first genetic study of *B. clarkeana* as a new plant resource of DT genes. Notably,

372 a large amount of EST data were available, enabling a better understanding of gene expression in

373 this species. *B. clarkeana* was compared with the related species *B. hygrometrica*, and both

374 plants showed KEGG enrichment of vegetative dehydration/desiccation pathways; these results

375 showed the common characteristics of metabolic pathways in DT plants. However, there were

376 some differences between the transcriptome data of these two species. First, the GC content

377 (45.43%) of *B. clarkeana* was higher than that of *B. hygrometrica*, which was close to the

378 distribution centered value of coding sequences (Matassi et al., 1989). Likely due to the lack of

379 dehydration stress, the annotated unigene percentage (78.82% of all assembled unigenes) and the

380 enrichment of GO (65.57% of all unigenes) and KEGG (49.12%) annotations in this study were

381 much greater than those of *B. hygrometrica* (Nr, 47.09%; GO, 28.71%; KEGG, 24.43% (Xiao et

382 al., 2015) and Nr, 66.6%; GO, 43.7%; KEGG, 15.1% (Zhu et al., 2015)). Second, due to different

383 sequencing depths or selective gene expression at various physiological stages, the structural

384 features of gene expression in *B. hygrometrica* was quite different under various environmental

385 pressures. Zhu et al. (2015) found that *B. hygrometrica* matched in the Nr database with *V.*

386 *vinifera*, *R. communis* and *P. trichocarpa*, whereas Xiao et al. (2015) found shared genes in the

387 genomes of *B. hygrometrica*, *Solanum tuberosum* and *Solanum lycopersicum* (Solanales). Thus,

388 Solanales and *V. vinifera* could both contain the main components of the protein-coding genes of

389 *B. hygrometrica*, which would be similar to the structural features of gene expression in *B.*

390   *clarkeana* in the present study.

391

392   ***EST-SSR characteristics of* B. clarkeana**

393   A significant number of SSRs were identified in the present study. A higher number of SSRs

394   indicated stronger environmental adaptation capabilities (Zalapa et al., 2012); therefore, *B.*

395   *clarkeana* should be highly adaptable to different environments due to the large number of SSRs

396   contained in its ESTs. Moreover, the 'Transcription', 'Replication, recombination and repair'

397   and 'Signal transduction mechanisms' reflect the strong ability of *B. clarkeana* to undergo

398   environmental adaptation.

399   Intrinsic features (such as repeat number, motif size, and length) could influence the rate

400   and probability of slippage. These features were the strongest predictors of microsatellite

401   mutability (Kelkar et al., 2008). The increased probability of slippage and mutation rates may be

402   due to, for example, a greater number of repeats (Ellegren, 2004; Kelkar et al., 2008), a greater

403   length irrespective of the repeat numbers (Webster, Smith & Ellegren, 2002), and lengths that

404   were inversely proportional to their motif sizes (Chakraborty et al., 1997). Additionally, the

405   mutation rates might vary among SSRs with different motif compositions due to the

406   dissimilarities of secondary DNA structures (Baldi & Baisnee, 2000). In this study, 37 pairs

407   (dinucleotide, 13; tri-, 13; penta-, 2; hexa-, 9) of 74 primer pairs (dinucleotide, 20; tri-, 38; penta-,

408   3; hexa-, 13) that were selected to validate the amplification results were polymorphic. The

409   percentage of polymorphism was 65% in dinucleotides (13 of the 20 selected were polymorphic),

410   34.21% (13 of 38) in trinucleotides, 66.67% (2 of 3) in pentanucleotides and 69.23% (9 of 13) in

411   hexanucleotides. As a result, in our study, SSRs with higher polymorphism rates were

412   concentrated on shorter motifs with a higher number of repeats (dinucleotides, 65%) and longer

413   motifs with fewer repeats (hexanucleotides, 69.23%; pentanucleotides, 66.67%). Our analysis

414   confirmed that mutability might increase with both increased repeat number and greater length,

415   as reported by Baldi & Baisnee (2000).

416   Compared with other SSR and EST-SSR reports (Choudhary et al., 2009; Li et al., 2012a,

417   2012b; Yuan et al., 2012; Fu, Wang & Shen, 2013), the observed number of polymorphic

418   primers was actually higher, but the polymorphism level of the markers and the H_O, H_E, HWE

419   and PIC values of the B. clarkeana population were still much lower in our study and were

420   similar to those of B. hygrometrica (Xiao et al., 2015).These results could be attributed to two

421   main reasons: first, the number of SSRs and polymorphisms of the DNA protein-coding

422   sequence was expected to be lower than that in non-coding sequences, and the mutation rate

423   within these regions was lower than that in other DNA sequences (Blanca et al., 2011; Zalapa et

424   al., 2012). Second, B. clarkeana is a plant with a short stature that requires scattered light. As

425   this plant grows on the north side of rock outcrops (mostly limestone) and in the shadow of trees

426   and shrubs (Chao et al., 2013), the long-distance dispersal potential of windborne seeds might be

427   significantly reduced. Furthermore, the occurrence of biparental inbreeding could be universal in

428   plants with high self-compatibility (Li & Wang, 2005), which would cause lower genetic

429   variability within populations of B. clarkeana.

430

431   **Conclusions**

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

433   were identified from 7,610 unigenes, 72,087 unigenes were successfully annotated to protein

434   databases, and polymorphic primer pairs of EST-SSRs were also developed. These results

435   indicated that transcriptome sequencing is a highly efficient method of EST-SSR identification in

436   non-model species that lack a reference genome and associations with functional genes.

