Analysis of genetic diversity, population structure and its association with quality traits in germplasm of Blumea balsamifera (L.) DC. based on EST-SSR markers

Xin Li
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Mei Huang
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Zhenxia Chen
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Dan Wang
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Kai Wang
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Xuan Hu
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Xiaolu Chen
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Lingliang Guan
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs
Hongrui Zhang
Henan Agricultural University
Yuxin Pang
Guizhou University of Traditional Chinese Medicine
Fulai Yu (fulai.yu@163.com)
Chinese Academy of Tropical Agricultural Sciences, Ministry of Agriculture and Rural Affairs

Research Article

Keywords: Blumea balsamifera, quality traits, EST-SSR, genetic diversity, association analysis

Posted Date: March 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2621060/v1

Additional Declarations: No competing interests reported.

Abstract

Background

In order to establish the link between quality traits and genotypes of *Blumea balsamifera* and specific allele variations, 51 *B. balsamifera* germplasm resources were used to evaluate the quality traits, EST-SSR markers were used to analyze the genetic diversity, population structure and the correlation between main quality traits and EST-SSR molecular markers.

Results

(1) There were abundant variations in the main quality traits of the tested materials. The highest coefficient of variation was Eriodictyol (85.5%), followed by carotenoid 3,3',5,7-tetrahydroxy-4'-methoxyflavanone and Sakuranetin, 11 excellent *B. balsamifera* germplasms were selected by principal component analysis; (2) Genetic diversity analysis showed that a total of 102 alleles were amplified from 22 pairs of primers, of which the effective alleles accounted for 53.52%, and the average polymorphism information content was 0.488, 9 pairs of primers had high polymorphism (PIC > 0.5), 11 pairs of primers had moderate polymorphism (0.25 < PIC < 0.5), and the proposed primers had strong effectiveness and good polymorphism; The average gene flow *Nm* was 0.203, which was far less than 1, indicating that there was almost no inbreeding between germplasms; The average *Nei* diversity index and Shannon information index were 0.542 and 1.023, which showed that the population had a high level of genetic diversity; (3) UPGMA cluster analysis and population structure analysis divided the 51 germplasms into 4 groups, and the germplasms from the same source were often gathered in a group. UPGMA cluster analysis showed that geographic distance affected the genetic relationship of germplasm to some extent; PCA analysis indicated that the genetic background of germplasm with similar geographical distance was similar, and the genetic relationship was closer; The analysis of population structure showed that the geographical origin of germplasm was closely related to the genetic relationship of germplasm, and again confirmed the accuracy of UPGMA cluster analysis; (4) The result of linkage disequilibrium(LD) analysis showed that the markers with D '>0.5 accounted for more than 50%, and the recombination probability between germplasm genes was low, indicating that the level of genetic diversity of the population was high, suggesting that the experimental materials were suitable for association analysis, and (5) The result of correlation analysis between quality traits and EST-SSR markers showed that 23 markers significantly correlated with 6 quality traits were detected, and the variance interpretation rate was 19.33% – 57.86%. Among them, the character Blumeatin had the best correlation with EST-SSR loci, and showed a very significant correlation with Bbf377 marker primer.

Conclusion

The results could lay a theoretical foundation for the selection and genetic improvement of excellent germplasm of *B. balsamifera* in the future.

Background

Blumea balsamifera (L.) DC belongs to the Asteraceae family, genus Blumea, which is widely distributed in China, Malaysia, Philippines, Thailand and Vietnam. It is commonly used in tropical and subtropical areas of Asia to treat eczema, dermatitis, rheumatism and skin lesions, etc. [1]. It is the only plant source of Aipian, which is recorded in the Pharmacopoeia of the People's Republic of China (2020 version). In recent years, due to its great medicinal value, its chemical composition has been widely studied, mainly focusing on volatile oil and flavonoids [2]. As the only plant source of Aipian, breeders pay increasing attention to its quality improvement. Germplasm resources are the premise and material basis of breeding, it is of great significance to study the genetic diversity of germplasm resources for efficient utilization and innovation of germplasm [3]. The research team summarized the germplasm resources of *B. balsamifera* in the early stage and found that due to the influence of industrial development, disordered harvesting and other factors in recent years, the germplasm resources were seriously reduced, the quality was uneven, and the output and quality of medicinal materials were difficult to control [4]. In

addition, the less development of genome information of *B. balsamifera* restricts the effective utilization of *B. balsamifera* germplasm resources [5, 6], so expanding the selection range of *B. balsamifera* and improving the scientific of breeding has become an urgent problem to be solved. In view of this, judging the genetic background and genetic relationship of existing germplasm at the molecular level can effectively reduce the germplasm combination with close genetic relationship and improve the feasibility of breeding.

With the continuous development of molecular marker technology, this technology has been widely used in the analysis of plant genetic diversity, fingerprints and quantitative trait loci (QTL), etc. [7–9]. Among them, simple sequence repeat (SSR) has the characteristics of codominance, stability, repeatability and polymorphism, and are one of the most widely used markers in many molecular markers. SSR is mainly used in the fields of genetic diversity and trait association analysis. Pang [5] used RAPD technology to study the clonal diversity of four wild populations of *B. balsamifera*, and found that the level of clonal diversity of *B. balsamifera* population was high. Zhang [6] used the commonly used SRAP and AFLP molecular markers to compare and analyze the genetic diversity of 35 *B. balsamifera* resources, and found that AFLP had diversity loci, it has many advantages, such as high molecular marker index, and was suitable for the evaluation of genetic diversity of *B. balsamifera*. The above researches were the previous research of the research group on the molecular level of *B. balsamifera*, however, there was no analysis on SSR marker analysis of *B. balsamifera*.

Plant quality traits are often controlled by micro effect polygenes, the genetic basis is complex and vulnerable to environmental impact. Traditional breeding is inefficient. Association analysis, also called association mapping, which is a method based on linkage disequilibrium and using natural populations as materials to detect the significant association frequency between markers in linkage disequilibrium and target traits in plants [10–12]. Relevance licensing only requires natural populations or germplasm resources, it has been widely used in wheat, soybean, corn and other crops [13–15], and has achieved good results. Although association analysis has been applied to many crops, it has not been reported that SSR markers can be used to carry out association analysis of active components in the germplasm resources of Chinese herbal medicine *B. balsamifera*.

Based on this, in this study, 51 *B. balsamifera* germplasm resources were used as materials to analyze their genetic diversity and population structure using EST-SSR molecular markers. Based on the analysis of population genetic structure, linkage disequilibrium analysis and quality trait association analysis were carried out., and EST-SSR markers related to 6 quality traits of *B. balsamifera* were determined, which provided a theoretical basis for molecular level genetic research and marker assisted breeding of *B. balsamifera*.

Results

Quality traits analysis of the B. balsamifera germplasm resource

Variability Analysis Of The Main Quality Traits

The contents of *l*-Borneol, 3,3',5,7-tetrahydroxy-4'-methoxyflavanone, Eriodictyol, 3,3',5-Trihydroxy-4',7-dimethoxyflavanone, Blumeatin and Sakuranetin in 51 samples were tested (Table 1). The highest coefficient of variation was Eriodictyol (85.5%), followed by 3,3',5,7-tetrahydroxy-4'-methoxyflavanone (68.2%) and 3,3',5-Trihydroxy-4',7-dimethoxyflavanone (53.1%). The highest average content was observed for *l*-Borneol (5.833 $mg \cdot g^{-1}$), followed by 3,3',5,7-tetrahydroxy-4'-methoxyflavanone (4.824 $mg \cdot g^{-1}$). The results showed that the variation in *B. balsamifera* was rich, and the six main quality traits of the samples were different among the different genotypes, the greater the genetic variation, the richer the genetic basis, the tested germplasm had great genetic improvement potential.

Table 1
Statistical analysis of six main quality traits for *B. balsamifera*

Quality traits	Amplitude of variation	Mean	SD	CV/%
l-Borneol / (mg·g^{-1})	1.191–14.463	5.833	3.576	0.613
3,3',5,7-tetrahydroxy-4'-methoxyflavanone / (mg·g^{-1})	1.098–16.086	4.824	3.292	0.682
Eriodictyol / (mg·g^{-1})	0.118–3.724	1.116	0.954	0.855
3,3',5-Trihydroxy-4',7-dimethoxyflavanone / (mg·g^{-1})	0.500–3.540	1.527	0.810	0.531
Blumeatin / (mg·g^{-1})	0.363–3.644	1.505	0.846	0.562
Sakuranetin / (mg·g^{-1})	0.010–0.225	0.055	0.036	0.650

Correlation Analysis Of The Main Quality Traits

It can be found in Table 2 that most of the quality traits were highly significantly correlated with each other, and only Sakuranetin was not significantly correlated with *l*-Borneol, 3,3',5,7-tetrahydroxy-4'-methoxyflavanone and Eriodictyol. In terms of correlation analysis of other quality traits, there was a significant positive correlation between *l*-Borneol, 3,3’,5,7-tetrahydroxy-4’-methoxyflavanone, Eriodictyol, 3,3’,5-Trihydroxy-4’,7-dimethoxyflavanone and Blumeatin, and Sakuranetin was significantly positively correlated with 3,3’,5-Trihydroxy-4’,7-dimethoxyflavanone and Blumeatin. The correlation between 3,3’,5,7-tetrahydroxy-4’-methoxyflavanone and *l*-Borneol was the highest (0.873), followed by Eriodictyol (0.822). The above studies showed that there were obvious correlations among most quality traits, and they were coordinated and interacted with each other, which was of great reference value for the subsequent association analysis of quality traits and EST-SSR markers.

