Genetic evaluation and core collection construction of crape myrtle accessions using newly developed EST-SSR markers

Yuanjun Ye[1], Lu Feng[1], Yiqian Ju[1], Ming Cai[1], Tangren Cheng[1], Jia Wang[1], Qixiang Zhang[1], Huitang Pan[1]

[1]Beijing Key Laboratory of Ornamental Plants Germplasm Innovation & Molecular Breeding, National Engineering Research Center for Floriculture, Beijing Laboratory of Urban and Rural Ecological Environment, Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants of Ministry of Education and College of Landscape Architecture, Beijing Forestry University, Beijing, 100083, China

Corresponding Author:
Huitang Pan[1]

Email address: htpan@bjfu.edu.cn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

ABSTRACT

Crape myrtle is an important ornamental woody plant, due to its long-lasting mid summer bloom and rich color. However, limited molecular markers on this species hinder the breeding and genetic studies. In this work, 8,652 EST-SSRs were identified from crape myrtle transcriptome data. Di-nucleotide repeats (57.1%) were the most abundant type followed by tri-, tetra-, penta-, hexa-nucleotide repeats, with the AG/CT motif occurring most frequently. Of the 1200 synthesized primer pairs, 761 EST-SSRs (63.4%) were successfully amplified and 245 EST-SSRs (20.4%) showed polymorphic. High cross-species transferabilities of these markers were observed except in *L. speciosa* (26.7%). The polymorphic information content (PIC) for each locus ranged from 0.210 to 0.813 with a mean of 0.589, suggesting a high level of informativeness. Using 30 polymorphic EST-SSRs, structure and cluster analyses roughly divided the 73 accessions into three major groups with some admixtures. Based on the SSR data and clustering analysis, a final core collection (20 accessions) was identified, which captured Na, Ne, I, and PIC value with a retention rate of 92.8%, 113.6%, 110.6% and 109.7%, respectively. Thus, this work contributes to the better understanding of the genetic diversity and germplasm resources conservation in *Lagerstroemia* species.

INTRODUCTION

Crape myrtle (*Lagerstroemia indica* L.) is native to Southeastern Asia and Australia that belongs to the Lythraceae family and comprises more than 50 species (*Graham et al., 2005*), which is regarded as an important ornamental woody flower for its diverse plant type, durable bloom period and rich color (*Knox, 1992*). In addition, crape myrtle is becoming an important source of income for flower enterprises and self-employed nursery growers (*Guidry & Einert, 1975*). Because of its value and wide range of uses, *Lagerstroemia* species have been planted in Indo-Malayan and Southeast Asia regions for many years (*Pooler, 2006*). Since the 1960s, many new cultivars with different plant architectures, beautiful flowers and disease resistance have been developed (*Egolf, 1981, 1986; Pooler, 2006; Wang et al., 2014*). Subsequent crape myrtle breeding is primarily focused on interspecific hybridization between *L. indica* and *L. fauriei*, especially for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

breeding cultivars with early blooming and disease resistance (*Pounders, Rinehart & Sakhanokho, 2007*). To date, hybridization is an important method to improve the ornamental traits of horticultural plants. However, it is hindered by the ambiguous germplasm resources evaluation and long selection cycle.

Genetic diversity estimate is critical for crape myrtle breeding, germplasm management, and conservation strategies because of inbreeding depression (*Pounders, Reed & Pooler, 2006*). Compared to the traditional morphological evaluations, DNA markers will reflect the real genetic diversity because of the lower environment influence. In *Lagerstroemia*, the first genetic diversity analysis was conducted on 12 clones of *L. fauriei* materials using RAPD and AFLP markers (*Pooler, 2003*). *He et al.* (*2012*) genotyped all the 96 *Lagerstroemia* samples and devided them into three distinct groups based on genetic distance. Thus, it provides a shortcut to perform the genetic studies and improvement of these valuable traits in crape myrtle by molecular markers (*Ye et al., 2016*).

Compared with RAPD and AFLP markers, SSRs have always been preferable to others for their co-dominant inheritance, multiple alleles, stable reproducibility and high transferability (*He et al., 2003; Agarwal, Shrivastava & Padh, 2008*). Dating back the genetic studies in *Lagerstroemia*, SSRs were extensively employed in the genetic evaluation (*Wang et al., 2010*), characterization and identification of germplasm resources (*Cai et al., 2010*), gene mapping (*Ye et al., 2015*) and genetic linkage map construction (*He et al., 2014*). However, fewer than 150 SSR markers in crape myrtle have been published to date, which hinders the development of in-depth genetic investigation in crape myrtle. Surveyed with *L. tomentosa*, the genome size of *Lagerstroemia* species was estimated to be up to 965 Mb (*Wang et al., 2015*). The whole genome of crape myrtle can not be covered and equally distributed by the limited markers. Therefore, it's pressing to develop massive SSR markers for marker-assissted selection in crape myrtle breeding.

Based on the original sequences of SSR markers, SSRs included genomic SSR (g-SSR) and expressed sequence tag SSR (EST-SSR). The tranditional techniques such as biotin-streptavidin capture method were usually based on a double-enriched microsatellite library and sequencing of the SSR colonies (*Eujayl et al., 2004*). Without genomic

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

information, the development of g-SSRs is time-consuming, expensive and laborious in productivity. In contrast, EST-SSR markers can be rapidly identified from expressed sequence at a low cost (*Zhou et al., 2014*). With the development of next-generation sequencing (NGS) technologies, we can obtain a large number of EST-SSRs from high-throughput transcriptome data cost-effectively. Recently, numerous EST-SSR markers were mined by transcriptome sequencing in various plants, which were proved to be effective and more conserved compared to non-coding sequences (*Wu et al., 2014; Liu et al., 2015; You et al., 2015*). Moreover, some important horticultural traits can be directly mapped using EST-SSRs due to their association with coding region (*Bouck & Vision, 2007*).

The management of precious accessions in germplasm collection is of importance to conserve their genetic diversity. However, the redundant genotypes, heterogeneous structure, and unavailable information on trait diversity affect the successful utilization of the genetic potential of these collections (*Xu et al., 2016*). Because it is difficult to characterize the whole samples completely due to the cost in labor, space and time, building a core collection with small asccessions capturing the genetic information of the intial collections is recommended. Based on the genetic diversity analysis, this core subset could cover the maximum genetic diversity of the resources using the representative genotypes (*Brown, 1989*). To date, dozens of core collections have been successfully established, including *Arabidopsis thaliana* (*McKhann et al., 2004*), *Solanum pimpinellifolim* (*Rao et al., 2012*), *Cucumis sativus* (*Lv et al., 2012*), *Cucumis melo* (*Hu et al., 2014*), *Malus* × *domestica* (*Richards et al., 2009; Liang et al., 2015*), and *Ziziphus jujuba* (*Xu et al., 2016*). To date, no core collection of crape myrtle have been constructed.

Herein, the aims of our study were to (1) develop EST-SSRs from transcriptome data on a large scale, (2) evaluate the genetic diversity of 73 accessions using these markers , (3) build a core collection as the representative germplasm resources of the entire population.

MATERIALS & METHODS

Plant materials and DNA extraction

To test the amplification efficiency of the selected EST-SSRs, the F1 segregating

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

population was employed in this study (*Ye et al., 2016*). Seven *Lagerstroemia* species were used to confirm the cross-species transferability of selected EST-SSR markers, including *L. indica*, *L. fauriei*, *L. speciosa*, *L. excelsa*, *L. caudata*, *L. limii* and *L. subcostata*. Meanwhile, the phylogenetic relationship of 73 accessions was analyzed using 30 highly informative EST-SSR markers. The list of the 73 accessions with their origin, growth habit and flower color were provided in Supplemental Table S1.

Total genomic DNA was isolated from fresh young leaf tissues using the CTAB method. The DNA quality and quantity were estimated by 1% agarose gel electrophoresis at 0.1μg/mL 1× TAE buffer and Unico UV-visible Spectrophotometer (Unico, USA), respectively. The DNA was adjusted to 50 ng/μL for polymerase chain reactions (PCR) amplification.