437   Therefore, by characterizing phenotypic features, these species can be identified (Li et al., 2002).

438   These data will accelerate our identification of functional genes and genetic variation in DT

439   plants, including *B. clarkeana*. In addition, polymorphic primer pairs can continue to be

440   developed from the remaining primers of EST-SSRs. The large-scale transcriptome dataset is a

441   powerful resource for functional gene marker-assisted selection and DT exploration in *Boea*

442   plants.

443

444   **Acknowledgments**

445   We are grateful to Cunhai Li and Fei Tan (Guanshan National Nature Reserve, JiangXi, China)

446   for their assistance with the sampling.

448 **References**

449 **Ai B, Gao Y, Zhang XL, Tao JJ, Kang M, Huang HW. 2015.** Comparative transcriptome

450   resources of eleven *Primulina* species, a group of 'stone plants' from a biodiversity hot

451   spot. *Molecular Ecology Resources* **15**: 619–632. DOI: 10.1111/1755-0998.12333.

452 **Alcazar R, Bitrian M, Bartels D, Koncz C, Altabella T, Tiburcio AF. 2011.** Polyamine

453   metabolic canalization in response to drought stress in Arabidopsis and the resurrection

454   plant *Craterostigma plantagineum*. *Plant Signaling & Behavior* **6(2)**: 243–250. DOI:

455   10.4161/psb.6.2.14317.

456 **Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G. 2008.** LOSITAN: a workbench to

457   detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics* **9**: 323.

458   DOI: 10.1186/1471-2105-9-323.

459 **Baldi P, Baisnee PF. 2000.** Sequence analysis by additive scales: DNA structure for sequences

460   and repeats of all lengths. *Bioinformatics* **16**: 865–889. DOI:

461   10.1093/bioinformatics/16.10.865.

462 **Beaumont MA, Nichols RA. 1996.** Evaluating loci for use in the genetic analysis of population

463   structure. *Proceedings of the Royal Society of London Series B, Biological Sciences* **263**:

464   1619–1626. DOI: 10.1098/rspb.1996.0237.

465 **Bernacchia G, Salamini F, Bartels D. 1996.** Molecular characterization of the rehydration

466   process in the resurrection plant *Craterostigma plantagineum*. *Plant Physiology* **111(4)**:

467   1043–1050. DOI: 10.1104/pp.111.4.1043.

468 **Bianchi G, Gamba A, Limiroli R, Pozzi N, Elster R, Salamini F, Bartels D. 1993.** The

469   unusual sugar composition in leaves of the resurrection plant *Myrothamnus flabellifolia*.

470   *Physiologia Plantarum* **87**: 223–226. DOI: 10.1111/j.1399-3054.1993.tb00146.x.

471   **Blanca J, Cañizares J, Roig C, Ziarsolo P, Nuez F, Picó B. 2011.** Transcriptome

472   characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo*

473   (Cucurbitaceae). *BMC Genomics* **12(3)**: 1–15. DOI: 10.1186/1471-2164-12-104.

474   **Bockel C, Salamini F, Bartels D. 1998.** Isolation and characterization of genes expressed

475   during early events of the dehydration process in the resurrection plant *Craterostigma*

476   *plantagineum*. *Journal of Plant Physiology* **152**: 158–166. DOI: 10.1016/S0176-

477   1617(98)80127-2.

478   **Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, Huang S, Weng**

479   **Y. 2010.** Genome-wide characterization of simple sequence repeats in cucumber

480   (*Cucumis sativus* L.). *BMC Genomics* **11(1)**: 1–18. DOI: 10.1186/1471-2164-11-569.

481   **Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997.** Relative mutation rates

482   at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy*

483   *of Sciences of the United States of America* **94**: 1041–1046.

484   **Chao TC, Zhou SB, Chang LL, Chen YS, Xu HJ, Zhou QK. 2013.** Effects of light intensity

485   on the leaf morphology and physiological parameters of *Boea clarkeana*. *Chinese*

486   *Journal of Ecology* **32**: 1161–1167.

487   **Choudhary S, Sethy NK, Shokeen B, Bhatia S. 2009.** Development of chickpea EST-SSR

488   markers and analysis of allelic variation across related species. *Theoretical and Applied*

489   *Genetics* **118(3)**: 591–608. DOI: 10.1007/s00122-008-0923-z.

490 **Christ B, Egert A, Suessenbacher I, Kraeutler B, Bartels D, Peters S, Hoertensteiner S.**

491 **2014.** Water deficit induces chlorophyll degradation via the 'PAO/phyllobilin'pathway in

492 leaves of homoio-(*Craterostigma pumilum*) and poikilochlorophyllous (*Xerophyta*

493 *viscosa*) resurrection plants. *Plant, Cell & Environment* **37(11)**: 2521–2531. DOI:

494 10.1111/pce.12308.

495 **Clarke K, Gorley R. 2001.** *PRIMER v5: user manual/tutorial*. Plymouth: Primer-E Ltd.

496 **Collett H, Butowt R, Smith J, Farrant J, Illing N. 2003.** Photosynthetic genes are

497 differentially transcribed during the dehydration-rehydration cycle in the resurrection

498 plant, *Xerophyta humilis*. *Journal of Experimental Botany* **54(392)**: 2593–2595. DOI:

499 10.1093/jxb/erg285.