Table 2

Correlation analysis of six main quality traits for *B. balsamifera*

Quality traits	*l*-Borneol	3,3’,5,7-tetrahydroxy-4’-methoxyflavanone	Eriodictyol	3,3’,5-Trihydroxy-4’,7-dimethoxyflavanone	Blumeatin	Sakuranetin
l-Borneol	1	0.873**	0.735**	0.780**	0.640**	0.269
3,3’,5,7-tetrahydroxy-4’-methoxyflavanone		1	0.822**	0.638**	0.454**	0.098
Eriodictyol			1	0.424**	0.397**	0.178
3,3’,5-Trihydroxy-4’,7-dimethoxyflavanone				1	0.821**	0.414**
Blumeatin					1	0.668**
Sakuranetin						1

*Note: **Significant correlation at the 0.01 level (bilateral).*

Principal Component Analysis Of Tested Materials

Principal component analysis of tested materials was carried out by SPSS 24.0 software, and the results of the KMO and Bartlett spherical tests showed that the value of KMO was 0.733 (KMO > 0.6), and the significant value of the Bartlett spherical test was 0.000 ($P < 0.05$), which indicated that the data was suitable for factor analysis. The principal component analysis

results showed that the cumulative contribution rate of the first two principal components was 85.474%, which basically contained most of the information of the original indicators, and could be used as comprehensive indicators for evaluating quality traits (Table 3). The characteristic vectors of *l*-Borneol, 3,3',5,7-tetrahydroxy-4'-methoxyflavanone, 3,3',5-Trihydroxy-4',7-dimethoxyflavanone and Blumeatin were relatively large in the first principal component, indicating that they were the main influencing factors of the first principal component. Sakuranetin was the main influencing factor of the second principal component. Taking the variance contribution rate corresponding to the top two principal components in the principal component analysis as the weight, the product of each principal component score and the corresponding weight was summed, and the cumulative contribution rate of the two principal components was divided. The following comprehensive score formula was obtained:

F(Total) = 63.93/85.474*F1 + 21.544/85.474*F2.

According to the formula, the comprehensive scores of 51 samples main quality traits were calculated and ranked (Table 4). The higher comprehensive score was, the better the comprehensive quality was. The F value of the top 11 germplasms was greater than 0.5, and the F values of 6 germplasms from Danzhou, including No. 17, 18, 8, and 9, were higher than 1. Germplasms No. 6, 10, 11, and 31 from Luodian, Guizhou, were higher than 0.5, which indicated that the quality traits of the germplasm from Luodian and Danzhou were better. According to the above study, germplasm 17, 9, 18, 8 and 11 showed excellent comprehensive performance in quality according to the comprehensive score, which could be used as the parents of excellent germplasm selection and the research materials of excellent genes in genetic improvement in the future.

Table 3

Principal component analysis of *B. balsamifera* germplasm resources.

Quality traits	1	2
l-Borneol	0.930	-0.216
3,3',5,7-tetrahydroxy-4'-methoxyflavanone	0.851	-0.461
Eriodictyol	0.763	-0.442
3,3',5-Trihydroxy-4',7-dimethoxyflavanone	0.870	0.196
Blumeatin	0.819	0.482
Sakuranetin	0.486	0.754
Eigenvalues	4.719	0.313
Contribution rate	63.930	21.544
Accumulative contribution rate	63.930	85.474

Table 4

Principal component scores of *B. balsamifera* germplasms.

Code	Scores			Ranking	Code	Scores			Ranking
	F1	F2	F(Total)			F1	F2	F(Total)	
17	3.06	0.39	1.94	1	41	0.01	-0.42	-0.17	27
9	2.76	0.26	1.71	2	7	-0.82	0.67	-0.2	28
18	2.47	-0.12	1.38	3	48	-0.74	0.53	-0.21	29
8	1.15	1.48	1.29	4	5	-0.06	-0.50	-0.24	30
11	2.05	-0.21	1.1	5	3	0.34	-1.26	-0.33	31
34	-0.43	2.56	0.83	6	19	-0.74	0.22	-0.34	32
6	0.43	1.35	0.82	7	15	-1.06	0.47	-0.42	33
13	0.98	0.42	0.74	8	23	-0.31	-0.64	-0.45	34
31	0.89	0.27	0.63	9	1	-0.33	-0.69	-0.48	35
22	-1.68	3.79	0.61	10	50	-0.58	-0.37	-0.5	36
10	1.00	0.01	0.59	11	47	-0.93	-0.10	-0.58	37
30	0.95	-0.17	0.48	12	49	-0.52	-0.67	-0.58	38
29	-0.01	1.03	0.43	13	32	-0.33	-1.06	-0.64	39
16	0.10	0.82	0.4	14	35	-0.82	-0.47	-0.67	40
14	-0.78	1.96	0.37	15	39	-0.50	-1.01	-0.71	41
26	0.62	0.01	0.36	16	43	-0.62	-0.88	-0.73	42
25	0.75	-0.25	0.33	17	46	-0.57	-0.96	-0.73	43
45	0.74	-0.36	0.28	18	44	-0.48	-1.10	-0.74	44
33	-0.09	0.64	0.22	19	36	-0.80	-0.77	-0.79	45
4	0.17	0.27	0.21	20	20	-0.96	-0.59	-0.81	46
2	-0.07	0.27	0.07	21	24	-1.07	-0.49	-0.83	47
51	-0.22	0.44	0.06	22	42	-0.59	-1.15	-0.83	48
27	0.41	-0.48	0.04	23	37	-0.71	-1.02	-0.84	49
40	-0.79	1.14	0.02	24	12	-0.65	-1.13	-0.85	50
28	0.29	-0.66	-0.11	25	38	-0.88	-1.17	-1	51
21	0.00	-0.28	-0.12	26					

Genetic diversity analysis based on EST-SSR markers

PCR amplification was carried out on the selected 22 pairs of primers, and the amplified products were analyzed by a DNA analyzer to accurately obtain the fragment sizes and corresponding electropherograms of alleles at each point of different germplasms, some of the detection results are shown in Fig. 1. The results of primer polymorphisms were listed in Table 5. A total of 102 allele number (Na) were detected with a range of 3–8, the range of effective allele number (Ne) was 1.329–4.682,

with an average of 2.482. The variation range of Shannon index (I) was 0.447–1.793, and the average value was 1.023. The polymorphism information content (PIC) ranged from 0.222–0.762, with an average of 0.488. There was moderate polymorphism (0.25 < PIC < 0.5), among which 9 pairs of primers had high polymorphism (PIC > 0.5) and 11 pairs of primers had moderate polymorphism (0.25 < PIC < 0.5). In this study, primers with high PIC value accounted for 41%, which indicated that these SSR loci can explain genotypic differences at the molecular level and have rich genetic differences. In addition, the average heterozygosity observed (Ho) was 0.494, and the average expected heterozygosity (He) was 0.548, which was slightly higher than the observed heterozygosity without excessive heterozygosity, the average value of gene flow among accessions was 0.203, and the average *Nm* value was less than 1, the above results indicated that gene exchange among the accessions was not frequent. To sum up, the primers screened in this study had high polymorphism, which can effectively reflect the genetic diversity of germplasm. In addition, the germplasm collected in the experiment has large genetic differences and rich genetic diversity, which can be used for association analysis of quality traits.

Table 5

Polymorphism of EST-SSR markers

Primer	Na	Ne	Ho	He	Nei	I	PIC	*Nm**
Bbf001	6	1.555	0.333	0.360	0.357	0.784	0.341	0.219
Bbf020	4	2.257	0.373	0.562	0.557	0.979	0.500	0.126
Bbf041	3	2.164	0.500	0.543	0.538	0.848	0.437	0.197
Bbf065	3	1.884	0.569	0.474	0.469	0.700	0.368	0.384
Bbf068	6	3.814	0.863	0.745	0.738	1.530	0.699	0.352
Bbf070	7	3.197	0.667	0.694	0.687	1.411	0.644	0.236
Bbf103	4	1.330	0.196	0.251	0.248	0.536	0.237	0.163
Bbf106	7	3.752	0.551	0.741	0.733	1.538	0.695	0.135
Bbf137	4	2.116	0.588	0.533	0.527	0.823	0.416	0.315
Bbf140	3	1.329	0.245	0.250	0.248	0.447	0.222	0.157
Bbf146	5	2.135	0.392	0.537	0.532	1.066	0.499	0.146
Bbf156	3	1.743	0.490	0.430	0.426	0.711	0.363	0.338
Bbf161	4	3.113	0.765	0.686	0.679	1.241	0.623	0.323
Bbf193	3	2.112	0.353	0.532	0.527	0.858	0.443	0.126
Bbf201	8	4.682	0.765	0.794	0.786	1.793	0.762	0.237
Bbf219	4	2.553	0.438	0.615	0.608	1.049	0.530	0.115
Bbf250	7	4.439	0.686	0.782	0.775	1.618	0.740	0.199
Bbf271	3	1.783	0.216	0.444	0.439	0.711	0.365	0.081
Bbf283	4	1.559	0.392	0.362	0.359	0.689	0.326	0.302
Bbf289	4	1.968	0.333	0.497	0.492	0.947	0.456	0.128
Bbf377	6	2.841	0.628	0.654	0.648	1.303	0.605	0.235
Bbf384	4	2.271	0.529	0.565	0.560	0.933	0.465	0.224
Total	102	54.595	—	—	—	—	—	—
Mean	4.636	2.482	0.494	0.548	0.542	1.023	0.488	0.203
SD	1.59	0.969	0.186	0.159	0.157	0.372	0.154	—

UPGMA Analysis And Principal Component Analysis Based On EST-SSR Markers

The UPGMA method was applied to cluster analysis of 51 accessions (Fig. 2), the genetic similarity coefficient ranged from 0.5 to 1, at a genetic similarity coefficient of 0.572, the collected germplasms could be divided into four groups. Group I included three germplasms from Guangdong and two germplasms from Danzhou. Group II included one germplasm from Danzhou, thirteen germplasms from Luodian, six germplasms from Tianlin, three germplasms from Xilin and one germplasm from Zhenfeng. Among them, a total of 11 accessions from Luodian, including No. 3–4, No. 10–13 and No. 26–31, were

clustered in one branch, and No. 50 and No. 51, which were from Wanning and Wuzhishan, Hainan, were clustered in one branch. This indicated that the genetic relationship of the germplasm had a certain correlation with geographical distribution. Group III contained two germplasms from Guangzhou, one germplasm from Nanning, and one germplasm from Wangmo. Group IV included all germplasms from Qiongzhong and Tunchang, Hainan, which again showed that geographical distance affects the genetic relationship between accessions.