SSR detection and primer design

In our previous studies, the F1 population was used to investigate the genetic inheritance of plant architecture traits (*Ye et al., 2015, 2016*). Herein, we performed the transcriptome analysis for the parents in order to screen the target genes, as well as the massive EST-SSR markers. Illumina sequencing was conducted at Novogene Bioinformatics Technology (Beijing, China). Total RNA was isolated from the young leaves using the RNeasy Plant Mini Kit (Qiagen, Germany) according to the manufacturer's instructions. The crape myrtle RNA was used to constructed the cDNA libraries with fragment length of 200bp (±25bp). Then paired-end was sequenced using Illumina HiSeq^{TM} 2500 (Illumina, San Diego, CA). After obtaining clean data, trinity software was used to assemble the transcriptome sequences for the high quality reads ($Q < 20$). Only these stringently compiled sequences were defined as unigenes. The Perl script MISA (http://www.pgrc.ipkgatersleben.de/misa) was used to identify SSRs from unigene database. The searching principle was that only di-, tri-, tetra-, penta- and hexa-nucleotides with a minimum of 8, 5, 5,4 and 3 repeats will be considered as SSR loci. The primers of each EST-SSRs were designed using Primer Premier 5.0 (Premier Biosoft International, CA, USA) with following criterias: PCR fragment size- 100-350 bp; primer length- 16-24 bp; Tm- 55-65 °C; GC content- 40-60 %.

Development of EST-SSR markers

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

The EST-SSRs were initially tested for PCR amplification by 1% agarose gel electrophoresis among 6 samples of F1 population (i.e. both parents and four individuals). The polymorphism of successfully amplified markers was estimated by 8% polyacrylamide gel electrophoresis (PAGE). And then a set of 30 primer pairs were selected according to their polymorphism and identification ability, in order to analyze the genetic diversity among 73 accessions, and transferability across seven *Lagerstroemia* species. The PCR amplification conditions and procedures were referenced from *Ye et al.* (*2016*). For analysis of genetic relationship, the forward primers of SSRs were elongated from M13 universal sequences appended to the 5'-end (*Schuelke, 2000*). The amplification conditions and IRDye label procedures were described by *He et al.* (*2012*). The PCR fragments (0.5 μL) with different sizes and fluorescent labels were pooled and analyzed on an ABI3730xl DNA Analyzer (Applied Biosystems, USA).

Statistical and genetic analysis

Microsatellite alleles were corrected using FlexiBin v 2 and GeneMarker v 2.20 (SoftGenetics, State College, Pennsylvania, USA). For polymorphism eavluation of each SSR locus, allele number and the polymorphic information content (PIC) were caculated using Popgene v 1.32 software (*Yeh, Yang & Boyle, 1999*). The summary statistics which reflected the degree of polymorphism, including the observed number of alleles per locus (Na), the number of effective alleles (Ne), the observed heterozygosity (Ho), the expected heterozygosity (He), and the shannon's information index (I) were analyzed using the Microsatellite toolkit v 3.1.1 and GenAlEx 6.5 (*Peakall & Smouse, 2012*).

The population structure of 73 crape myrtle accessions was analyzed using STRUCTURE v 2.3 (*Pritchard, Stephens & Donnelly, 2000*), based on the Bayesian clustering analysis according to the expected Hardy-Weinberg equilibrium and absence of linkage disequilibrium between the analyzed loci in each population. For each possible value of K (2-8), ten repetitive runs were performed with 500,000 Markov chain Monte Carlo (MCMC) iterations following a burn-in period of 200,000 steps. To identify the optimal number of the clusters, the delta K method (*Evanno et al., 2005*) was employed in STRUCTURE HARVEST (*Earl & Vonholdt, 2012*). The barplot of the probability of the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

membership from the results of STRUCTURE was visualized by the CLUMPAK (*Kopelman et al., 2015*). Genetic distances were calculated using shared allele distance to create a matrix by PowerMarker v 3.25 (*Liu & Muse, 2005*). Cluster analysis was conducted to show the relationship among 73 accessions using an unweighted pair group method with an arthmetic mean (UPGMA) and Nei's unbiased genetic distance with the FreeTree program and the TreeView software package.

Core collection construction

The construction of a core collection was conducted as described by *Xu et al.* (*2016*) with some modifications. Based on the number of accessions, we chose a progressive sample strategies to identify the core subset, in which 9 core collections were developed to confirm the optimal size. Amongst all the core subsets, several important accessions in this species were selected as the retained accessions (e.g. *L. fauriei*, *L. speciosa*, *L. subcostata* and *L. caudata*). To insure the accuracy of the core subset construction, five repetitive runs were processed using two different methods by PowerMarker v 3.25 software, including simulated annealing and random search. The PowerCore software (*Kim et al., 2007*) was employed to further screen for the results. The analyses were repeated 1000 times until the representativeness met the requirement of a core subset or the appropriate number of accessions was achieved. Finally, a T-test for Na, Ne, I, Ho, He and PIC value was performed to determin the correlation between the core subset and the initial collection using SPSS v18.0 (SPSS, Chicago, IL, USA).

RESULTS

Summary and characterization of EST-SSR markers

A total of 8,652 SSRs were identified from the 93,161 examined EST sequences, of which 1,432 sequences contained at least one SSR. Among the diverse types of repeats, di-nucleotide motifs were the most abundant (4,932, 57.1%) with tri- (3,456, 39.9%), tetra- (212, 2.5%), penta- (29, 0.3%) and hexa- (13, 2%) nucleotide being the next most common in consecutive order (Table 1).

The frequency distribution of major repeats with di- and tri-nucleotide units was also analyzed in the present study. Among the di-nucleotide motifs, AG/CT (42.1%) was the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

highest abundant repeat types, followed by GA/TC (34.0%), AT/TA (13.7%), AC/GT (5.1%), CA/TG (4.8%) and CG/GC (0.3%) (Fig. 1a). Among the tri-nucleotide repeats, the richest motifs were GAA/TTC (11.6%) and GGA/TCC (9.8%) (Fig. 1b). The repeat motif number of these SSRs ranged from 5 to 12, and SSRs with six repeats (32.5%) were the most abundant followed by five (26.7%), seven (17.1%), eight (11.3%), nine (8.0%) and others (4.4%) (Fig. 1c).

Development of polymorphic EST-SSRs

A total of 1,200 EST-SSR primers were finally synthetized based on the program criterions, including 331 (27.6%) di-, 714 (59.5%) tri- and 155 (12.9%) other type motifs. Of the tested markers, 761 primer pairs (63.4%) (Supplemental Table S2) were successfully amplified with the expected sizes, whereas 49 PCR products showed larger than the expected sizes, indicating that an intron may exist within the amplified regions. Given the remaining SSRs could not generate any bands, they were not chosen for further analysis. Of the successfully amplified EST-SSR markers, 245 primer pairs (20.4%) showed polymorphic in six crape myrtle accessions (Fig. 2).

Polymorphism dectection and cross-species transferability

The 30 informative EST-SSRs were selected to analyze polymorphism in 73 accessions, and showed high discriminating capacity, as deduced from a low cumulative identity probability (PI) of 1.3E-24 (Table 2). Of all the SSR markers, YYJ-283 yielded the highest identity probability of 4.9E-02 and YYJ-129 yielded the lowest identity probability of 6.2E-01.

A total of 223 polymorphic bands were discovered, ranging from 4 (YYJ-92/YYJ-337) to 13 (YYJ-693) with an average of 7.433 per primer (Table 3). The mean value of Ne, Ho, He and I were 3.242, 0.536, 0.626 and 1.321, respectively. Moreover, PIC as an important index reveals the genetic diversity of the test markers. The PIC value ranged from 0.210 (YYJ-695) to 0.813 (YYJ-283) with a mean of 0.589, indicating that the highly polymorphic EST-SSRs would be employed to perform the genetic analysis in *Lagerstroemia* species.

The newly developed EST-SSRs were then used to assess cross-species conservation

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

and transferability (Table 4; Fig. 3). Of all the *Lagerstroemia* species, the markers showed high transferability except in the *L. speciosa* with a transferability ratio of 26.7%. Successful cross-species amplification was accomplished in other species, with about 93.3% of the markers in *L. indica*, *L. excelsa*, *L. fauriei* and *L. subcostata*, 86.7% in *L. caudata* and 83.3% in *L. limii*. Seven markers (YYJ-706, YYJ-199, YYJ-187, YYJ-201, YYJ-166, YYJ-148 and YYJ-356) exhibited perfect cross-species transferability in all the *Lagerstroemia* species, indicating these markers could be employed as anchored markers for parentage identification and genetic evolution studies in the Lythraceae family.