500 **Collett H, Shen A, Gardner M, Farrant JM, Denby KJ, Illing N. 2004.** Towards transcript

501 profiling of desiccation tolerance in *Xerophyta humilis*: construction of a normalized 11 k

502 *X. humilis* cDNA set and microarray expression analysis of 424 cDNAs in response to

503 dehydration. *Physiologia Plantarum* **122**: 39–53. DOI: 10.1111/j.1399-

504 3054.2004.00381.x.

505 **Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005.** Blast2GO: a

506 universal tool for annotation, visualization and analysis in functional genomics research.

507 *Bioinformatics* **21(18)**: 3674–3676. DOI: 10.1093/bioinformatics/bti610.

508 **Cooper K, Farrant JM. 2002.** Recovery of the resurrection plant *Craterostigma wilmsii*from

509 desiccation: protection versus repair. *Journal of Experimental Botany* **53(375)**: 1805–

510 1813. DOI: 10.1093/jxb/erf028.

511   **Durand J, Bodénès C, Chancerel E, Frigerio J-M, Vendramin G, Sebastiani F, Buonamici**

512   **A, Gailing O, Koelewijn H-P, Villani F. 2010.** A fast and cost-effective approach to

513   develop and map EST-SSR markers: oak as a case study. *BMC genomics* **11(1)**: 570. DOI:

514   10.1186/1471-2164-11-5.

515   **Ellegren H. 2004.** Microsatellites: simple sequences with complex evolution. *Nature Reviews*

516   *Genetics* **5(6)**: 435-445. DOI: 10.1038/nrg1348.

517   **Farrant JM, Brandt W, Lindsey GG. 2007.** An overview of mechanisms of desiccation

518   tolerance in selected angiosperm resurrection plants. *Plant Stress* **1**: 72–84.

519   **Fu N, Wang Q, Shen H-L. 2013.** *De novo* assembly, gene annotation and marker development

520   using Illumina paired-end transcriptome sequences in celery (*Apium graveolens* L.). *PloS*

521   *One* **8(2)**: e57686. DOI: 10.1371/journal.pone.0057686.

522   **Gaff DF. 1971.** Desiccation-tolerant flowering plants in Southern Africa. *Science* **174(4013)**:

523   1033–1034. DOI: 10.1126/science.174.4013.1033.

524   **Gao B, Zhang D, Li X, Yang H, Wood AJ. 2014.** *De novo* assembly and characterization of the

525   transcriptome in the desiccation-tolerant moss *Syntrichia caninervis*. *BMC Research*

526   *Notes* **7**: 490. DOI: 10.1186/1756-0500-7-490.

527   **Gechev TS, Benina M, Obata T, Tohge T, Sujeeth N, Minkov I, Hille J, Temanni MR,**

528   **Marriott AS, Bergstrom E, Thomas-Oates J, Antonio C, Mueller-Roeber B,**

529   **Schippers JH, Fernie AR, Toneva V. 2013.** Molecular mechanisms of desiccation

530   tolerance in the resurrection glacial relic *Haberlea rhodopensis*. *Cellular and Molecular*

531   *Life Sciences* **70(4)**: 689–709. DOI: 10.1007/s00018-012-1155-6.

532 **Gechev TS, Dinakar C, Benina M, Toneva V, Bartels D. 2012.** Molecular mechanisms of

533   desiccation tolerance in resurrection plants. *Cellular and Molecular Life Sciences* **69(19)**:

534   3175–3186. DOI: 10.1007/s00018-012-1088-0.

535 **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,**

536   **Raychowdhury R, Zeng Q. 2011.** Full-length transcriptome assembly from RNA-Seq

537   data without a reference genome. *Nature Biotechnology* **29**: 644–652. DOI:

538   10.1038/nbt.1883.

539 **Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS. 2003.** Transferable EST-SSR

540   markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular*

541   *Genetics & Genomics* **270(4)**: 315–323. DOI: 10.1007/s00438-003-0921-4.

542 **Illing N, Denby KJ, Collett H, Shen A, Farrant JM. 2005.** The signature of seeds in

543   resurrection plants: a molecular and physiological comparison of desiccation tolerance in

544   seeds and vegetative tissues. *Integrative & Comparative Biology* **45(5)**: 771–787. DOI:

545   10.1093/icb/45.5.771.

546 **Jenks MA, Wood AJ, eds. 2007.** *Plant desiccation tolerance*. Oxford, UK: Blackwell

547   Publishing Ltd.

548 **Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima**

549   **S, Okuda S, Tokimatsu T, Yamanishi Y. 2008.** KEGG for linking genomes to life and

550   the environment. *Nucleic Acids Research* **36(Database issue)**: D480–484. DOI:

551   10.1093/nar/gkm882.

552 **Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008.** The genome-wide

553   determinants of human and chimpanzee microsatellite evolution. *Genome Research* **18(1)**:

554   30–38. DOI: 10.1101/gr.7113408.

555   **Le TN, Blomstedt CK, Kuang J, Tenlen J, Gaff DF, Hamill JD, Neale AD. 2007.**

556   Desiccation-tolerance specific gene expression in leaf tissue of the resurrection plant

557   *Sporobolus stapfianus*. *Functional Plant Biology* **34**: 589–600. DOI: 10.1071/FP06231.