The genetic similarity coefficient matrix obtained was analyzed by principal component analysis (PCA) using Ntsys software. Two dimensional and three-dimensional scattering graphs were obtained (Fig. 3.). From the distribution of germplasm on the map, it can be found that the geographically distant germplasm was relatively far away on the map, and most of the geographically close germplasm was also relatively close on the map. The distribution characteristics of UPGMA cluster germplasm were also roughly reflected in PCA. For example, the germplasm of No.47 and No.48 from Guangzhou, Guangdong, were still gathered together, and were far away from the germplasm of other provinces, germplasms from Guangxi are also mostly gathered together, as shown in the yellow dotted circle in Fig. 3, and most of the germplasms from Tunchang, Qiongzhong and Danzhou in Hainan were also gathered together, as shown in the blue dotted circle in Fig. 3. This showed that the genetic background of germplasm with similar geographical origin was similar, and the genetic relationship between germplasm was closer, it also confirmed the accuracy of UPGMA cluster analysis.

In combination with the average gene flow *Nm* value in Table 7, the values were far less than 1, which indicated that there was genetic differentiation among the germplasms. Both the UPGMA tree clustering and the diversity parameters showed that there was genetic differentiation between the different geographically distributed germplasm of the genus of *B. balsamifera*. Among them, geographical distance affected the distance of genetic relationship between the germplasms, indicating that the genetic relationship between the germplasms with closer geographical distance was closer, which verified that the genetic communication ability between the germplasm resources with longer geographical distance was weaker.

Analysis Of The Population Structure Of The Tested Materials

The genetic structure of 51 *B. balsamifera* germplasms was analyzed by Structure 2.3.3 software based on the genotypic data obtained from 22 pairs of primers. According to the analysis of the population structure division by the genotype data of germplasm, when the characteristic type value K = 4 of the allelic variation frequency of the population sample was subject to the Hardy Weinberg equilibrium, it was determined that the most appropriate subgroup value of the sample is K = 4, as shown in Fig. 4, and the sample population can be divided into four groups, and the corresponding variety population structure map (Fig. 5) was drawn according to the K value, the natural population structure was relatively simple and clear. Group 1 was the red part, with a total of 21 germplasms; group 2 was the green part, with a total of 21 germplasms; group 3 was the blue part, with a total of 7 germplasms; and group 4 was the yellow part, with a total of 2 germplasms. Analyzing the situation of the germplasm source, we found that: The two germplasms from Guangzhou, Guangdong, were still gathered together. Some germplasms from Danzhou, Hainan, as well as those from Tunchang, Qiongzhong, Wanning and Wuzhishan, Hainan, are concentrated in group 1. Germplasms 1–7 and 12 from Wangmo, Zhenfeng and Luodian, Guizhou, are concentrated in group 2. Other germplasms 26–31 and 10, 11 from Luodian, Guizhou, are separately gathered in group 3, which may be the result of breeding after group selection, the above groups had significant regional characteristics and were similar to the classification results of UPGMA and PCA, which once again proved the accuracy of classification and the close relationship between genetic relationship and geographical origin of germplasm.

The diversity information of the four genetic groups was evaluated by the number of effective alleles (Ne) and Shannon's information index (I) (Table 6) by GenALEx software. The maximum number of effective alleles observed in Group 1 (2.2538) and the minimum was observed in Group 4 (1.4091), with an average value of 1.9388 in the overall population. The average Shannon's information index (I) was 0.6883 for the total flax population. The average expected heterozygosity (He) value was 0.4009, which indicated that the population had low genetic consistency and rich genetic diversity. AMOVA showed that 17% genetic variability occurred among the genetic groups, 83% variance was within the genetic groups, and the genetic variation among individuals was 0% (Fig. 6). The fixation index (Fst) for the population was 0.1752, which indicated that the large

genetic variation differentiation and long genetic distance in the current *B. balsamifera* population. The gene flow among the genetic groups was 1.7772 (*Nm* > 1), which explained the significant changes within the genome.

Table 6
Estimated genetic diversity parameters among the genetic groups as revealed by structure analysis

Pop	N	Na	Ne	I	Ho	He	uH
Group 1	20.7727	3.5455	2.2538	0.8706	0.5133	0.4880	0.5001
Group 2	20.9091	3.8182	2.0593	0.8550	0.4645	0.4648	0.4762
Group 3	7.0000	2.7273	2.0330	0.7440	0.5455	0.4462	0.4805
Group 4	1.9545	1.4091	1.4091	0.2836	0.4091	0.2045	0.2727
Mean	12.6591	2.8750	1.9388	0.6883	0.4831	0.4009	0.4324

Table 7
Analysis of molecular variance among the different genetic groups

Source	df	SS	MS	Est. Var.	Fst	P (rand ≥ data)	*Nm*
Among Pops	3	90.2535	30.0845	1.1376	—	—	—
Among Indiv	47	249.6190	5.3110	0.0000	—	—	—
Within Indiv	51	275.5000	5.4020	5.4020	—	—	—
Total	101	615.3725	—	6.5395	0.1752	0.001	1.7772

LD of EST-SSR molecular markers

LD analysis was the basis of association analysis. In plant genomics research, LD mapping used natural populations, which could be used to investigate all alleles at a population locus [16]. A total of 231 combinations were generated between 22 EST-SSR loci pairwise in this study. Figure 7 showed the results of using two different parameters D' and R^2 to describe the LD matrix, where D' described the recombination history of the population, and R^2 reflected the mutation history and recombination history of the population. As seen from the figure, the color of the LD matrix described by the two different parameters was lighter and the LD was not strong, but the LD described by the D' value was more obvious than that described by R^2. D '>0.5 represents a high degree of linkage disequilibrium, in this study, the value of D' mainly ranges from 0.3 to 0.7, accounted for 70.13% of all EST-SSR marker pairs, and the value of D '>0.5 accounts for 51.08%, as shown in Figure A in Fig. 7 shows that the part of the right half surrounded by the yellow dotted line is the area with a high concentration of linkage imbalance. which indicated that the linkage status of this part of the mark was strong. The results indicated that the recombination probability and mutation probability between genes in the selected germplasm materials were low, which was consistent with the AMOVA results in the population structure analysis, which may be related to the fact that the 51 *B. balsamifera* germplasm resources tested were from wild materials.

Association analysis between the main quality traits and EST-SSR molecular markers in *B. balsamifera*

To avoid false-positive correlation, the General linear model (GLM)and Mixed linear model (MLM) in TASSEL 2.1 software were used, and the corresponding Q values and Q + K values of each material were taken as covariables. The target trait data and EST-SSR polymorphism marker data were substituted into the GLM and MLM of Tassel2.1 software for association

analysis (Table 8). In general, the variation explanation rate of the MLM was lower than that of the GLM. In the GLM model, 4 marker loci were significantly associated with 6 quality traits ($P < 0.01$), 13 marker loci were significantly associated with 6 quality traits ($P < 0.05$), and the variation interpretation rate ranged from 19.33–57.86%. Bbf137 was significantly correlated with l-Borneol, with a variation interpretation rate of 25.77%, Bbf065, Bbf106 and Bbf377 were significantly correlated with Blumeatin, and the explanatory rates of variation were 45.13%, 57.86% and 57.11%, respectively. In the MLM model, 2 marker loci were significantly associated with 6 quality traits ($P < 0.01$), 4 marker loci were significantly associated with 6 quality traits ($P < 0.05$), and the variation interpretation rate ranged from 20.82–42.86%, Bbf201 was significantly correlated with l-Borneol, Bbf377 was significantly correlated with Blumeatin, and the variation explanation rate was 28.1%, Bbf065 was significantly correlated with sakuranetin, and the variation explanation rate was 42.86%.

The above results showed that Blumeatin was highly correlated with EST-SSR loci in GLM model, and showed extremely significant positive correlation with three EST-SSR loci. 3,3',5,7-Tetrahydroxy-4'-methoxyflavanone and 3,3',5-trihydroxy-4',7-dimethoxyfla-vanone showed significant positive correlation with four EST-SSR loci. In the MLM model, Blumeatin and Bbf377 showed an extremely significant positive correlation. In summary, the trait Blumeatin had the best correlation with EST-SSR loci, and had an extremely significant correlation with Bbf377 in both models.

Table 8

The EST-SSR loci associated with six quality traits in *B. balsamifera* and their explained proportion of phenotypic variation

General linear model (Q)				Mixed linear model (Q + K)			
Trait	Allele	P Value	R^2(%)	Trait	Allele	P Value	R^2(%)
l-Borneol	Bbf068	0.0468	34.5*	*l*-Borneol	Bbf201	0.0233	24.53*
l-Borneol	Bbf070	0.0265	41.89*	3,3',5,7-tetrahydroxy-4'-methoxyflavanone	Bbf137	0.0339	23.42*
l-Borneol	Bbf137	0.0073	25.77**	3,3',5,7-tetrahydroxy-4'-methoxyflavanone	Bbf201	0.0405	20.83*
l-Borneol	Bbf161	0.0466	29.56*	Blumeatin	Bbf377	1.99E-05	28.10**
3,3',5,7-tetrahydroxy-4'-methoxyflavanone	Bbf001	0.0200	33.32*	Sakuranetin	Bbf065	0.0067	42.86**
3,3',5,7-tetrahydroxy-4'-methoxyflavanone	Bbf137	0.0179	22.42*	Sakuranetin	Bbf106	0.0405	20.82*
3,3',5,7-tetrahydroxy-4'-methoxyflavanone	Bbf140	0.0206	19.33*	—	—	—	—
3,3',5,7-tetrahydroxy-4'-methoxyflavanone	Bbf161	0.0349	30.89*	—	—	—	—
Eriodictyol	Bbf068	0.0451	34.68*	—	—	—	—
3,3',5-Trihydroxy-4',7-dimethoxyflavanone	Bbf001	0.0380	30.51*	—	—	—	—
3,3',5-Trihydroxy-4',7-dimethoxyflavanone	Bbf020	0.0104	30.5*	—	—	—	—
3,3',5-Trihydroxy-4',7-dimethoxyflavanone	Bbf068	0.0451	34.68*	—	—	—	—
3,3',5-Trihydroxy-4',7-dimethoxyflavanone	Bbf377	0.0123	42.77*	—	—	—	—
Blumeatin	Bbf065	2.86E-06	45.13**	—	—	—	—
Blumeatin	Bbf106	4.68E-04	57.86**	—	—	—	—
Blumeatin	Bbf271	0.0270	20.82*	—	—	—	—
Blumeatin	Bbf377	1.42E-04	57.11**	—	—	—	—

Note:*: Significant correlation (P 0. 05); **: Extremely significant correlation (P<0. 01).