Population structure and cluster analysis

In the Bayesian model-based cluster analysis of population structure, the delta K approach suggested a clear peak at K = 3 (Fig. 4a), where the whole individuals were divided into three major groups. In addition, a large number of accessions showed mixed ancestry (membership values lower than 80%) (Fig. 4b). Group 3 contained the highest number of samples (32), followed by Group 1 (24) and Group 2 (17). The individuals in Group 1 were referred as the hybrids between *L. fauriei* and *L. indica*, whereas the individuals in Group 3 possessed the only ancestry of *L. indica*. Particularly, Group 2 accounted for several important accessions in this species, from which most of them were big arbors. The highest value of genetic parameters including Nt, Na, Ne, Ho, He, % P and I were identified in Group 2, whereas the lowest were found in Group 1 (Table 5). At population level, 80 private alleles (Np) were detected at 28 loci distributed in the three populations, with the frequencies ranged 0.016 to 0.400. For all the individuals, 12 private alleles were detected in *L. subcostata*, followed by *L. excelsa* (10 private alleles), *L. limii* and *L. caudata* (8 private alleles). Approximately 56% of the private alleles were examined in Group 2, indicating that individuals in this group possessed informative genetic diversity and may have a unique ancestry type.

The genetic relationship between the accessions was performed based the 30 EST-SSR loci, in which 73 accessions were divided into three clusters (Fig. 5). Overall, the dendrogram corroborated the results of STRUCTUTE analysis with some exceptions in three clades. The accessions in Cluster 2 were distributed at the extremely advanced

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

position of the dendrogram, suggesting a special evolutionary relationship in this species. Cluster 1 and 3 consisted of all the cultivars of the *L. fauriei* hybrid, *L. indica* hybrid and *L. limii* hybrid, which exhibited a great consistency with the origins and previous studies.

The construction of core subset

Nine core collections were constructed using two sampling strategies based on the SSR data and cluster analysis, which accounted for approximately 8%, 11%, 14%, 16%, 19%, 22%, 25%, 27% and 30% of the total accessions, respectively (Table 6). Aiming to identify a core subset with a minimal amount of accessions that retain the maximal genetic information and best represent the entire genotypes, we evaluated the Na, Ne, I, and PIC value of each subset to select the most suitable core collection. Compared with other candidate collections, core collection 8 (20 accessions) captured a higher Na, Ne, I, and PIC value with a retention rate of 92.8%, 113.6%, 110.6% and 109.7%, respectively. The 20 core individuals were devided into two clusters, in which five species were grouped into Cluster 1, whereas Cluster 2 comprised fourteen *Lagerstroemia* cultivars (Fig. 6). All the genetic parameters were calculated using SPSS 18.0 program, indicating no significant differences between the core and entire collection ($P < 0.05$). The allele frequency in the core subset and entire collection was highly correlated ($R^2 = 0.925$), demonstrating the best representation of core collection (Fig. 7).

DISCUSSION

Transcriptome analysis by next-generation sequencing has been widely used to discover new genes and develop molecular markers in many plants. Particularly, such a powerful technique characterized by high throughput, high accuracy and low cost can be employed in model or non-model species (*Deng et al., 2016*). To date, the limited SSR markers has seriously hindered the development of marker-assissted breeding in crape myrtle. In the present study, large-scale EST-SSRs in the transcriptome of *L. indica* were developed and characterized. Our results showed that a total number of 8,652 SSRs were identified from the 93,161 examined EST sequences and the di-nucleotide motifs were the highest abundant, which were consistent with the studies in *Pinus contorta* (*Parchman et al., 2010*) and blueberry (*Rowland et al., 2012*). However, *Wang et al.* (*2015*) found the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

tetra-nucleotide microsatellite repeats were the most frequent type in the crape myrtle geneme, followed by the di-nucleotide motifs. Similar studies have reported the different results with tri-nucleotide motifs being the most abundant type (*Qiu et al., 2010; Niu et al., 2013*), suggesting that the dominant type of SSRs may vary among different strategies and species. Of all the di-nucleotide repeats, the highest abundant repeat motifs were AG/CT, which was also revealed in sweet potato (*Wang et al., 2011*) and radish (*Zhai et al., 2014*). Furthermore, AT-rich and GC-rich repeats were detected in intron and exon regions for the splice site recognition in plant genes (*Amit et al., 2012*). Our result showed a low frequency of GC repeat units, which was consistent with the findings in various species (*Aggarwal et al., 2007; Zeng et al., 2010*). Nevertheless, the AT content (13.7%) was not agreement with the results of previous findings in crape myrtle (60.8%). The scenario can be speculated that the frequency of SSR motifs strongly depends on the size of analyzed databases, SSR search creteria and inequable mining tools (*Varshney, Graner & Sorrells, 2005; Biswas et al., 2012*).

Experimental analysis for 761 SSRs in this study showed a higher rate of successful amplification with expected fragment (63.4%) than revealed in other species, such as *Taxodium* (*Cheng et al., 2015*) (51.1%), tree penoy (*Wu et al., 2014*) (47.3%), suggesting that the transcriptome sequencing was accurate and the assembled unigenes were of high quality. However, 49 PCR products showed larger than the expected sizes, which probably be due to the existence of long intervening introns, large insertion fragments or repeat number variations, or assembly errors (*Wei et al., 2013*). Of the successfully amplified EST-SSR markers, 245 primer pairs (20.4%) showed polymorphic in both parents and one mapping population. The ratio of polymorphic EST-SSR markers was lower than genomic-SSRs in crape myrtle (36.4%, *Cai et al., 2010*; 27.9%, *Wang et al., 2015*), which may be due to the highly conservative coding region of EST sequences . However, the mean number of alleles (Na) of 30 SSR loci (7.433) had a higher degree of polymorphism compared with the g-SSRs (5.58, *Cai et al., 2010*; 5.58, *Wang et al., 2015*). The reasons can be explained by the sample numbers and the different geographic origins.

The selected EST-SSRs were used to perform genetic analysis between the 73

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

accessions, which showed high discriminating capacity as deduced from a low cumulative identity probability (PI) of 1.3E-24. Generally, the PIC value reflects the informativeness degree of the markers and are classified as high (PIC > 0.5), moderate (0.5 < PIC > 0.25), and low (PIC < 0.25) (*Bostein et al., 1980*). A high PIC value among 22 primer pairs (73.3% of all loci) indicated that these markers could be useful for assessing the population structure, genetic diversity and relationship in *Lagerstroemia* species. The abundance of polymorphism probably be due to the complicated genetic background of collected germplasms or the contingency of highly polymorphic SSR markers being selected.

The newly developed EST-SSRs were then selected to assess cross-species conservation and transferability in 7 species of *Lagerstroemia* genus. In total, perfect cross-species amplifications were accomplished in most species, with about 93.3% of the markers in *L. indica*, *L. excelsa*, *L. fauriei* and *L. subcostata*, 86.7% in *L. caudata* and 83.3% in *L. limii*. This perfect transferability of EST-SSRs in crape myrtle was partly resulted from the moderate conservation of the sequences flanking the SSR among these 7 accessions. However, only 8 out of 30 primer pairs successfully amplified expected fragments in *L. speciosa*, indicating that it differed from other species evolutionarily. Seven markers (YYJ-706, YYJ-199, YYJ-187, YYJ-201, YYJ-166, YYJ-148 and YYJ-356) exhibited perfect cross-species transferability in all the *Lagerstroemia* species.Therefore, the novel and powerful EST-SSRs can be employed as an effective tool for comparative mapping, parentage identification and genetic evolution analyses in the future study.

The population structure and genetic diversity were investigated using 30 polymorphic EST-SSRs in the entire collection of 73 accessions. Our results showed that the grouping in STRUCTURE was greatly consistent with the cluster analysis. Despite several exceptions existing in the three populations, the cluster differentiation is convictive. The dendrogram analysis demonstrated that all the accessions were mainly clustered together based on their growth habit and origin, which was partly similar to the findings of *Pooler* (*2003*), and *He et al.* (*2012*). Similarly, near all the crape myrtle cultivars were grouped together while the *Lagerstroemia* species and their interspecific hybrids were clustered together, indicating the clustering was likely to reveal the shared pedigrees or the same breeding strategies.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

However, several exceptions still exist, i.e., the *L. indica* cultivar No. 36 was clustered into the interspecific hybrids (Cluster 1) and No. 8, 26, 32 and 33 were clustered into the *L. indica* cultivars (Cluster 3). Because they were purchased, transported and propagated at the same time, mislabeling could have occurred due to the indiscernible flower and plant type. Wild crape myrtle species such as *L. excelsa* and *L. limii* are precious genetic resources that should be carefully stored and evaluated. Among the seven species, only *L. fauriei* was devided into Cluster 1 with interspecific hybrids, whereas the remaining species were devided into Cluster 2 with arbor trees. The conclusion was completely consistent with the findings of *He et al.* (*2012*), in which *L. subcostata* and *L. limii* were grouped together, while *L. caudata* and *L. speciosa* were clustered closely.