558   **Lehner A, Chopera DR, Peters SW, Keller F, Mundree SG, Thomson JA, Farrant JM.**

559   **2008.** Protection mechanisms in the resurrection plant *Xerophyta viscosa*: cloning,

560   expression, characterisation and role of XvINO1, a gene coding for a myo-inositol 1-

561   phosphate synthase. *Functional Plant Biology* **35**: 26–39. DOI: 10.1093/jxb/erm056.

562   **Li DJ, Deng Z, Qin B, Liu XH, Men ZH. 2012a.** *De novo* assembly and characterization of

563   bark transcriptome using Illumina sequencing and development of EST-SSR markers in

564   rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genomics* **13**: 192. DOI:

565   10.1186/1471-2164-13-192.

566   **Li M, Zhu L, Zhou CY, Lin L, Fan YJ, Zhuang ZM. 2012b.** Development and

567   characterization of EST-SSR markers from *Scapharca broughtonii* and their

568   transferability in *Scapharca subcrenata* and *Tegillarca granosa*. *Molecules* **17**: 10716–

569   10723. DOI: 10.3390/molecules170910716.

570   **Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002.** Microsatellites: genomic distribution,

571   putative functions and mutational mechanisms: a review. *Molecular Ecology* **11 (12)**:

572   2453–2465. DOI: 10.1046/j.1365-294X.2002.01643.x.

573   **Li ZY, Wang YZ. 2005.** *Plants of Gesneriaceae in China*. Henan, China: Henan Science and

574   Technology Publishing House.

575 **Li ZY. 1996.** The georaphical distribution of the subfamily Cyrtandroideae Endl. emend. Burtt

576   (Gesneriaceae). *Acta Phytotaxomica Sinica* **34**: 341–360.

577 **Liang C, Liu X, Yiu SM, Lim BL. 2013.** *De novo* assembly and characterization of *Camelina*

578   *sativa* transcriptome by paired-end sequencing. *BMC Genomics* **14**: 146. DOI:

579   10.1186/1471-2164-14-146.

580 **Liu G, Hu Y, Zhao F. 2007.** Molecular cloning of BcWRKY1 transcriptional factor gene from

581   *Boea crassifolia* Hemsl and its preliminary functional analysis. *Acta Scientiarum*

582   *Naturalium-Universitatis Pekinensis* **43**:446–452.

583 **Liu K, Muse SV. 2005.** PowerMarker: an integrated analysis environment for genetic marker

584   analysis. *Bioinformatics* **21(9)**: 2128–2129. DOI: 10.1093/bioinformatics/bti282.

585 **Liu ZP, Chen TL, Ma LC, Zhao ZG, Zhao PX, Nan ZB, Wang YR. 2013.** Global

586   transcriptome sequencing using the Illumina platform and the development of EST-SSR

587   markers in *Autotetraploid alfalfa*. *PLoS One* **8(12)**: e83549. DOI:

588   10.1371/journal.pone.0083549.

589 **Matassi G, Montero LM, Salinas J, Bernardi G. 1989.** The isochore organization and the

590   compositional distribution of homologous coding sequences in the nuclear genome of

591   plants. *Nucleic Acids Research* **17(13)**: 5273–5290. DOI: 0.1093/nar/17.13.5273.

592 **Mowla SB, Thomson JA, Farrant JM, Mundree SG. 2002.** A novel stress-inducible

593   antioxidant enzyme identified from the resurrection plant *Xerophyta viscosa* Baker.

594   *Planta* **215(5)**: 716–726. DOI: 10.1007/s00425-002-0819-0.

595   **Mulako I, Farrant JM, Collett H, Illing N. 2008.** Expression of Xhdsi-1VOC, a novel member

596   of the vicinal oxygen chelate (VOC) metalloenzyme superfamily, is up-regulated in

597   leaves and roots during desiccation in the resurrection plant *Xerophyta humilis* (Bak) Dur

598   and Schinz. *Journal of Experimental Botany* **59**: 3885–3901. DOI: 10.1093/jxb/ern226.

599   **Mundree SG, Whittaker A, Thomson JA, Farrant JM. 2000.** An aldose reductase homolog

600   from the resurrection plant *Xerophyta viscosa* Baker. *Planta* **211(5)**: 693–700. DOI:

601   10.1007/s004250000331.

602   **Neale AD, Blomstedt CK, Bronson P, Le TN, Guthridge K, Evans J, Gaff DF, Hamill JD.**

603   **2000.** The isolation of genes from the resurrection grass *Sporobolus stapfianus* which are

604   induced during severe drought stress. *Plant, Cell & Environment* **23**: 265–277. DOI:

605   10.1046/j.1365-3040.2000.00548.x.

606   **Oliver MJ, Guo LN, Alexander DC, Ryals JA, Wone BWM, Cushman JC. 2011a.** A sister

607   group contrast using untargeted global metabolomic analysis delineates the biochemical

608   regulation underlying desiccation tolerance in *Sporobolus stapfianus*. *Plant Cell* **23**:

609   1231–1248. DOI: 10.1105/tpc.110.082800.

610   **Oliver MJ, Jain R, Balbuena TS, Agrawal G, Gasulla F, Thelen JJ. 2011b.** Proteome

611   analysis of leaves of the desiccation-tolerant grass, *Sporobolus stapfianus*, in response to

612   dehydration. *Phytochemistry* **72(10)**: 1273–1284. DOI: 10.1016/j.phytochem.2010.10.020.