Discussion

Quality traits evaluation of B. balsamifera germplasm resources

The quality traits of medicinal plants were often affected by the interaction between environmental diversity and genetic diversity and were closely related to the quality of medicinal plants [17]. The coefficient of variation of crop traits often reflects the difference in the influence of crop traits on germplasm quality, which can be used as the basis for germplasm improvement [18], the greater the difference in coefficient of variation, the richer the genetic basis, and the more likely there

was excellent allelic variation in the population composed of test materials [19]. In this study, the results showed that there were high genetic differences in 6 quality traits of 51 *B. balsamifera* germplasms, and the coefficient of variation reached 53.1% − 85.5%, indicated that *B. balsamifera* germplasms had large differences among germplasm resources, great improvement potential, and more openness in germplasm innovation, such as parent selection. The correlation analysis among the quality characteristics can reveal the influence of the proportion of characteristics on the quality of *B. balsamifera*. Table 4 showed that all the characteristics of *B. balsamifera* are positively correlated, with a significant and extremely significant correlation accounting for 80%, which indicated that there was no mutual restriction between the characteristics, the characteristics can complement each other, which can be used together as the evaluation conditions for excellent germplasms of *B. balsamifera*. Principal component analysis can simplify screening indicators and select key indicators [20]. In this study, two principal components were simplified with a cumulative contribution rate of 85.474%, which indicated that these two principal components were the main characteristics of germplasms, and the contribution rate of the first principal component was the largest (63.93%), which indicated that it was easier to improve it. Whether the germplasm resources are the best cannot be judged by the single prominent trait index. Only the germplasm resources with the most prominent comprehensive traits are the most suitable resources for promotion or innovative utilization [21]. A comprehensive score model was established by combining the membership function with the principal component analysis to evaluate 51 germplasms. The F value can be used to directly evaluate the quality traits. The F value of NO.17 germplasm was the highest, and the comprehensive performance was the best. The F value of the germplasms ranked NO.2–5 was greater than 1, which can be used as germplasm resources for variety breeding and innovation.

Genetic diversity analysis and population structure analysis of B. balsamifera germplasm resources based on EST-SSR markers

SSR markers could efficiently and quickly evaluate the genetic diversity of species, and high genetic diversity could represent the health of the species population with strong adaptability [22]. Previous studies on *B. balsamifera* have used SRAP, AFLP and RAPD molecular markers [5, 6], however, SSR molecular markers have not been used to explore the genetic diversity of *B. balsamifera* germplasm. In this study EST-SSR markers were used in *B. balsamifera* for the first time, and a total of 22 pairs of primers and 106 loci were used to analyze the genetic diversity of 51 germplasms. The results showed that the number of primer alleles was high (102), and the effective alleles accounted for 53.5%, representing a wide range of germplasm sources, rich types and more representative samples. The PIC value of 0.488 was a moderate polymorphism, representing a high primer polymorphism, which was conducive to the differentiation of tested germplasms with different genetic backgrounds. Gene flow (*Nm*) was 0.203, and there was almost no genetic drift between representative germplasms, with a high degree of genetic differentiation. The above analysis results showed that the germplasm selected in this experiment had a wide range of sources, rich types, extensive genetic basis, and strong representativeness. The method used was accurate and sensitive, which can effectively evaluate the genetic diversity of *B. balsamifera* germplasm resources, and the primers screened by the research were highly effective and had a good discrimination effect. At the same time, in the future research process, the collection scope of germplasm can be appropriately expanded, and the introduction and utilization of germplasm can be strengthened to expand the genetic diversity of parents.

Genetic diversity analysis was the basis of SSR molecular marker and quality trait association analysis [13], genetic diversity and population of germplasm resources (or natural populations) analysis of somatic genetic structure is the premise of association mapping [23–25]. The UPGMA method and structure analysis were used to verify each other in this study, all germplasms were divided into four groups, and each group was highly related to the germplasm source. Principal component analysis also verified that the genetic relationship of the germplasm was closely related to the geographical source. The two-dimensional and three-dimensional PCA maps were highly coincident, which indicated that the principal component analysis clustering was feasible and accurate and could be used to identify the genetic background and genetic relationship of *B. balsamifera* germplasms [26]. In addition, the diversity analysis of the grouped population showed that the Ne value accounts for 67.44% of the Na value, and the Ho value of all groups was greater than the He value, which indicated that the genetic diversity among populations was rich, which may be related to the diversity of climate and environment in China's vast regions and provides favorable conditions for germplasm resource innovation. The structural analysis has explored that the

genetic relationship of germplasm resources is closely related to the geographical origin, and the genetic background of those with far genetic relationship was often different. Therefore, when improving the germplasm of *B. balsamifera* in the future, the germplasm with far genetic relationship and excellent quality traits can be selected for breeding, which can not only obtain offspring with rich genetic differences, but also improve the genetic diversity of the germplasm to a certain extent.

Correlation analysis between quality traits and EST-SSR molecular markers in B. balsamifera

Population LD was often affected by genetic drift, natural selection, population stratification and other factors in association analysis, among which population stratification would interfere with association results and was considered to be the main factor of false association and false association [27]. In order to avoid false positive associations, the Q and Q + K corresponding to each population of *B. balsamifera* germplasms was taken as the covariate during the analysis, and the quality traits were associated based on GLM and MLM models, effectively reducing and eliminating the impact of population stratification on the association results. Compared with LD, association analysis did not need to build a mapping population, and has the advantages of multiple sites and good positioning effect [28]. This study was the first time to associate the quality traits of *B. balsamifera* with EST-SSR marker analysis, and compared six significant association markers of quality traits. It was found that Bbf137 and Bbf377 were repeated in the same character markers, indicating that the markers were repeatable and stable. Markers significantly associated with multiple traits are Bbf137, Bbf161, Bbf001, and Bbf377, among which Bbf137 and Bbf377 are highly significantly associated with *l*-Borneol, 3,3’,5,7-tetrahydroxy-4’-methoxyflavanone, 3,3’,5-Trihydroxy-4’,7-dimethoxyflavanone, and Blumeatin. Therefore, these two markers may have a good role in the improvement of comprehensive quality traits, which provided some reference for the application of molecular design and breeding. In addition, the study found that some traits detected in GLM, such as Eriodictyol and 3,3 ', 5-Trihydroxy-4', 7-dimethoxyflavanone, were not detected in MLM system analysis. The reason was that the influence of population structure (Q value) and kinship (K value) on association analysis was considered in MLM analysis, so the number of association markers obtained by MLM analysis was less, but the results were more accurate, which may also be related to population size [29–31]. In addition, six quality traits were detected by the two models, which indicated that the number of markers used in this study was sufficient and the coverage in the *B. balsamifera* gene was high, and the materials provided by the experiment can provide sufficient genetic variation. As there was no relevant literature report, the reliability and authenticity of relevant markers can be verified in the future by increasing the population number, expanding the collection range of germplasm resources [29] and providing more abundant genetic variation.

Conclusions

In this study, 22 EST-SSR molecular markers were used to analyze the genetic diversity and population structure of 51 *B. balsamifera* germplasm quality traits, and LD analysis was carried out. Through GLM (Q) and MLM (Q + K) two analysis models, the correlation analysis of *B. balsamifera* quality traits and EST-SSR markers was carried out. The results showed that the coefficient of variation of six quality traits was 53.1% − 85.5%, with wide variation range and rich genetic basis. Each SSR marker had an average of 4.636 alleles, and the average polymorphism information content was 0.488, indicating that the primer had high polymorphism and strong discrimination ability, and could accurately evaluate the genetic diversity of all germplasm. UPGMA clustering divided the germplasm into 4 groups, and the clustering results of each group suggested that the aggregation characteristics of germplasm were related to geographical origin; PCA analysis showed that the genetic background of germplasm with similar geographical origin was similar, and the genetic relationship between germplasm was closer; The population structure analysis showed that the samples were divided into four groups when $\triangle K = 4$, and the germplasm grouping was closely related to the genetic relationship of the germplasm, which again confirmed the accuracy of UPGMA classification. The results of LD analysis showed that the recombination probability and mutation probability between genes in germplasm materials were low, and the genetic diversity among materials was rich. In association analysis, 23 loci related to quality traits were detected by the two analysis models, and their variance interpretation rates were 19.33–57.86%. Among them, the trait Blumeatin had the best correlation with EST-SSR loci, and had a greater correlation with Bbf377.

To sum up, this study not only confirmed the feasibility of applying EST-SSR markers to the genetic diversity analysis of *B. balsamifera*, but also provided a theoretical basis for applying EST-SSR markers to the correlation analysis of quality traits in *B. balsamifera* germplasm resources, and also provided a molecular technical means for the subsequent molecular identification of *B. balsamifera* germplasm and molecular assisted breeding of excellent germplasm.

Materials And Methods

Plant materials

The experimental samples were 51 germplasms from 12 counties (cities) of 4 provinces (Table 9). The specimen was identified by the associate researcher Fu-lai Yu and was planted at the germplasm resource nursery of *B. balsamifera*, Danzhou, China.

Table 9
Source of germplasms

Oringin	Accessions	Number	Germplasm type
Luodian, Guizhou	7	1–6, 11	Wild
Luodian, Guizhou	3	7, 10, 12	Cultivate
Luodian, Guizhou	6	26–31	Group selection
Wangmo, Guizhou	1	49	Wild
Zhenfeng, Guizhou	1	45	Wild
Longlin, Guangxi	1	35	Wild
Tianlin, Guangxi	6	36–40, 43	Wild
Xilin, Guangxi	3	41, 42, 44	Wild
Nangning, Guangxi	1	46	Cultivate
Guangzhou, Guangdong	2	47, 48	Cultivate
Danzhou, Hainan	5	8, 9, 13 16, 17	Variation
Danzhou, Hainan	4	14, 15, 18, 19	Individual selection
Danzhou, Hainan	1	25	Cultivate
Qiongzhong, Hainan	6	20–24, 34	Wild
Tunchang, Hainan	2	32, 33	Wild
Wanning, Hainan	1	51	Wild
Wuzhishan, Hainan	1	50	Wild

Determination Of Quality Traits

Determination of *l*-Borneol

HP-5 quartz capillary column (0.32 mm × 30 m, 0.25 μm); Took 80 ℃ as the starting temperature and kept it for 2 min; Raised the temperature to 100 ℃ at 5 ℃/min, and then to 200 ℃ at 20 ℃/min; The temperature of injection port was 220 ℃; FID

detector, the detector temperature was 240 °C; The injection volume was 0.6 μ. The split ratio was 9: 1. According to the above conditions, determine methyl salicylate + *l*- Borneol reference substance and 51 tested samples of *B. balsamifera* [32].