Moreover, we found that Group 2 possessed the higher value of genetic diversity than Group 1 and 3 in STRUCTURE analysis, indicating that individuals in this group captured abundant genetic information. From all the loci detected, 12 private alleles were detected in *L. subcostata*, followed by *L. excelsa* (10 private alleles), *L. limii* and *L. caudata* (8 private alleles). Given the extremely advanced position of the dendrogram, it can be speculated these accessions shares a unique ancestry type in this species. As a consequence, necessary strategies need to more sharply focused on protecting these rare alleles and utilizing the precious germplasm resources in the future work.

Virtually, it is expensive and difficult to investigate the whole phenotypic characters and genetic diversity in wide germplasm collections. Thus, a core subset with minimal repetitiveness should be constructed to represent maximal genetic diversity of the entire collections. In this study, a core subset with 27.4% sampling ration was established, which captured the largest Na, Ne, I, and PIC value with a retention rate of 92.8%, 113.6%, 110.6% and 109.7%, respectively. To the best of our knowledge, this is the first report to establish a core collection for the *Lagerstroemia* species.

Na, Ne, I and PIC have been popularly employed as the indexes to evaluate the genetic diversity of the core collection. *Zhao et al.* (*2016*) constructed a core subset which possessed highest Na, Ne, I and PIC with a retention rate of 81.31%, 121.08%, 111.86%, and 113.99%. *Xu et al.* (*2016*) selected the five parameters of Na, Ne, Ho, He and PIC for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

evaluating the genetic information of the core collection. Thus, a highest genetic diversity retention rate with low sampling ratio of the intial population is recommended. Compared with other candidate collections, we choose core collection 8 (20 accessions) to be the best core germplasm of crape myrtle.

Previous studies suggested that the suitable sampling ratio should be based on the characteristics of different germplasm collections. Generally, 10-30% of the sample size should have covered the vast majority of genetic diversity of the intial population (*Wang et al., 2011*). Nine core collections were constructed in this work, which accounted for approximately 8%, 11%, 14%, 16%, 19%, 22%, 25%, 27% and 30% of the total accessions, respectively. Meanwhile, several important accessions in this species were selected as the retained accessions through the above population structure and cluster analysis (e.g. *L. fauriei*, *L. speciosa*, *L. subcostata* and *L. caudata*). Our results showed that the core subset with a 27.4% sample size captured a high allelic retention (92.8%), in which five species were grouped into Cluster 1, whereas Cluster 2 comprised fourteen *Lagerstroemia* cultivars. Based on the dendrogram of 20 core collections, we concluded that *Lagerstroemia* species and *L. indica* progenies possessed the majority of genetic diversity, followed by the hybrids between *L. fauriei* and *L. indica*. Results of t-tests of Na, Ne, I and PIC between the core collection and the entire collection revealed no significant differences ($P < 0.05$), indicating that the core collection developed in the present study effectively represented the whole germplasm collections.

In summary, the core subset identified in this work is very useful for crape myrtle breeding, which will serve as a primary source for efficient sampling of the available germplasm accessions and mining novel genes in genetic association and functional analyses. However, identification of genenotype information only to construct a core subset may not be reliable for capturing the entire genetic alleles of the intial population. The sampling size of entire collections and the limited genetic marker data can also influence the quality of the core subset. Therefore, future studies should be focused on improving this core collection by characterizing the accessions morphologically, incorporating additional individuals and enriching the genetic information in *Lagerstroemia* species.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

REFERENCES

Agarwal, M., Shrivastava, N. & Padh, H. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep.* **27,** 617-631 (2008).

Aggarwal, R. K. *et al*. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor. Appl. Genet.* **114(2),** 359-372 (2007).

Amit, M. *et al*. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* **1,** 543-556; doi: 10.1016/j.celrep.2012.03.013 (2012).

Biswas, M. K. *et al*. Exploiting BAC-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in *Citrus*. *Mol. Biol. Rep.* **39,** 5373-5386 (2012).

Bostein, D., White, R. L., Sholnick, M. & David, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphism. *Am. J. Hum. Genet.* **32,** 314-331 (1980).

Bouck, A. M. Y. & Vision, T. The molecular ecologist's guide to expressed sequence tags. *Mol. Ecol.* **16,** 907-924 (2007).

Brown, A. H. D. Core collections: a practical approach to genetic resources management. *Genome* **31,** 818-824 (1989).

Cai, M. *et al*. Isolation and characterization of microsatellite markers from *Lagerstroemia caudata* (Lythraceae) and cross-amplification in other related species. *Conserv. Genet. Resour.* **2,** 89-91 (2010).

Cheng, Y. L. *et al*. Development and Characterization of EST-SSR Markers in *Taxodium* 'zhongshansa'. *Plant Mol. Biol. Rep.* **33,** 1-11 (2015).

Deng, T. *et al*. De Novo Transcriptome assembly of the Chinese swamp buffalo by RNA sequencing and SSR marker discovery. *PLoS ONE* **11(1),** e0147132 (2016).

Earl, D. A. & Vonholdt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4,** 359-361 (2012).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

Egolf, D. R. ‘Acoma’, ‘Hopi’, ‘Pecos’ and ‘Zuni’ *Lagerstroemia*. *HortScience* **21,** 1250-1252 (1986).

Egolf, D. R. ‘Muskogee’ and ‘Natchez’ *Lagerstroemia*. *HortScience* **16,** 576-577 (1981).

Eujayl, I. *et al*. *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theor. Appl. Genet.* **108,** 414-422 (2004).

Evanno, G. *et al*. Detecting the number of clusters of individuals using the sofware STRUCTURE: a simulation study. *Mol. Ecol.* **14,** 2611–2620 (2005).

Graham, S. A., Hall, J., Sytsma, K. & Shi, S. Phylogenetic analysis of the Lythraceae based on four gene regions and morphology. *Int. J. Plant Sci.* **166,** 995-1017 (2005).

Guidry, R. K. & Einert, A. E. Potted dwarf crape myrtles: a promising new floriculture crop. *Florists Rev.* **157,** 30 (1975).

He, D. *et al*. Genetic diversity of *Lagerstroemia* (Lythraceae) species assessed by simple sequence repeat markers. *Genet. Mol. Res.* **11,** 3522-3533 (2012).

He, D., Liu, Y., Cai, M., Pan, H. T. & Zhang, Q. X. The first genetic linkage map of crape myrtle (*Lagerstroemia*) based on amplification fragment length polymorphisms and simple sequence repeats markers. *Plant Breeding* **133,** 138-144 (2014).

He, G. *et al*. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). *BMC Plant Biol.* **3,** 3 (2003).

Hu, J. *et al*. Microsatellite diversity, population structure, and core collection formation in melon germplasm. *Plant Mol. Biol. Rep.* **33,** 439-447 (2014).

Kim, K. W. *et al*. PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* **23,** 2155-2162 (2007).

Knox, G. W. *University of Florida/IFAS Extension* http://edis.ifas.ufl.edu/mg266 (1992).

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **15,** 1179-1191 (2015).

Liang, W. *et al*. Genetic diversity, population structure and construction of a core collection of apple cultivars from Italian germplasm. *Plant Mol. Biol. Rep.* **33,** 458-473 (2015).

Liu, K. & Muse, S. PowerMarker: new genetic data analysis sofware. Version 3.23

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

Bioinformatics **21(9),** 2128-2129 (2005).

Liu, T. M. *et al*. Large-scale development of expressed sequence tag-derived simple sequence repeat markers by deep transcriptome sequencing in garlic (*Allium sativum* L.). *Mol. Breeding* **35,** 204 (2015).

Lv, J. *et al*. Genetic diversity and population structure of cucumber (*Cucumis sativus* L.). *PLoS ONE* **7,** e46919 (2012).

McKhann, H. I. *et al*. Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. *Plant J.* **38,** 193-202 (2004).

Niu, S. H. *et al*. Transcriptome characterization of *Pinus tabuliformis* and evolution of genes in the *Pinus phylogeny*. *BMC Genomics* **14,** 263; doi: 10.1186/1471-2164-14-263 (2013).

Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W. & Buerkle, C. A. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11,** 180; doi: 10.1186/1471-2164-11-180 (2010).

Peakall, R. & Smouse, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic sofware for teaching and research – an update. *Bioinformatics* **28,** 2537-2539 (2012).

Pooler, M. R. 'Arapaho' and 'Cheyenne' *Lagerstroemia*. *HortScience* **41,** 855-856 (2006).

Pooler, M. R. In *Flower Breeding and Genetics* (ed. Anderson, N. O.). 439-457 (Springer Publishing, 2006).

Pooler, M. R. Molecular genetic diversity among 12 clones of *Lagerstroemia fauriei* revealed by AFLP and RAPD markers. *HortScience* **38,** 256-259 (2003).