613   **Peakall R, Smouse PE. 2006.** GENALEX 6: genetic analysis in Excel. Population genetic

614   software for teaching and research. *Molecular Ecology Notes* **6**: 288-295. DOI:

615   10.1111/j.1471–8286.2005.01155.x.

616 **Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J,**

617   **Cheung F, Parvizi B, Tsai J, Quackenbush J. 2003.** TIGR Gene Indices clustering

618   tools (TGICL): a software system for fast clustering of large EST datasets.

619   *Bioinformatics* **19(5)**: 651–652. DOI: 10.1093/bioinformatics/btg034.

620 **Porembski S, Barthlott W. 2000.** Granitic and gneissic outcrops (inselbergs) as centers of

621   diversity for desiccation-tolerant vascular plants. *Plant Ecology* **151**: 19–28. DOI:

622   10.1023/A:1026565817218.

623 **Proctor MCF, Pence VC. 2002.** Vegetative tissues: bryophytes, vascular resurrection plants and

624   vegetative propagules. In: Black M, Pritchard HW, eds. *Desiccation and survival in*

625   *plants: drying without dying*. New York: CABI Publishing, 207–237.

626 **Rodriguez MC, Edsgard D, Hussain SS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH,**

627   **Bartels D, Mundy J. 2010.** Transcriptomes of the desiccation-tolerant resurrection plant

628   *Craterostigma plantagineum*. *Plant Journal* **63(2)**: 212–228. DOI: 10.1111/j.1365-

629   313X.2010.04243.x.

630 **Schneider H, Manz B, Westhoff M, Mimietz S, Szimtenings M, Neuberger T, Faber C,**

631   **Krohne G, Haase A, Volke F. 2003.** The impact of lipid distribution, composition and

632   mobility on xylem water refilling of the resurrection plant *Myrothamnus flabellifolia*.

633   *New Phytologist* **159**: 487–505. DOI: 10.1046/j.1469-8137.2003.00814.x.

634 **Sherwin HW, Farrant JM. 1998.** Protection mechanisms against excess light in the

635   resurrection plants *Craterostigma wilmsii* and *Xerophyta viscosa*. *Plant Growth*

636   *Regulation* **24**: 203–210. DOI: 10.1023/A:1005801610891.

637 **Wang HB, Jiang JF, Chen SM, Qi XY, Peng H, Li PR, Song AP, Guan ZY, Fang WM, Liao**

638 **Y, Chen FD. 2013.** Next-generation sequencing of the *Chrysanthemum nankingense*

639 (Asteraceae) transcriptome permits large-scale unigene assembly and SSR marker

640 discovery. *PLoS One* **8(4)**: e62293. DOI: 10.1371/journal.pone.0062293.

641 **Webster MT, Smith NG, Ellegren H. 2002.** Microsatellite evolution inferred from human-

642 chimpanzee genomic sequence alignments. *Proceedings of the National Academy of*

643 *Sciences of the United States of America* **99(13)**: 8748–8753. DOI:

644 10.1073/pnas.122067599.

645 **Xiang XY, Zhang ZX, Wang ZG, Zhang XP, Wu GL. 2015.** Transcriptome sequencing and

646 development of EST-SSR markers in *Pinus dabeshanensis*, an endangered conifer

647 endemic to China. *Molecular Breeding* **35**: 1–10. DOI: 10.1007/s11032-015-0351-0.

648 **Xiao LH, Yang G, Zhang LC, Yang XH, Zhao S, Ji ZZ, Zhou Q, Hu M, Wang Y, Chen M.**

649 **2015.** The resurrection genome of *Boea hygrometrica*: a blueprint for survival of

650 dehydration. *Proceedings of the National Academy of Sciences of the United States of*

651 *America* **112**: 5833–5837. DOI: 10.1073/pnas.1505811112.

652 **Ye J, Fang L, Zheng HK, Zhang Y, Chen J, Zhang ZJ, Wang J, Li ST, Li RQ, Bolund L.**

653 **2006.** WEGO: a web tool for plotting GO annotations. *Nucleic Acids Research* **34(Web**

654 **Server issue)**: W293–W297. DOI: 10.1093/nar/gkl031.

655 **Yuan N, Sun Y, Nakamura K, Qiu YX. 2012.** Development of microsatellite markers in

656 heterostylous *Hedyotis chrysotricha* (Rubiaceae). *American Journal of Botany* **99(2)**:

657 e43–45. DOI: 10.3732/ajb.1100304.

658 **Zalapa JE, Cuevas H, Zhu HY, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R,**

659   **Simon P. 2012.** Using next-generation sequencing approaches to isolate simple sequence

660   repeat (SSR) loci in the plant sciences. *American Journal of Botany* **99(2)**: 193–208. DOI:

661   10.3732/ajb.1100394.

662 **Zhang DD, Zhou SB, Zhou H, Liu F, Yang SY, Ma ZH. 2016.** Physiological response of *Boea*

663   *clarkeana* to dehydration and rehydration. *Chinese Journal of Ecology* **35**: 72–78.

664 **Zhu Y, Wang B, Phillips J, Zhang ZN, Du H, Xu T, Huang LC, Zhang XF, Xu GH, Li WL.**

665   **2015.** Global transcriptome analysis reveals acclimation-primed processes involved in the

666   acquisition of desiccation tolerance in *Boea hygrometrica. Plant & Cell Physiology* **56**:

667   1429–1441. DOI: 10.1093/pcp/pcv059.