Determination Of Five Components

Chromatographic column: Agilent TC-C18 column (250 mm × 4.6 mm, 5 μm); The mobile phase was acetonitrile (A) − 0.5% phosphoric acid solution (B), gradient elution (0–15 min, 15% − 22% A; 15–21 min, 22% − 24% A; 21–30 min, 24% − 30% A; 30–45 min, 30% − 33% A; 45–55 min, 33% − 38% A; 55–70 min, 38% − 55% A); Detection wavelength 285 nm; Column temperature 35 °C; Volume flow 0.9 mL/min; Injection volume 10 μL.

Weighed 0.5 g of the sample, added 10 mL of 90% methanol, weighed the mass, conducted ultrasonic treatment for 45 min, placed it at room temperature, weighed the mass again, used 90% methanol to supplement the lost mass, shook it up, filtered it, took the continued filtrate, and determined the contents of the five components with the above methods [33].

EST-SSR Molecular Marker Technique

Healthy young leaves were collected and frozen in liquid nitrogen. The DNA of the sample was extracted using the magnetic bead genomic DNA extraction kit. The concentration and purity of nucleic acids were detected with a Nano DROP 8000 ultra-micro spectrophotometer, and the DNA concentration was adjusted to 50–200 ng L^{-1}. The mixed DNA samples of germplasm with large genetic background differences were selected from all samples, the synthesized EST-SSR primers were screened, rescreened and verified, and 22 pairs of EST-SSR primers with clear bands, high polymorphism and good stability were screened (Table 10).

Table 10
Information of 22 EST-SSR primers

Primers	Repeat motifs	Primer sequence (5′ – 3′)	Size of bands Amplified (bp)	Fluorescent labelled
Bbf001	$(CT)_7$	F:TGTCACTGCACCACCTTTG	342–346	HEX
		R:GGCACAAACGGCTATGGTC		
Bbf020	$(AAG)_4$	F:CCCTCGCCCTATTTCACCG	222–226	FAM
		R:CATCCCGGCGATCCTTCC		
Bbf041	$(CT)_7$	F:CTGTCAGCACCATTTCCCG	285–289	HEX
		R:GTGGAACAACGCTATCTGGC		
Bbf065	$(CGG)_5$	F:CGCCGTAGCTGAAATCGTG	263–269	FAM
		R:CGCTTGCGAACCCATAGTC		
Bbf068	$(AG)_8$	F:TGTTGCATTCCATGATCCACC	311–319	FAM
		R:CTTTACTGAAGCTGTGCAACTC		
Bbf070	$(GT)_6$	F:CTGGTCTTGTAAGCGGTGC	321–333	FAM
		R:AGCACATGGTTTCCTGGATTAG		
Bbf103	$(ATT)_4$	F:ACGGCACATGGACCAAATC	255–266	FAM
		R:TCCATCCAATTCGTTTCCTCG		
Bbf106	$(AC)_8$	F:GGGCAAATCCACTTTGGGAG	222–240	HEX
		R:TCATGTCGGAAATAGAAGCGG		
Bbf137	$(AAAC)_4$	F:GAGTGTTGCTCATCCAGGC	176–188	HEX
		R:TGCAAGTCATCAGGCAAAGG		
Bbf140	$(ACG)_4$	F:AGCATTCTGGCAGCTCAAC	297–306	HEX
		R:CAGCAGGCATTGAAGACGG		
Bbf146	$(CT)_{10}$	F:GCAAAGGAGTGGTGAAGCC	248–256	ROX
		R:TGAACACGGATTTGGGCAAG		
Bbf156	$(AAG)_4$	F:TCCACTAGGCACGGCATAC	222–228	HEX
		R:TCGTTTCTTGAGCGTCACC		
Bbf161	$(CT)_8$	F:TCATCAAAGCTGCAACCCG	275–281	FAM
		R:GTTCGGTTTCGCCTAGTGG		
Bbf193	$(ATC)_6$	F:AGTCGATCGGGTGTGCTTC	272–286	HEX
		R:GAACGAGGGACGGATTTGC		
Bbf201	$(CT)_8$	F:TCTCTCCACCGCAGGATTG	237–254	ROX
		R:ACTGCTGCTCCGAGATACC		
Bbf219	$(AG)_7$	F:TGAAACTGACCATCGACGC	233–239	HEX

Primers	Repeat motifs	Primer sequence (5′ – 3′)	Size of bands Amplified (bp)	Fluorescent labelled
		R:GGCAGCCTCCTTCCACAG		
Bbf250	$(AG)_8$	F:CTCCGACAGTTCGAAGTCC	228–239	FAM
		R:CGAAGCTCCGATTGGTACG		
Bbf271	$(CT)_7$	F:TACTGCGTTACGGCCCTC	172–176	FAM
		R:CGATGACTTCAAGCGCCAG		
Bbf283	$(ACC)_5$	F:AACCGCAATCCCTCACCAG	390–397	FAM
		R:CCTTGTTGACTCAGATTCCCG		
Bbf289	$(AGC)_5$	F:TCAACCACAGAAGTGCCAG	205–221	FAM
		R:ATGGAAGCCCGGAATGGAG		
Bbf377	$(ACTC)_6$	F:AGATTACCATTGCTGCAGACC	205–374	TAMRA
		R:AAACAAGAAGCCGGAACCC		
Bbf384	$(ATGT)_4$	F:ACCGAGTGCACCCTTATAATC	263–434	TAMRA
		R:ACAACTTGCACTTCGTGGC		

PCR amplification reaction system: 5.0 μL 2 × Taq PCR Master Mix, 1 μL genomic DNA, 0.5 μL upstream primer, 0.5 μL downstream primer, 3.0 μL ddH_2O. The reaction procedure was as follows: pre denaturation at 95°C for 5 min; denaturation at 95°C for 30 s, gradient annealing at 62 ~ 52°C for 30 s, extension at 72°C for 30 s, 10 cycles, each cycle decreasing by 1°C; denaturation at 95℃ for 30 s, annealing at 52℃ for 30 s, extension at 72℃ for 30 s, 25 cycles; 72℃ for 20 min and storage at 4℃.

3 μL fluorescent PCR product and 1% agarose gel electrophoresis were used to detect whether the PCR band was single and whether the fragment size was consistent with the expectation. If the band was single and consistent in size, the concentration of the DNA marker was compared for quantification, and all products were diluted to the same concentration range. 10 μL formamide and 0.5 μL internal standard were mixed well, and 1 μL PCR diluted product was added for capillary electrophoresis detection.

Data Processing Method

Data analysis

Excel software was used to analyze the original data, and SPSS 24 was used to analyze the differences and correlations. GenALEx 6.502 converts the results of EST-SSR molecular markers into TXT text format.

Genetic Diversity And Cluster Analysis

Popgene 32[34] and PowerMarker 3.25[35] calculated the number of alleles (Na), number of effective alleles (Ne), expected heterozygosity (He), observed heterozygosity (Ho), Shannon index (I), polymorphism information content (PIC) and other genetic diversity indicators. The UPGMA method was used for cluster analysis, and use MEGA [36] software to make the cluster diagram. The genetic similarity coefficient was calculated by Ntsys [37], and the principal component analysis (PCA) was used to construct the two-dimensional scatter diagram and three-dimensional scatter diagram of principal components.

Group Structure Analysis

Structure 2.3.3 [38] analyzed the material group structure and determines the optimal group number K. The optimum number of populations was determined by running Ks from 2 to 8 with five independent runs with a burn-in period of 100,000. The inference of the best K was determined using the delta K method using the structure harvester programme (https://taylor0.biology.ucla.edu/structureHarvester/) among the germplasms. GenALEx 6.502 [39] analyzed the population structure.

Correlation Analysis

Linkage disequilibrium (LD) analysis, General linear model (GLM) and Mixed linear model (MLM) was used by Tassel 2.1 software [40] to estimate P value and R^2 value. The P value significance threshold was 0.05, and the extremely significant value was 0.01.

Declarations

Acknowledgments

Not applicable.

Authors' Contributions

Xin Li: interpretation of data and wrote the draft; Mei Huang: experiments, interpretation of data and revising the article; Zhen-Xia Chen: experiments; Dan Wang: experiments; Kai Wang: experiments; Xiao-Lu Chen: experiments; Ling-Liang Guan: experiments; Hong-Rui Zhang: contributed to conception and design of the article; Yu-Xin Pang: contributed to conception and design of the article; Fu-Lai Yu: contributed to conception and design of the article, revising the article and obtained funding and overall responsibility. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Hainan provincial Natural Science Foundation of China No.2019CXTD414 and Central Public-interest Scientific Institution Basal Research Fund for Chinese Academy of Tropical Agricultural Sciences No.1630032022008 1630032022022 .

Ethics approval and consent to participate

All 51 plant materials used in this study were wild plants. They were collected and cultivated at the germplasm nursery of the Chinese Academy of Tropical Agricultural Sciences, Danzhou,Hainan, with permission. All methods used for studying the plants and their materials met the requirements of relevant institutional, national and international guidelines and legislation.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Data Availability Statement

We ensure that the experimental data is truly available and that data supporting the results of this study can be obtained from the first author upon reasonable request.