Pounders, C., Reed, S. & Pooler, M. Comparison of self- and cross-pollination on pollen tube growth, seed development, and germination in crapemyrtle. *HortScience* **41,** 575-578 (2006).

Pounders, C., Rinehart, T. & Sakhanokho, H. Evaluation of interspecific hybrids between *Lagerstroemia indica* and *L. speciosa*. *HortScience* **42,** 53-68 (2007).

Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945-959 (2000).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

Qiu, L., Yang, C., Tian, B., Yang, J. B. & Liu, A. Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). *BMC Plant Biol.* **10,** 278; doi: 10.1186/1471-2229-10-278 (2010).

Rao, E. S., Kadirvel, P., Symonds, R. C., Geethanjali, S. & Ebert, A. W. Using SSR markers to map genetic diversity and population structure of *Solanum pimpinellifolium* for development of a core collection. *Plant Genet. Resour.* **10,** 38-48 (2012).

Richards, C. M. *et al*. Selection of stratified core sets representing wild apple (*Malus sieversii*). *J. Am. Soc. Hortic. Sci.* **134,** 228-235 (2009).

Rowland, L. J. *et al*. Generation and analysis of blueberry transcriptome sequences from leaves, developing fruit, and flower buds from cold acclimation through deacclimation. *BMC Plant Biol.* **12,** 46; doi: 10.1186/1471-2229-12-46 (2012).

Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18,** 233-234 (2000).

Varshney, R. K., Graner, A. & Sorrells, M. E. Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* **23,** 48-55 (2005).

Wang, J. *et al*. Genomic sequencing using 454 pyrosequencing and development of an SSR primer database for *Lagerstroemia indica* L.. *Plant Omics J.* **8,** 17-23 (2015).

Wang, X. M. *et al*. *Lagerstroemia indica* ‘Xiangyun’, a seedless crape myrtle. *HortScience* **49,** 1590-1592 (2014).

Wang, X. W. *et al*. Development of microsatellite markers from crape myrtle (*Lagerstroemia* L.). *HortScience* **45,** 842-844 (2010).

Wang, Y. Z., Zhang, J. H., Sun, H. Y., Ning, N. & Yang, L. Construction and evaluation of a primary core collection of apricot germplasm in China. *Sci. Hortic.* **128 (3),** 311-319 (2011).

Wang, Z. *et al*. Characterization and development of EST-derived SSR markers in cultivated sweetpotato (*Ipomoea batatas*). *BMC Plant Biol.* **11(1),** 139 (2011).

Wei, W. L. *et al*. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

markers. *BMC genomics* **12,** 451 (2011).

Wu, J., Cai, C. F., Cheng, F. Y., Cui, H. L. & Zhou, H. Characterisation and development of EST-SSR markers in tree peony using transcriptome sequences. *Mol. Breeding* **34,** 1853-1866 (2014).

Xu, C. Q. *et al*. Identifying the genetic diversity, genetic structure and a core collection of *Ziziphus jujuba* Mill. Var. jujuba accessions using microsatellite markers. *Sci. Rep.* **6,** 31503; doi: 10.1038/srep31503 (2016).

Xu, Y. *et al*. Developing a core collection of *Pyropia haitanensis* using simple sequence repeat markers. *Aquaculture* **452,** 351-356 (2016).

Ye, Y. J. *et al*. Identification and validation of SNP markers linked to dwarf traits using SLAF-Seq technology in *Lagerstroemia*. *PLoS ONE* **11,** e0158970 (2016).

Ye, Y. J. *et al*. Screening of molecular markers linked to dwarf trait in crape myrtle by bulked segregant analysis. *Genet. Mol. Res* **14,** 4369-4380 (2015).

Yeh, F. C., Yang, R. C. & Boyle, T. J. B. POPGENE Version 1.32 Microsoft Windows-based freeware for population genetic analysis. *University of Alberta and the Center for International Forestry Research* (1999).

You, Y. N. *et al*. Development and characterisation of EST-SSR markers by transcriptome sequencing in taro (*Colocasia esculenta* (L.) Schoot). *Mol. Breeding* **35,** 134 (2015).

Zeng, S. H. *et al.* Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics* **11,** 94-105 (2010).

Zhai, L. *et al*. Novel and useful genic-SSR markers from de novo transcriptome sequencing of radish (*Raphanus sativus* L.). *Mol. Breeding* **33,** 749-754 (2014).

Zhao, J., Tong, Y. Q., Ge, T. M. & Ge, J. W. Genetic diversity estimation and core collection construction of *Sinojackia huangmeiensis* based on novel microsatellite markers. *Biochem. Syst. Ecol.* **64,** 74-80 (2016).

Zhou, C. P. *et al*. Development of 240 novel EST-SSRs in *Eucalyptus* L'He´rit. *Mol. Breeding* **33,** 221-225 (2014).

ACKNOWLEDGEMENTS

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

This work was financially supported by The 12th Five Years Key Programs for Science and Technology Development of China (no. 2013BAD01B07), the National Science Foundation of China (no. 31470695), the Fundamental Research Funds for the Central Universities (No.BLYJ201612) and Scientific Program from Fujian Province (no. 2014N3012).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

Tables

Table 1 Summary of EST-SSRs identified in crape myrtle transcriptome.

Items	Numbers
Total number of sequences examined	93,161
Total size of examined sequences (bp)	77,921,920
Total number of identified SSRs	8,652
Number of sequences containing more than 1 SSR	1,432
Di-nucleotide	4,942
Tri-nucleotide	3,456
Tetra-nucleotide	212
Penta-nucleotide	29
Hexa-nucleotide	13

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

542 **Table 2 Probability of identity analyzed from 73 accessions using GenAlex 6.5 on 30**
543 **EST-SSR markers.**

SSR marker	Number of identical pairs of genotypes	Probability of identity	Cumulative probability of identity
YYJ-283	1771	4.9E-02	4.9E-02
YYJ-281	1361	5.7E-02	2.8E-03
YYJ-706	1114	5.9E-02	1.7E-04
YYJ-682	546	6.3E-02	1.1E-05
YYJ-693	381	6.4E-02	7.0E-07
YYJ-643	364	7.1E-02	5.0E-08
YYJ-327	353	7.6E-02	3.8E-09
YYJ-199	273	7.6E-02	2.9E-10
YYJ-656	230	9.8E-02	2.8E-11
YYJ-187	140	1.0E-01	2.8E-12
YYJ-68	69	1.3E-01	3.6E-13
YYJ-40	22	1.4E-01	5.0E-14
YYJ-413	16	1.4E-01	7.0E-15
YYJ-579	16	1.4E-01	9.8E-16
YYJ-201	13	1.5E-01	1.5E-16
YYJ-646	11	1.5E-01	2.3E-17
YYJ-297	2	1.8E-01	4.1E-18
YYJ-166	2	2.0E-01	8.2E-19
YYJ-365	2	2.1E-01	1.7E-19
YYJ-180	2	2.2E-01	3.7E-20
YYJ-148	2	2.2E-01	8.1E-21
YYJ-331	2	2.6E-01	2.1E-21
YYJ-228	1	3.1E-01	6.5E-22
YYJ-92	1	3.2E-01	2.1E-22
YYJ-81	1	3.2E-01	6.7E-23

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

YYJ-337	0	3.6E-01	2.4E-23
YYJ-356	0	3.6E-01	8.6E-24
YYJ-118	0	4.3E-01	3.7E-24
YYJ-129	0	5.8E-01	2.1E-24
YYJ-695	0	6.2E-01	1.3E-24

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

544 **Table 3 Polymorphic information of 30 EST-SSRs in 73 accessions.**

Locus	Na	Ne	I	Ho	He	PIC
YYJ-283	9	6.054	1.881	0.903	0.835	0.813
YYJ-281	8	5.389	1.849	0.719	0.814	0.792
YYJ-706	8	5.508	1.795	0.612	0.818	0.793
YYJ-682	8	5.274	1.771	0.500	0.810	0.783
YYJ-693	13	5.007	1.930	0.688	0.800	0.776
YYJ-643	7	4.984	1.671	0.636	0.799	0.768
YYJ-327	10	4.735	1.765	0.754	0.789	0.758
YYJ-199	11	4.580	1.768	0.544	0.782	0.753
YYJ-656	8	4.047	1.608	0.958	0.753	0.716
YYJ-187	9	3.649	1.647	0.596	0.726	0.701
YYJ-68	7	3.330	1.422	0.607	0.700	0.655
YYJ-40	9	3.313	1.485	0.640	0.698	0.652
YYJ-413	5	3.402	1.317	0.657	0.706	0.650
YYJ-579	5	3.410	1.326	0.672	0.707	0.651
YYJ-201	10	2.921	1.462	0.685	0.658	0.623
YYJ-646	5	3.289	1.277	0.375	0.696	0.636
YYJ-297	7	2.816	1.277	0.314	0.645	0.596
YYJ-166	6	2.393	1.226	0.589	0.582	0.552
YYJ-365	5	2.636	1.116	0.526	0.621	0.559
YYJ-180	7	2.617	1.176	0.619	0.623	0.545
YYJ-148	12	2.240	1.366	0.538	0.554	0.537
YYJ-331	9	2.401	1.100	0.415	0.583	0.500
YYJ-228	7	1.880	1.031	0.318	0.468	0.443
YYJ-92	4	1.866	0.881	0.414	0.464	0.427
YYJ-81	8	1.840	1.000	0.228	0.456	0.429
YYJ-337	4	1.781	0.803	0.415	0.438	0.397
YYJ-356	6	1.730	0.866	0.429	0.422	0.393
YYJ-118	6	1.559	0.775	0.290	0.359	0.338
YYJ-129	5	1.328	0.558	0.212	0.247	0.237
YYJ-695	5	1.290	0.476	0.221	0.225	0.210

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

545 **Table 4 Transferability of SSR loci of *Lagerstroemia indica* to related species.** *Note:* +,
546 successful amplification. –, no amplification.