669   **Tables**

670   **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

671

673   **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:GGCTTCTGATCAGGTTTCTGAAT								
BC2	F:GAGATCCCAGATCCAGATCTTCT	$(TC)_6$	3	160-164	0.038	0.023	0.192 n.s	0.423	Pr032805689
	R:AACATTAATGGAAACACGTCGTC								
BC3	F:ATTCGCTCTCTTGGTATGACTGT	$(TA)_6$	5	170-184	0.054	0.045	0.380n.s	0.664	Pr032805690
	R:CCCAATTTGAAGTGTTGCTTTAC								
BC4	F:TATCAGCGTGTGTGAATAGTTGC	$(TA)_7$	4	157-163	0.097	0.045	0.004**	0.491	Pr032805691
	R:TAACCTAAATTCGAATCCATCCA								
BC5	F:CAAACTTGGCTTAATACCATTCG	$(TG)_9$	3	119-125	0.079	0.083	0.713 n.s	0.469	Pr032805692
	R:CCATGATCATCTCTATTTCAGGC								
BC6	F:CCTTAAGGAGATGCATTGTGAAT	$(TC)_9$	3	159-169	0.000	0.000	- n.c.	0.299	Pr032805693
	R:GTATGAAGGGCATCAACAATAGG								
BC7	F:GCTGAAAGTTGGTGATTGCTAGT	$(AT)_9$	4	166-178	0.120	0.125	0.087 n.s	0.526	Pr032805694
	R:AGTTATGTCTTCGCTTGCTTCAG								
BC8	F:AACGTGAGAGTGCTAGTTCGGTA	$(TGA)_5$	3	167-173	0.014	0.000	0.041*	0.17	Pr032805695
	R:TCTTCCTCACTTTATCATCCACG								
BC9	F:AGAAGAGGTACGACAGTTTGCTG	$(GCG)_5$	2	156-159	0.059	0.064	1.000 n.s	0.195	Pr032805696
	R:TTCACGTCCGAATTCTTAGTCTC								
BC10	F:CACTGCACATAGAAGGAGGAGTT	$(GCG)_6$	5	108-129	0.081	0.076	0.146 n.s	0.581	Pr032805681
	R:GTAATCGCCTACATGATTCATCC								
BC11	F:CAGCAGTATGTCGGGATTATTTC	$(TTTCT)_4$	2	123-133	0.000	0.000	-n.c.	0.155	Pr032805682
	R:CCTCTGGTCATATTGCTGTTACC								
BC12	F:AACAAGAGGGTCAGCTACAACAG	$(CAGCAA)_4$	4	160-178	0.104	0.095	0.184 n.s	0.549	Pr032805683
	R:CAGCAATGGTATTAGCAGAGGAC								
BC13	F:ACCTTGACGATCCTTCATCTTCT	$(GGTGCG)_4$	6	132-174	0.124	0.095	0.161 n.s	0.701	Pr032805684
	R:TTATGTTCTCCATATCCGTCAGC								
BC14	F:GGCAGCAATATAGCTCAAATACG	$(GACAAG)_4$	4	170-188	0.196	0.083	0.000***	0.516	Pr032805685
	R:ACCTGATCGTTCACAACTTCATC								
BC15	F:TCTTATTCAACACAACAGCCTGA	$(ATGATA)_4$	5	151-175	0.157	0.140	0.228 n.s	0.528	Pr032805686
	R:GCTGCAGTTGATAATGAGAAGGA								
BC16	F:ACCAATGGTCTATATTTCAACGG	$(ATTACT)_4$	6	149-179	0.132	0.125	0.174 n.s	0.643	Pr032805687
	R:TGTGCCCCACATAGCTTCTATCTA								
BC17	F:TGACGAGGCTTCTACAGAATGAG	$(CATCCT)_4$	2	137-143	0.034	0.045	1.000 n.s	0.186	Pr032805688
	R:ACAAACAACAAGATGGGAATCAT								

674   *Note:* N_A = number of alleles per locus across all populations; H_E = expected heterozygosity

675   (mean value); H_O = observed heterozygosity (mean value); PIC, polymorphic information

676   content; HWE = Hardy-Weinberg equilibrium. [a]After Bonferroni correction, significant

677   departures from Hardy-Weinberg equilibrium are indicated by * $P<0.05$, ** $P<0.01$, ***

678     $P<0.001$. n.s. = not significant; n.c. = not calculated (Clarke & Gorley, 2001).

679   **Table 3 GO, COG and KEGG annotation of 17 SSR-containing ESTs.**

Unigene*	GO**			COG	KEGG
	b. p.	c. c.	m. f.		
BC1	-	6	-	function unknown	-
BC2	1	-	-	-	plant hormone signal transduction
BC3	-	-	-	-	-
BC4	7	-	3	general function prediction only signal transduction mechanisms transcription replication,recombination, and repair	plant hormone signal transduction plant-pathogen interaction
BC5	-	-	-	-	-
BC6	-	-	-	general function prediction only	-
BC7	2	2	-	general function prediction only	biosynthesis of secondary metabolites amino sugar and nucleotide sugar metabolism
BC8	12	4	4	general function prediction only signal transduction mechanisms transcription replication, recombination and repair	-
BC9	-	-	2	secondary metabolites biosynthesis, transport and catabolism	biosynthesis of secondary metabolites flavonoid biosynthesis flavone and flavonol biosynthesis sesquiterpenoid and triterpenoid biosynthesis isoflavonoid biosynthesis
BC10	2	4	1	cell cycle control, cell division, chromosome partitioning	metabolic pathways endocytosis ether lipid metabolism
BC11	-	-	-	-	-
BC12	10	3	2	general function prediction only posttranslational modification, protein turnover, chaperones	plant hormone signal transduction
BC13	-	-	-	transcription	-
BC14	-	-	-	-	-
BC15	9	5	2	-	-
BC16	-	-	-	-	-
BC17	-	-	-	-	-