References

1. Pang YX, Wang D, Fan ZW, Chen XL, Yu FL, Hu X, Wang K, Yuan L. Blumea balsamifera-A phytochemical and pharmacological review. Molecules. 2014;19:9453–77.
2. Yuan Y, Hang M, Pang YX, Yu FL, Chen C, Liu LW, Chen ZX, Zhang YB, Chen XL, Hu X. Variations in Essential Oil Yield, Composition, and Antioxidant Activity of Different Plant Organs from Blumea balsamifera (L.) DC. at Different Growth Times. Molecules. 2016;21:8.
3. Xiao YF, Huang M, Yu FL, Chen ZX, Liao L, Pang YX. Analysis on genetic diversity of phenotypic traits of Blumea balsamifera germplasm resources. Fujian J Agric. 2021;36:157–67.
4. Xie XL, Chen ZX, Pang YX, Guan LL, Guo KM. Research Progress of Blumea balsamifera Resources. World Science and Technology Modernization of Traditional Chinese Medicine. 2017; 19: 2024–2029.
5. Zhang YB, Yuan Y, Pang YX, Wang D, Hu X. Comparative analysis of genetic diversity between SRAP and AFLP in Blumea balsamifera. J South Agric. 2016;47:1261–7.
6. Pang YX, Wang WQ, Zhang YB, Mo TH, Yuan Y. Study on the clonal diversity and structure of wild populations of Blumea balsamifera. Guangxi Plants. 2010;30:209–14.
7. Huang KM, Zou YJ, Shi YQ, Ying YN, Yan SB, Bao JS. Pumpkin cultivation SSR molecular marker analysis and diversity of agronomic traits in cultivars. Nuclear Agron News. 2021;35:2746–55.
8. Li JF, Zhang B, Quan JZ, Wang YF, Zhang XM, Zhao Y, Yuan XL, Jia XP, Dong ZP. Detection of major agronomic traits associated loci in millet based on SSR markers and allelic variation analysis. Chin Agricultural Sci. 2019;52:4453–69.
9. Ye XR, Liu JT, Li YP, Zhu HS, Wen QF. Based on EST-SSR markers Identification of germplasm resources of wax gourd by MCID method. J Nuclear Agric. 2021;35:780–8.
10. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
11. Prichard JK, Stephens M, Rosenberg NA, Peter D. Association mapping in structured populations. Am J Hum Genet. 2000;67:170–81.
12. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu JM, Arnett DK, Ordovas JM, Buckler ES. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42:355–60.
13. Babu BK, Mathur RK, Venu MVB, Shil S, Ravichandran G, Anita P, Bhagya HP. Genome-wide association study (GWAS) of major QTLs for bunch and oil yield related traits in Elaeis guineensis L.Plant Science. 2021;305.
14. Rasool S, Mahajan R, Nazir M, Bhat KA, Shikari AB, Ali G, Bhat B, Bhat BA, Shah MD, Murtaza I, Nazir N, Sofi PA, Bhat MA, Zargar SM. SSR and GBS based GWAS study for identification of QTLs associated with nutritional elemental in common bean (Phaseolus vulgaris L).Scientia Horticulture. 2022;306.
15. Ha RK, Kyu JS, Min NG, Ki JP, Si HR, Chang YM, Ju KL. Genetic characterization and association mapping in near isogenic lines of waxy maize using seed characteristics and SSR markers. Genes & Genomics. 2021;43:1.
16. Liao GL, Li ZY, Huang CH, Zhong M, Tao JJ, Qu XY, Chen L, Xu XB. Genetic diversity of inner quality and SSR association analysis of wild kiwifruit (Actinidia eriantha).Scientia Horticulturae. 2019; p.248.
17. Huang QL, Xie LMX, Yang BS, Deng KZ, Ge F, Wu B, Xu ZB, Li J. Study on the Correlation between agronomic Traits and Quality of Pueraria. Chin Materia Med. 2021;44:1563–8.
18. Hou JF, Wang CL, Hong XJ, Zhao JM, Xue CC, Guo N, Gai JY, Xing H. Association analysis of vegetable soybean quality traits with SSR markers.Plant Breeding. 2011;130.
19. Li WZ, Yao XQ, Xu CX, Min L, Chen QS, Hu GH. Correlation analysis between yield and quality correlation and simple repeat sequence (SSR) of maize inbred lines. J Agricultural Biotechnol. 2012;20:1369–77.
20. Shen BY, Li L, Zong XX. Analysis and evaluation of main quality traits of pea germplasm resources. Shandong Agricultural Sciences. 2022;54:1–17.

21. He S, Zhang J, Xie Q, Ma WX, Ni YJ, Ye QH, Chen QX. Quality Analysis and Comprehensive Evaluation of Nelumbo roxburghii Germplasm Resources in Different Regions. Res Dev Nat Prod. 2018;30:1415–22.
22. Amos W, Harwood J. Factors affecting levels of genetic diversity in natural populations. Phil Trans R Soc Lond B. 1998;353:177–86.
23. Harris BP, Stokes bury KDE. The spatial structure of local surficial sediment characteristics on Georges Bank, USA. Cont Shelf Res. 2010;30:1840–53.
24. Wang ML, Zhu CS, Barkley NA. Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection. Theor Appl Genet. 2009;120:13–23.
25. Kline JB, Moore DJ, Clevenger CV. Activation and association of the Tec tyrosine kinase with the human prolactin receptor: Mapping of a Tec/Vav-receptor binding site. Mol Endocrinol. 2001;15:832–41.
26. Li Q, Geng Z, Yang QC, Shu WT, Li JH, Chang SH, Zhang DH, Zhang BL. SSR Genetic Diversity Analysis of 50 Soybean Germplasm Resources in Huang Huaihai. Seed. 2021;40:39–44.
27. Wen ZX, Zhao TJ, Zheng YZ, Liu SH, Wang SE, Wang F, Gai JY. Association analysis of agronomic and quality traits wih SSR markers in Glycine max and Glycine soja in China: â…¡. Exploration of elite alleles. Acta Agronomical Sinica. 2008;34:1339–49.
28. Ashfaq M, Rasheed A, Sajjad M, Ali M, Rasool B, Javed MA, Allah SU, Shaheen S, Anwar A, Ahmad MS, Mubashar U. Genome wide association mapping of yield and various desirable agronomic traits in Rice.Molecular biology reports. 2022.
29. Tong ZJ, Chen XJ, Fang DH, Zeng JM, Wu XF, Xiao BG. Genetic diversity of SSR markers in 231 flue-cured tobacco germplasm resources and their association with agronomic traits and chemical components. Chin J Tob. 2017;23:31–61.
30. Yan M, Han SY, Dong WZ, Huang BY, Liu H, Zhang ZX, Xu J, Sun ZQ, Du P, Tang FS, Zhang XY. Association Analysis of Quality Traits of Major Peanut Cultivars in China. J Plant Genetic Resour. 2013;14:1064–71.
31. Xu SX, Wang XW, Tian Q, Kong XH, Liu L, Si AJ, Wang J, Yu Y. SSR Association Analysis of Genetic Diversity and Fiber Quality Traits of Xinjiang Early maturing Upland Cotton Germplasm Resources. J Cotton. 2020;32:233–46.
32. Yu FL, Zhao N, Wu ZS, Huang M, Wang D, Zhang YB, Hu X, Chen XL, Huang LQ, Pang YX. NIR Rapid Assessments of *Blumea balsamifera* (Ai-na-xiang) in China.Molecules. 2017; p.22.
33. Hu X, Yu WHF, Yuan FL, Huang C, Zhang M, Chen YB, Wang ZX, Pang D. Quality evaluation of *Blumea balsamifera* based on fingerprint and multicomponent content determination. Chin Herb Med. 2021;52:3679–88.
34. Ye FC, Yang R, Boyle T, Popgene. Microsoft Window-based freeware for population genetic analysis, version 1. Volume 31. Edmonton: University of Alberta and Center for International Forestry Research; 1999.
35. Liu KJ, Muse SV. PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics. 2005;21:2128–9.
36. Sudhir K, Glen S, Koichiro T. MEGA7: molecular evolutionary genetics analysis Version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4.
37. He RL, Han YL, Wang F, Liu JQ, Wang LH. Development of transcriptome SSR markers and genetic diversity of Astragalus mongolicus. Chin J Materia Medica. 2018;43:1838–43.
38. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
39. Peakall R, Smouse PE. GenALEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes. 2006;6:288–95.
40. Meng YX, Meng YL, Wang JC, Erjing S, Zhang HJ, Ren PR, Ma XL, Li BC, Yang K, Wang HJ. Genetic diversity of highland barley and its association with agronomic traits and SSR markers. Acta Agron Sinica. 2016;42:180–9.