SSR marker	*L. indica* 'Pocomoke'	*L. fauriei*	*L. caudata*	*L. speciosa*	*L. limii*	*L. excelsa*	*L. subcostata*
YYJ-283	+	+	+	–	+	+	+
YYJ-281	+	+	+	–	–	+	+
YYJ-706	+	+	+	+	+	+	+
YYJ-682	+	+	+	–	–	–	+
YYJ-693	+	+	+	–	+	+	+
YYJ-643	+	+	+	–	+	+	+
YYJ-327	+	+	+	–	+	+	+
YYJ-199	+	+	+	+	+	+	+
YYJ-656	+	+	+	–	+	+	+
YYJ-187	+	+	+	+	+	+	+
YYJ-68	+	+	+	–	–	+	–
YYJ-40	+	+	+	–	+	+	–
YYJ-413	+	+	–	–	+	+	+
YYJ-579	+	+	+	–	+	+	+
YYJ-201	+	+	+	+	+	+	+
YYJ-646	–	+	+	–	+	+	+
YYJ-297	+	+	–	–	+	+	+
YYJ-166	+	+	+	+	+	+	+
YYJ-365	+	–	+	–	+	+	+
YYJ-180	+	+	+	–	+	+	+
YYJ-148	+	+	+	+	+	+	+
YYJ-331	+	+	+	–	+	+	+
YYJ-228	+	+	–	–	+	–	+
YYJ-92	+	+	+	+	–	+	+
YYJ-81	+	+	+	–	+	+	+
YYJ-337	+	+	+	–	–	+	+
YYJ-356	+	+	+	+	+	+	+
YYJ-118	–	+	+	–	+	+	+
YYJ-129	+	–	+	–	+	+	+
YYJ-695	+	+	–	–	+	+	+

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

Table 5 The genetic analysis of three population for the 73 accessions. N = No. of individuals; Nt = No. of different alleles (total) in each population; Na = No. of different alleles per locus; Ne = No. of effective alleles; Np = No. of private alleles; Ho = Observed heterozygosity; He = Expected heterozygosity; I = Shannon information index.

Population	N	Nt	Na	Ne	Np	Ho	He	I
Pop1	24	149	4.967	2.864	15	0.596	0.586	1.148
Pop2	17	168	5.600	3.746	45	0.461	0.693	1.423
Pop3	32	152	5.100	2.734	20	0.503	0.558	1.002
Total	73	223	7.433	3.242	80	0.536	0.626	1.321

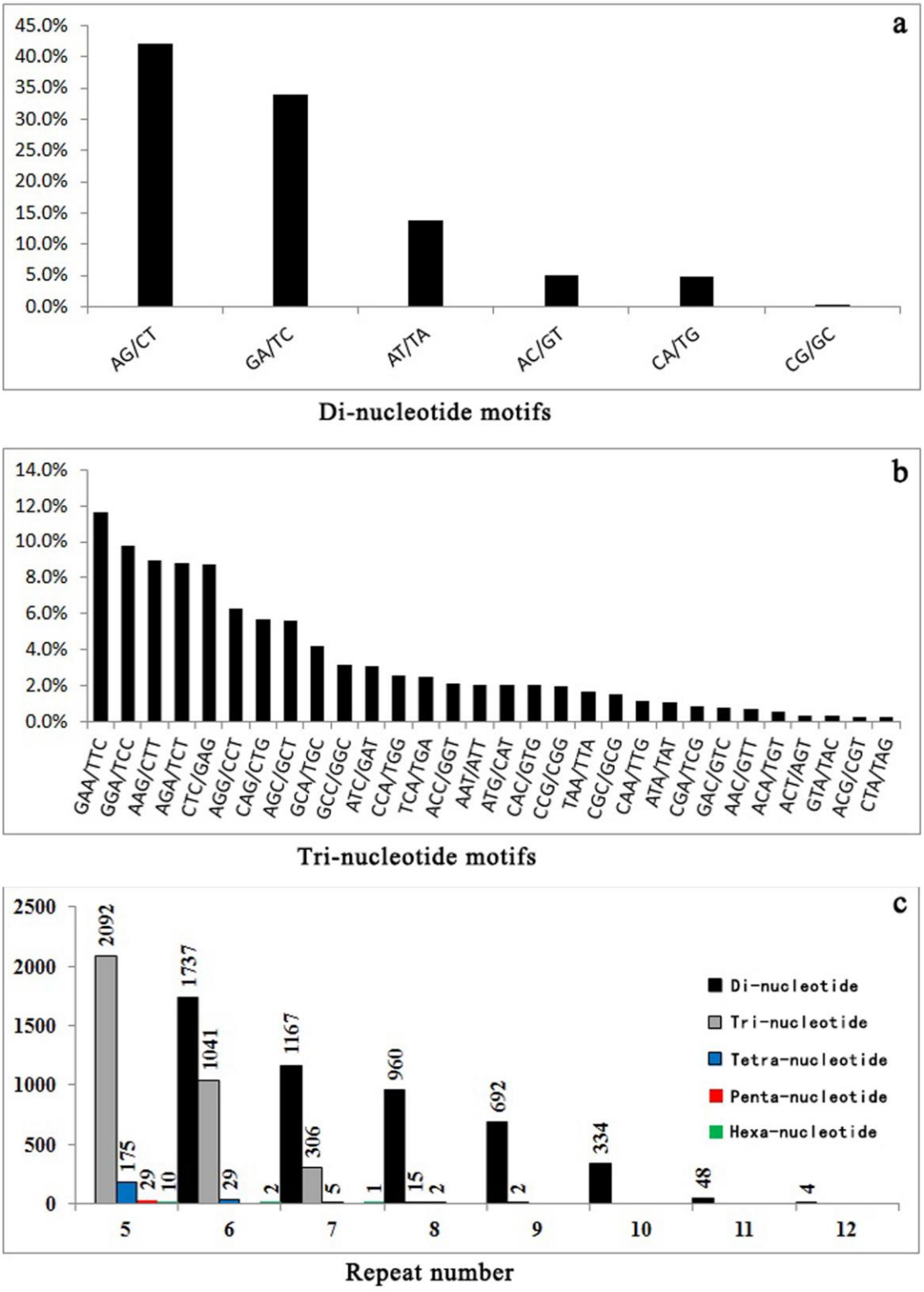
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

551 **Table 6 Comparison of the genetic diversity statistics among different sampling**
552 **groups of crape myrtle.**

Population	Number of individuals	Na	Ne	I	Ho	He	PIC
Entire collection	73	7.433	3.242	1.321	0.536	0.626	0.589
Core collection 1	6	4.500	3.555	1.310	0.419	0.676	0.631
Core collection 2	8	5.267	3.787	1.409	0.451	0.693	0.648
Core collection 3	10	5.767	3.870	1.448	0.467	0.695	0.659
Core collection 4	12	6.100	3.866	1.469	0.491	0.696	0.662
Core collection 5	14	6.100	3.866	1.469	0.491	0.696	0.648
Core collection 6	16	6.400	3.591	1.426	0.486	0.672	0.638
Core collection 7	18	6.600	3.608	1.439	0.495	0.674	0.641
Core collection 8	20	6.900	3.682	1.461	0.521	0.678	0.646
Core collection 9	22	6.967	3.566	1.436	0.516	0.667	0.635

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

553 **Figure captions**



554

555 **Figure 1 Characterization of EST-SSRs in *Lagerstroemia* transcriptome.** (a) Frequency