680   *Note:* b. p.= biological_process; c. c.= cellular_component; m. f.= molecular_function. * The
681   name of each unigene is replaced with the name of the EST-SSR maker it contains. ** This table
682   shows only the number of the GO terms in the ontology for unigenes. The details of the GO
683   classification in three ontologies are shown in Supplemental Table S1.
684

685

Figures

687

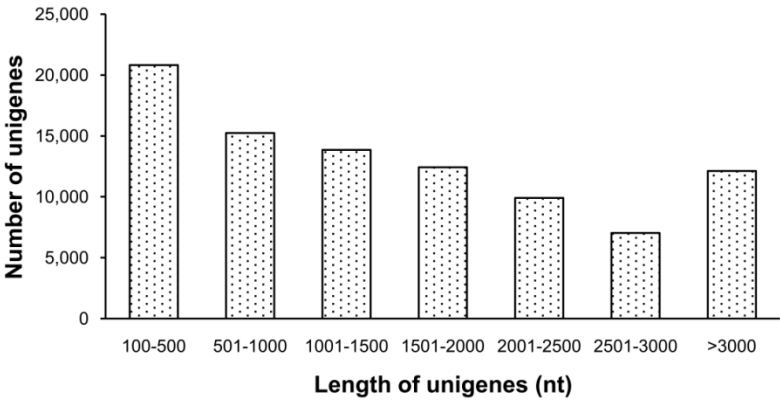



Figure 1 The length distribution of unigenes.

689

690

691

692

Figure 2 The species distribution of Nr annotations.

694

695

Figure 3 Gene ontology classification of unigenes between *B. hygrometrica* and *B. clarkeana*.

697

698 GO functions are shown in the X-axis. The Y-axis shows the percentage of genes annotated with

699 the GO function. RWC = relative water content; HD= hydrated.

700

701

702

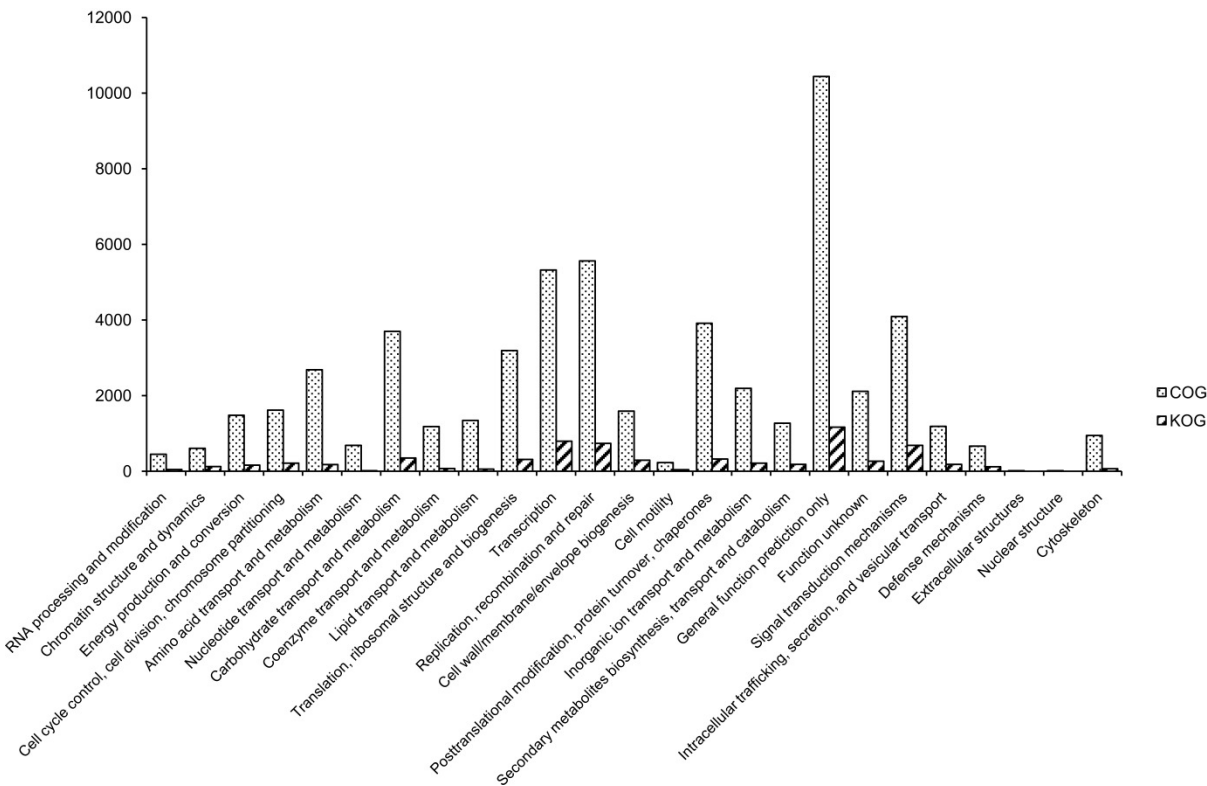
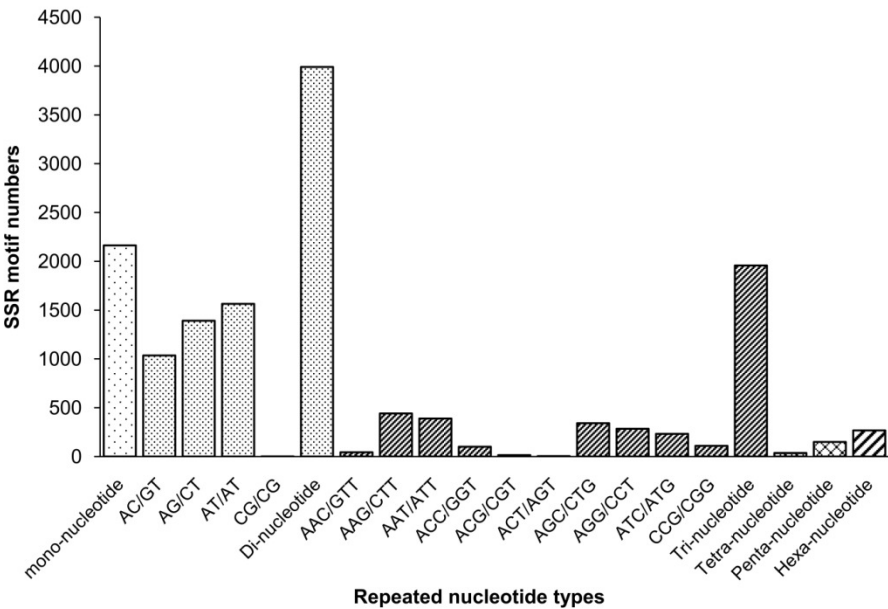