Figures

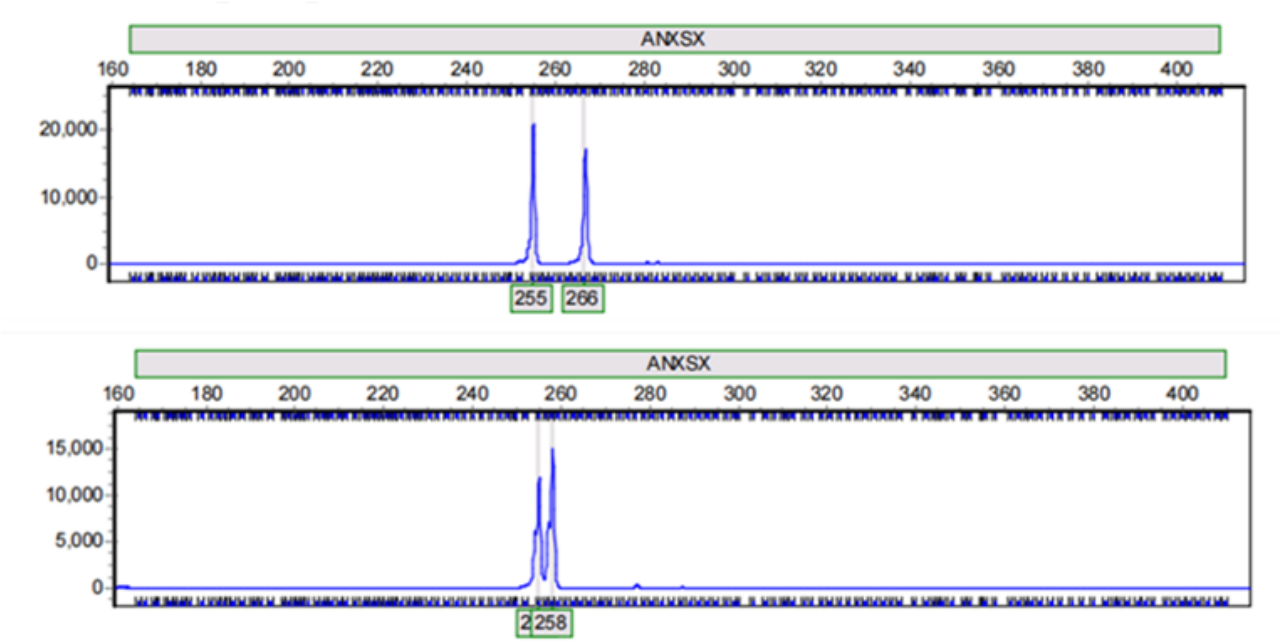

Figure 1

The amplicons of germplasms No. 21 and No. 22 were 255, 258, 266 bp

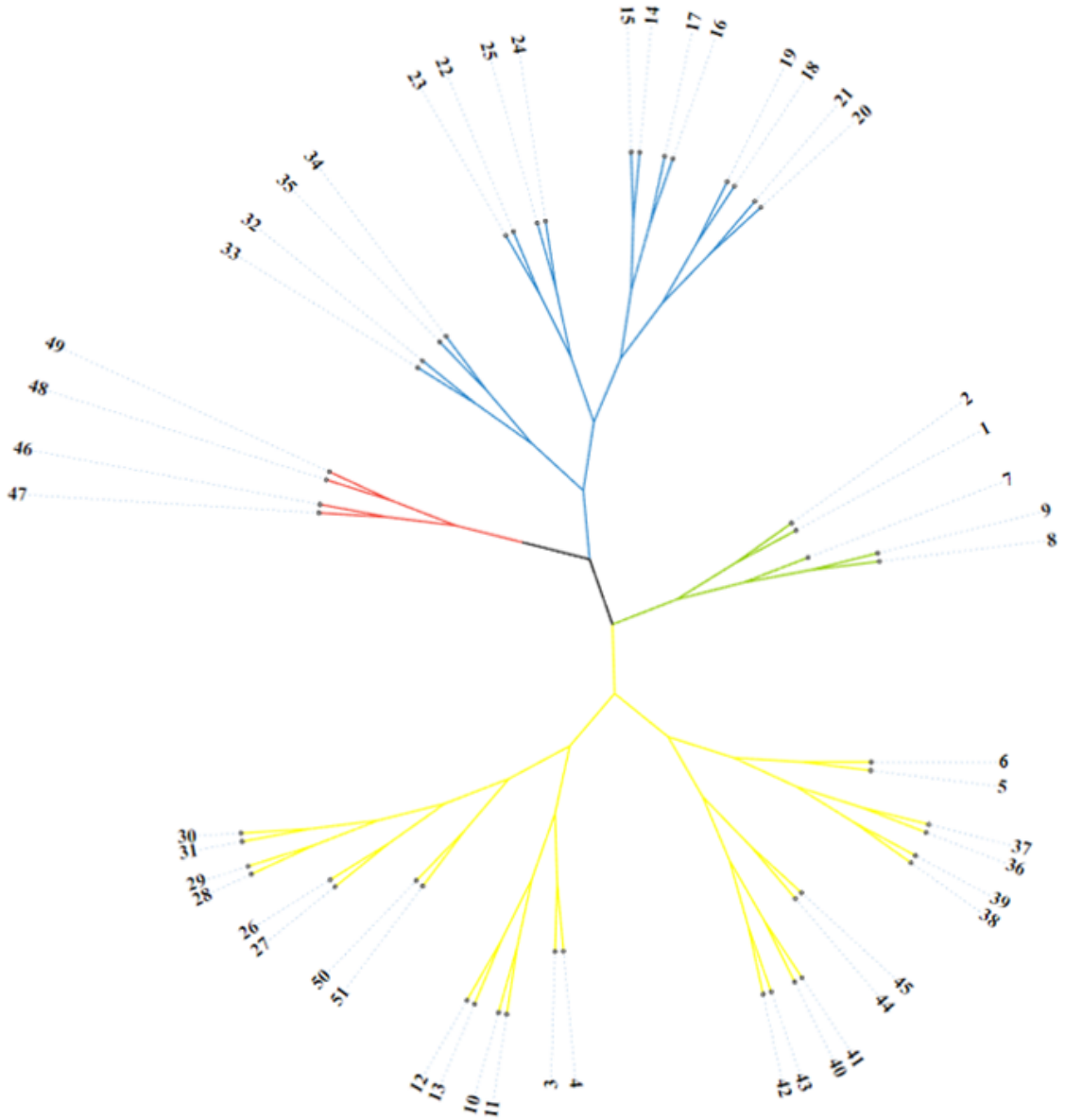

Figure 2

UPGMA clustering of *B. balsamifera* based on the genetic similarity coefficient

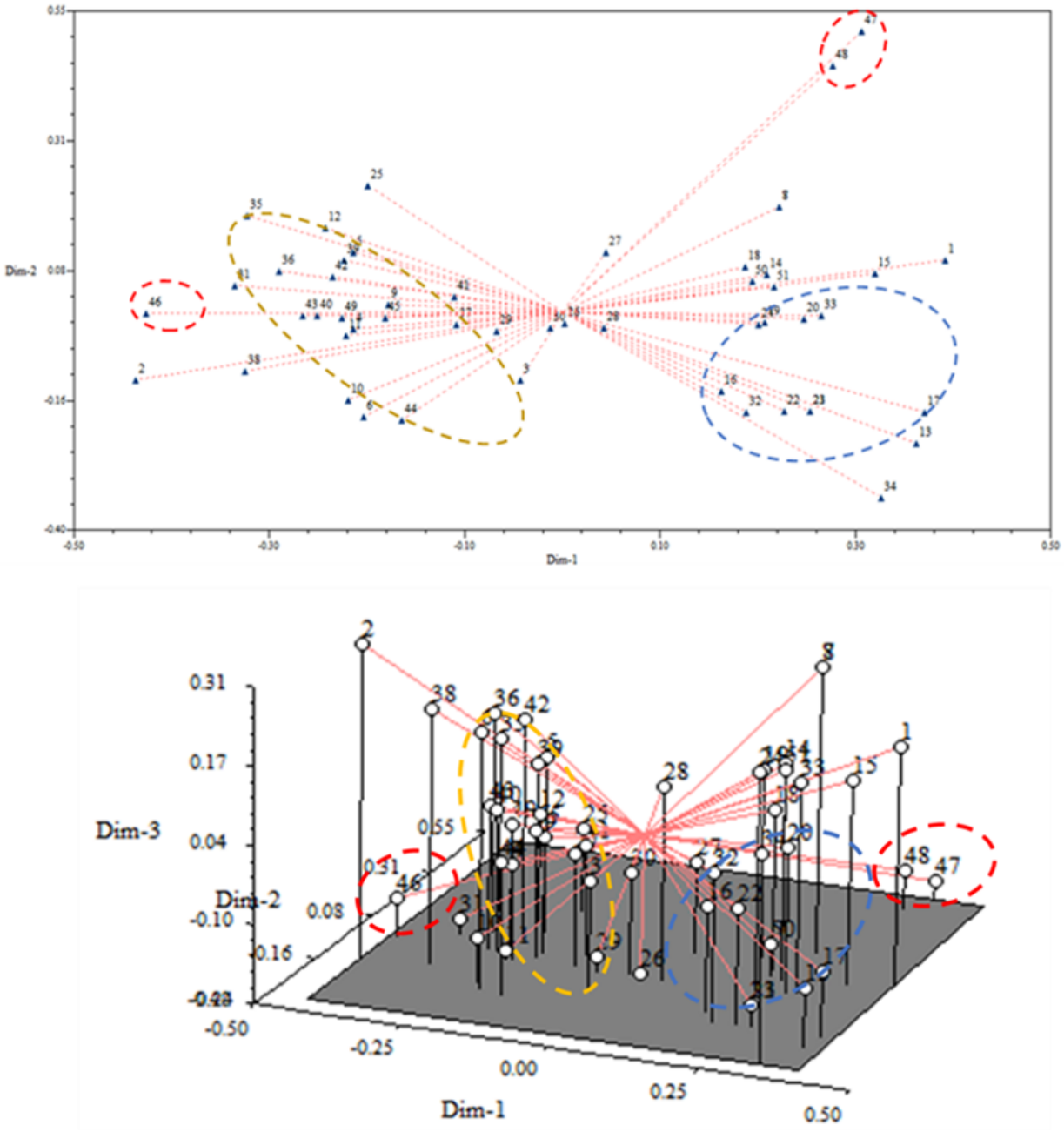

Figure 3

Two-dimensional scatter diagram and three-dimensional scatter diagram of principal coordinate analysis of EST-SSR markers