556 distribution of di-nucleotide SSRs based on motif type. (b) Frequency distribution of

557 tri-nucleotide SSRs based on motif type. (c) Number of different repeat units.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

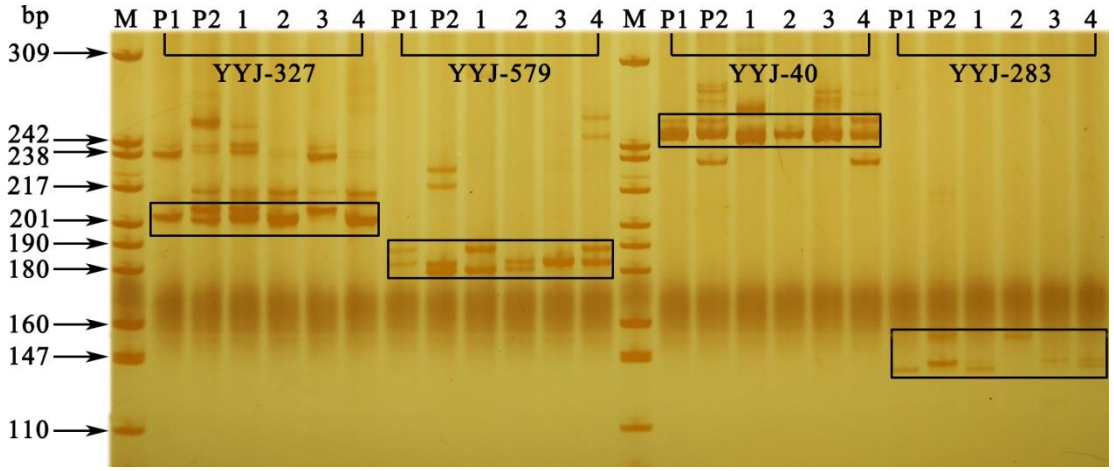


Figure 2 The polyacrylamide gel electrophoresis of 4 typical polymorphic markers among 6 samples in the F1 population. *Note:* the bands in the black rectangles represent the expected fragments. M, pBR322 DNA marker (TianGen Biotech, Beijing, China); P1, *L. indica* 'Pocomoke'; P2, *L. fauriei*; 1-4, four individuals selected from the F1 population.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017

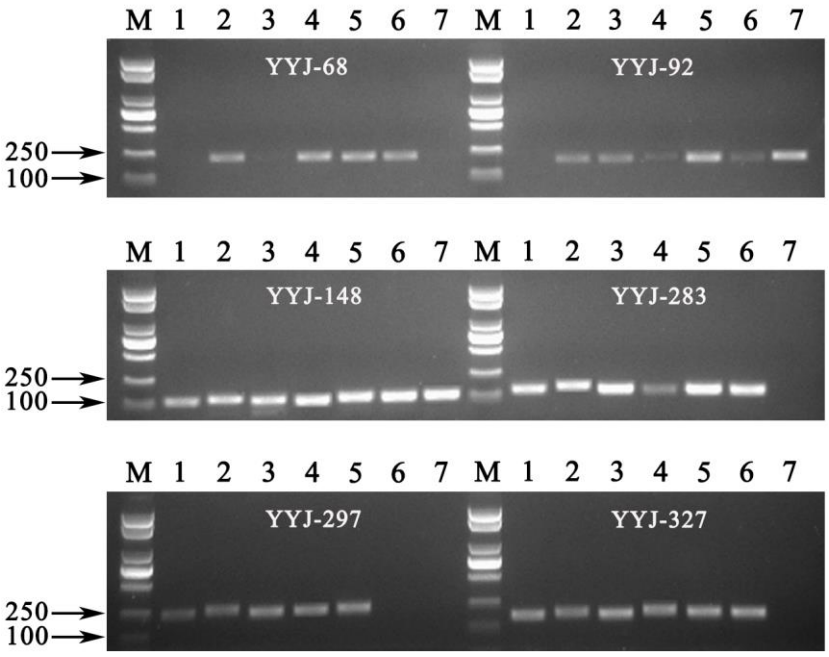
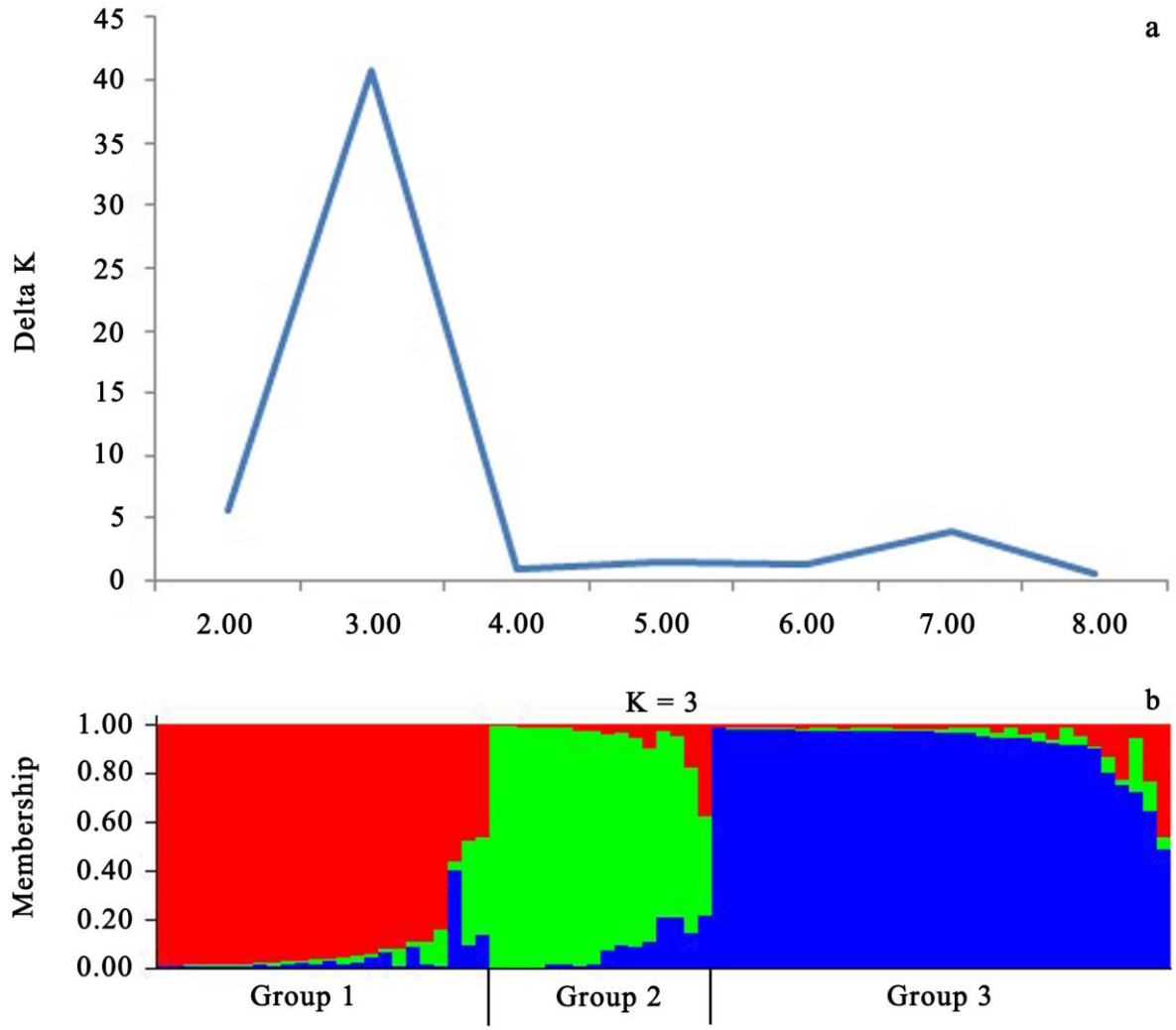


Figure 3 The transferability of selected EST-SSR markers across 7 crape myrtle species by 1% agarose gel electrophoresis. M, DL2000 DNA ladder marker (TianGen Biotech, Beijing, China); 1-7, each represents *L. limii*, *L. excelsa*, *L. subcostata*, *L. indica* 'Pocomoke', *L. fauriei*, *L. caudata* and *L. speciosa*.