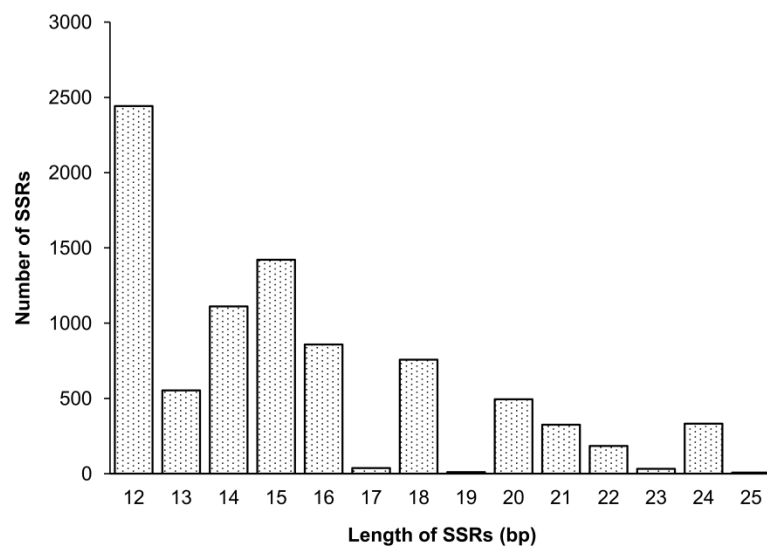
Figure 4 COG and KOG functional classification of unigenes.

The horizontal coordinates are functional classes of COG and KOG, and the vertical coordinates
are numbers of unigenes in one class.

706

707

708 **Figure 5 The distribution of the most repeated nucleotide types.**

709

710

711 **Figure 6 The distribution of SSRs of different lengths.**

712

713

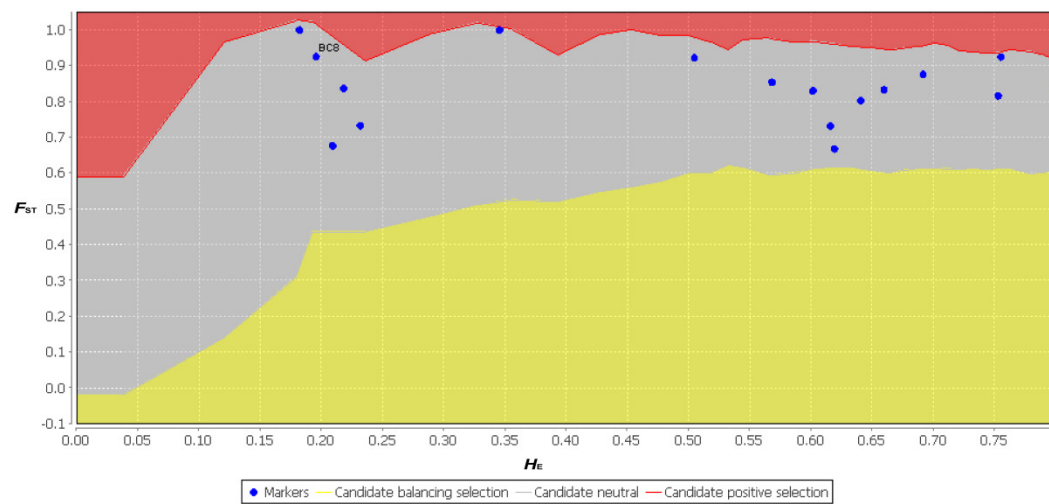



Figure 7 The neutral test results for 17 primer pairs using F_{ST} and H_E from 11 populations

715 **using LOSITAN.**

716