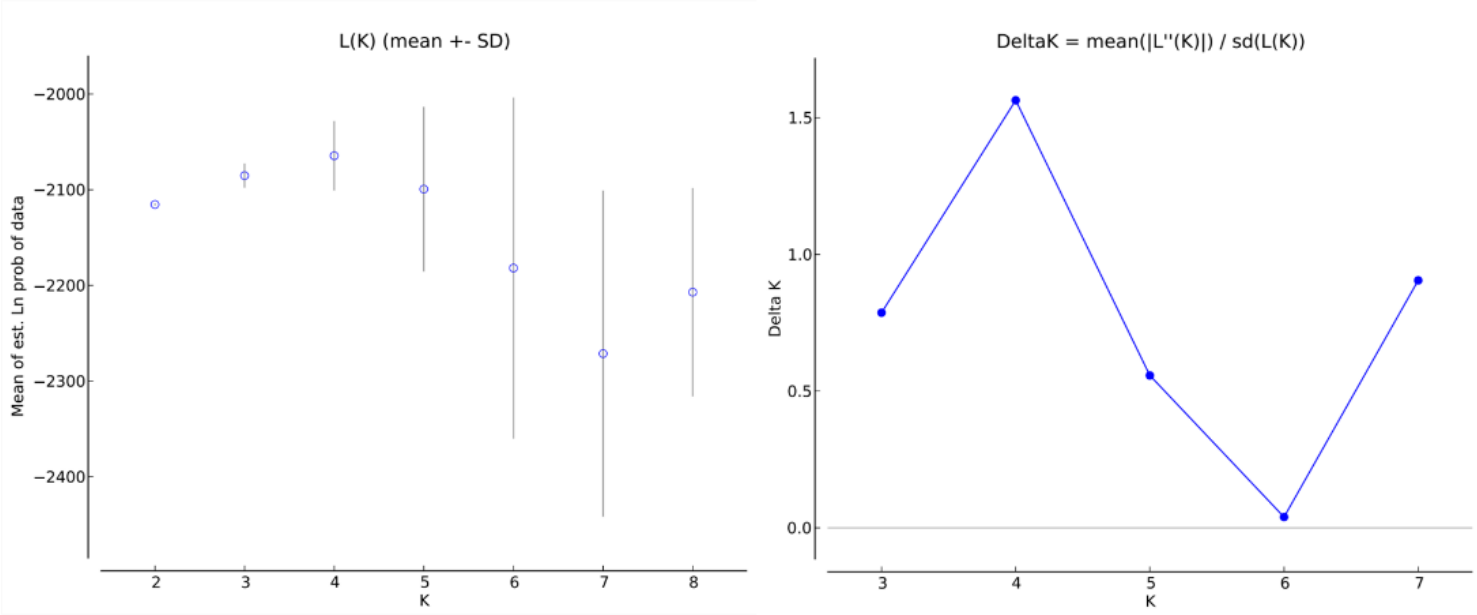

Figure 4

Estimation of the average number of genetic groups using STRUCTURE HARVESTER

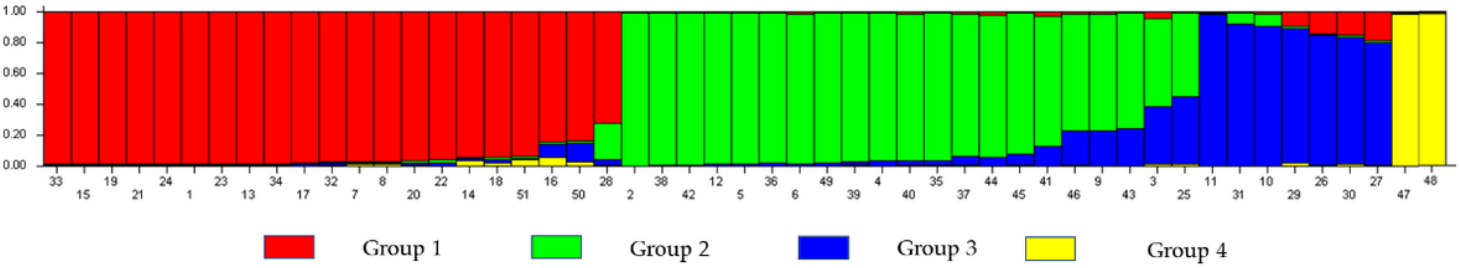

Figure 5

Population structure model of 51 *B. balsamifera* samples based on the genotypic data revealed from EST-SSRs

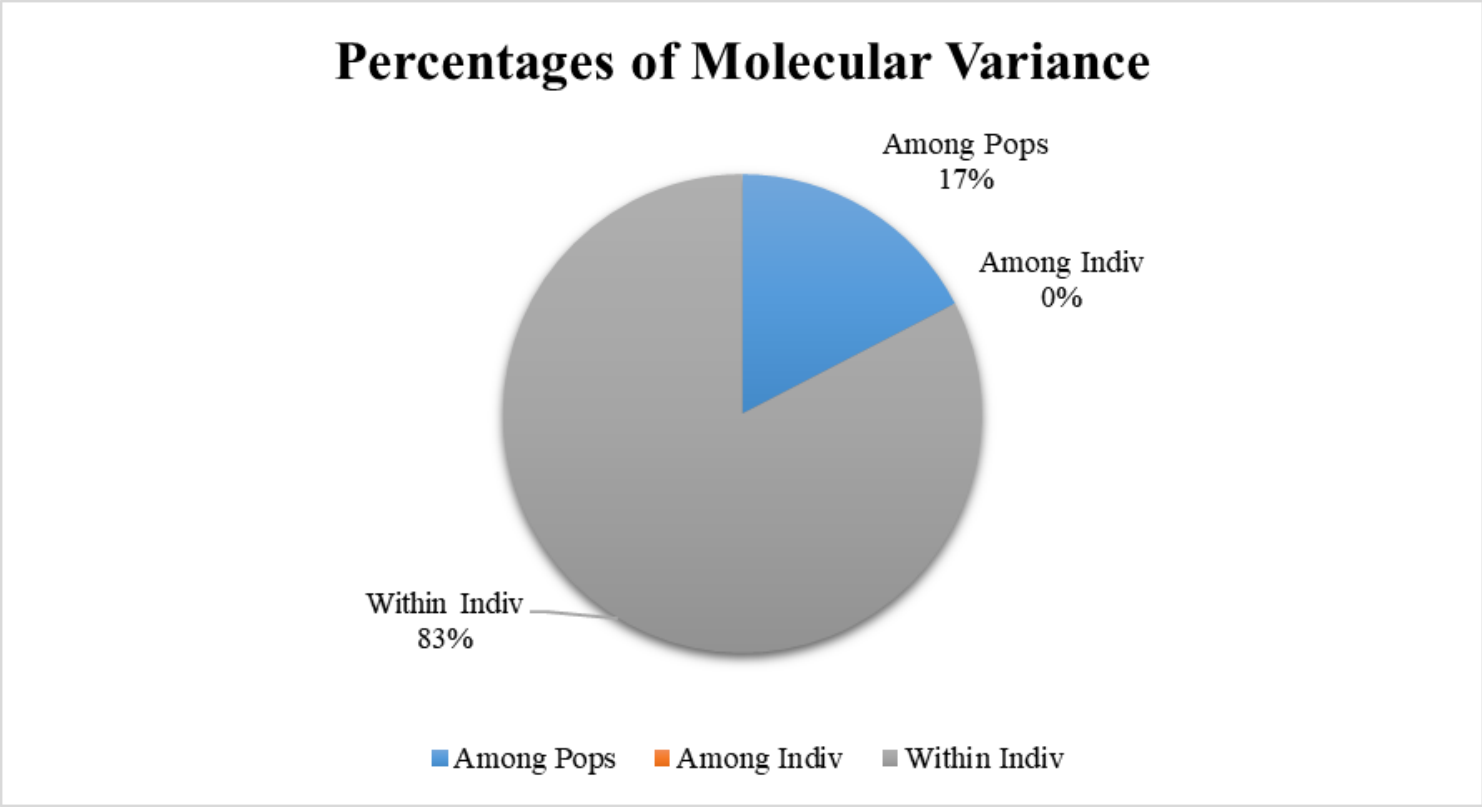

Figure 6

Percent of molecular variation in *B. balsamifera* populations

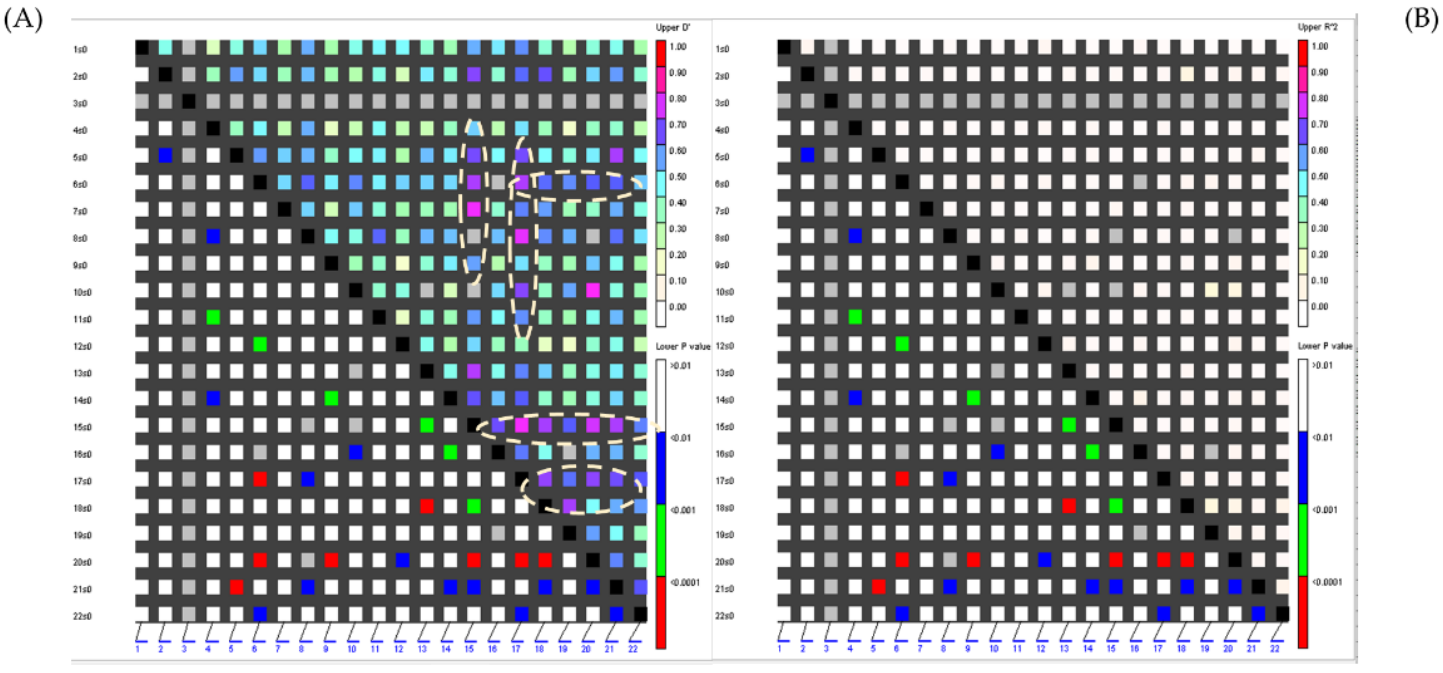

Figure 7

Distribution of LD among 22 EST-SSR loci on 51 *B. balsamifera* germplasms

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GENAlexAnalysisresultsofpopulationgeneticdiversity.xlsx
- GLMandMLMrawdata.xlsx
- Germplasmspecificinformation.xlsx
- QualitytraitdataandSSRgeneticdiversitydata.xlsx
- ThesourceofFigure1.Theamplificationmapsofgermplasm21and22arebothfromthisdocument.pdf