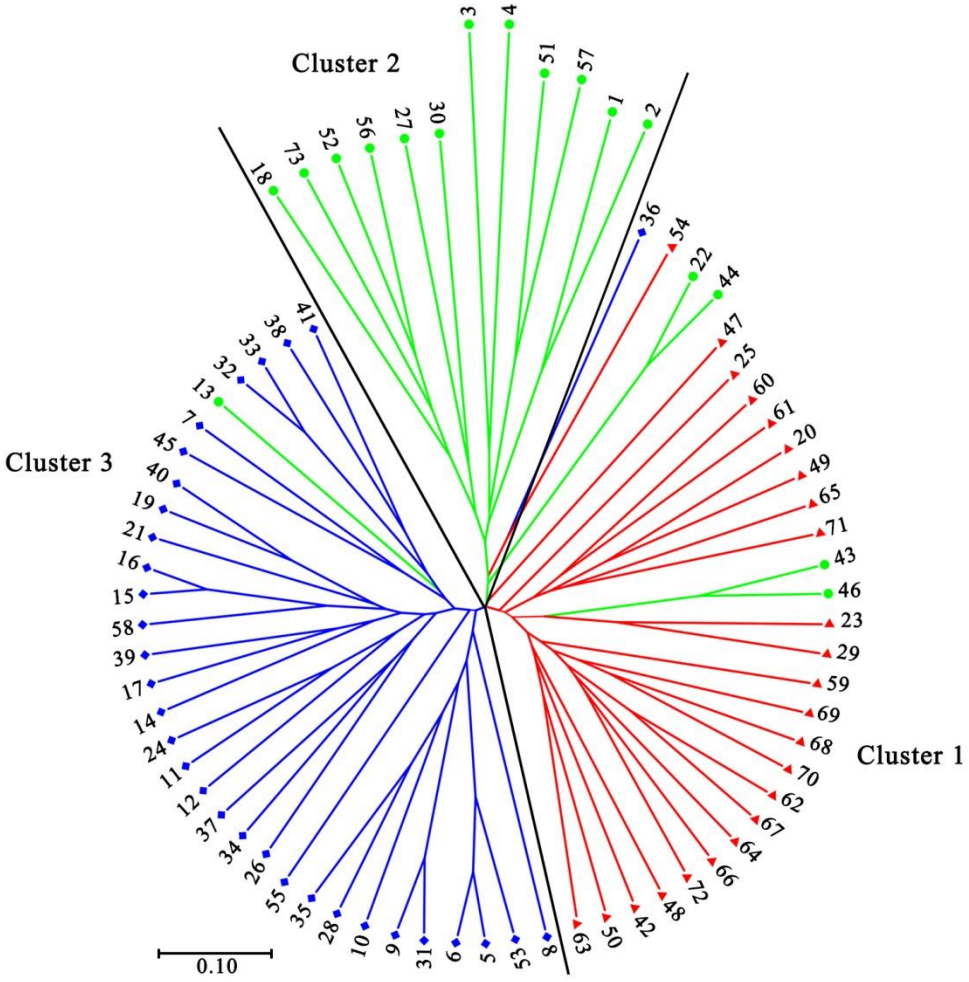
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017



568

569 **Figure 4 Model-based structure analysis of crape myrtle accessions.** (a) Determination
570 of the real number of groups based on delta K. (b) Population structure for three clusters,
571 following the result of the delta K estimation.

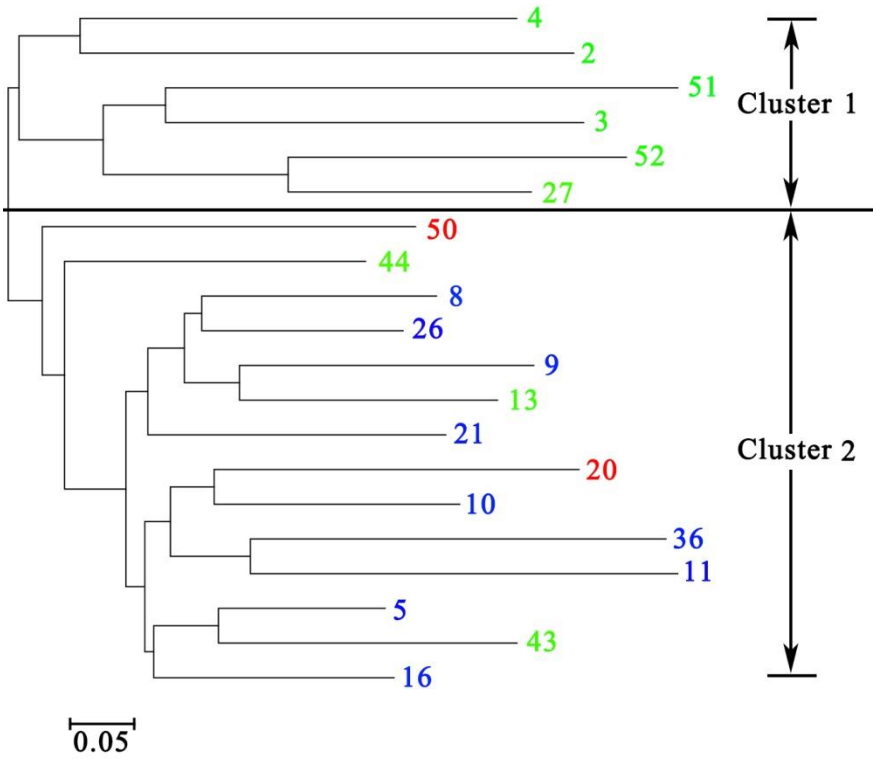
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017



572

573 **Figure 5 UPGMA dendrogram of the 73 accessions.** *Note:* UPGMA dendrogram was
574 performed using the Powermarker v 3.25 based on the data of 30 EST-SSR markers. The
575 cluster results corresponded to those of the STRUCTURE groups with the same color. The
576 information of the code represented in the figure can be seen in the Supplemental Table S1.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017



577

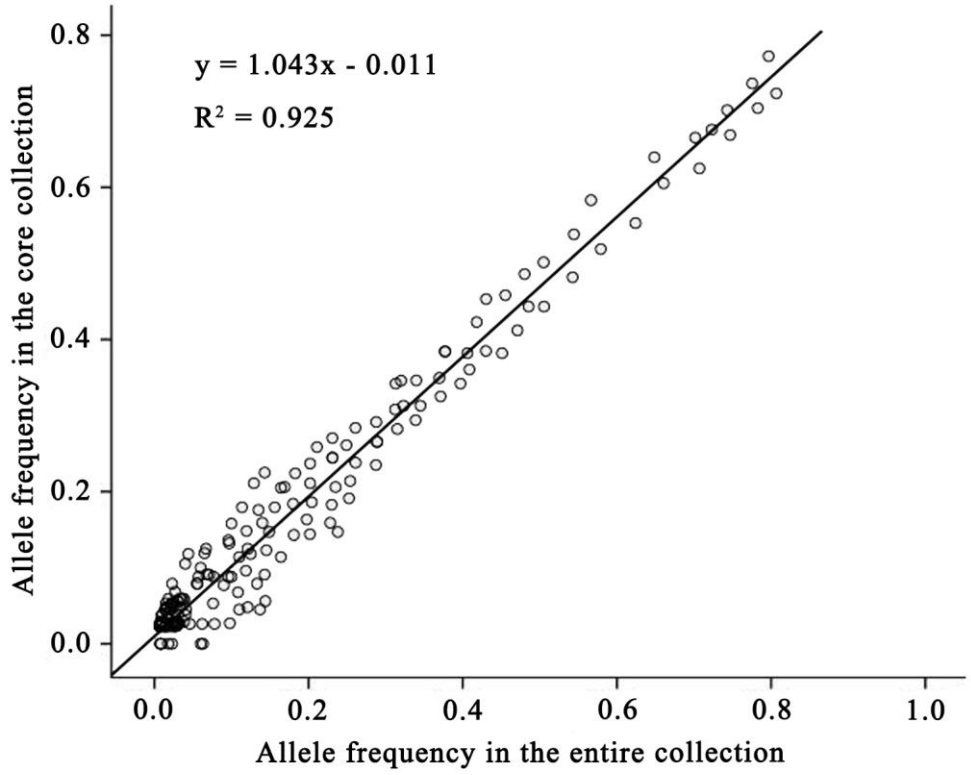
578 **Figure 6 Cluster analysis of 20 core individuals based on the data of 30 EST-SSR**

579 **markers.** The same color was used for each sample corresponding to those of the

580 STRUCTURE groups. The information of the code represented in the figure can be seen in

581 the Supplemental Table S1.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017



582

583 **Figure 7 Scatter Plot of allele frequency distribution between the core subset and**

584 **entire collection using SPSS v 18.0.** *Note:* The dots in the figure represent 207 alleles

585 shared by the 20 core individuals and 73 intial accessions (223 alleles).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3141v1 | CC BY 4.0 Open Access | rec: 11 Aug 2017, publ: 11 Aug 2017