Development and characterization of EST-SSR markers in Rhodomyrtus tomentosa based on transcriptome

Lina Sun
Guangxi Zhuang Autonomous Region Forestry Research Institute
Jinhua Li
Guangxi Zhuang Autonomous Region Forestry Research Institute
Kaidao Sun
Guangxi Zhuang Autonomous Region Forestry Research Institute
Huaxin Wang
Guangxi Zhuang Autonomous Region Forestry Research Institute
Kaitai Yang
Guangxi Zhuang Autonomous Region Forestry Research Institute
Qi Chen (34422070@qq.com)
Nanning Goldtech Bioscience Ltd. Co. https://orcid.org/0000-0002-2000-5269
Mao Lin
Guangxi Zhuang Autonomous Region Forestry Research Institute https://orcid.org/0000-0002-0408-848X

Research Article

Keywords: Transcriptome, EST-SSR, Rhodomyrtus tomentosa, Genetic diversity, Polymorphism

Posted Date: May 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1610424/v1

Version of Record: A version of this preprint was published at Genetic Resources and Crop Evolution on February 13th, 2023. See the published version at https://doi.org/10.1007/s10722-022-01528-x.

Abstract

Rhodomyrtus tomentosa (Ait.) Hassk is a flowering evergreen plant with traditional medicine and ornamental usages. This study aimed to develop a set of EST-SSRs markers for genetic diversity analysis of R. tomentosa. Transcriptome sequencing was performed to obtain the expressed sequence tags in different plant tissues. A total of 51,486 unigenes with a mean length of 1,173 bp were achieved. 18,879 SSRs were identified in 14,132 unigenes, in which 3,541 unigenes contained more than one SSR. The top three SSR repeat types were mononucleotide (8,126, 43.04%), dinucleotide (5,846, 31.06%) and trinucleotide (4,610, 24.42%). The most abundance motifs were A/T (7,816, 41.40%), followed by AG/CT (5,002, 26.50%) and AAG/CTT (934, 4.95%). Of these SSRs, 11,726 SSRs were eligible for designing of flanking primers. Among the 100 randomly selected primers, 50 primers generated corresponding PCR products, whereas 23 primers were polymorphic. Thirteen primers with good reproducibility were used to characterize the genetic relationship of sixteen natural R. tomentosa plants. The mean value of observed heterozygosity (Ho), expected heterozygosity (He), fixation index (Fst), and Shannon information index (I) were 0.240, 0.414, 0.413, and 0.641, respectively. UPGMA (unweighted pair-group method with arithmetic averages) dendrogram divided the sixteen R. tomentosa plants into five groups, which were able to distinguish the plants according to geographical origin. This study presented a set of EST-SSRs that could be useful for population genetic relationship analysis in R. tomentosa.

Introduction

Rhodomyrtus tomentosa (Ait.) Hassk., an evergreen shrub plant, is widely distributed in South Asia(Hue et al., 2015). It belonged to the Myrtaceae family, including over 3,000 species (Freire et al., 2018). *R. tomentosa* has a long history of herbal usage in many countries (Zhao et al., 2020). It has been reported to possess antibacterial activities, anti-inflammatory effects, antioxidant activity, and antitumor activity (Limsuwan et al., 2011;Jeong et al., 2013;Wu et al., 2015;Tayeh et al., 2017). *R. tomentosa* is also widely cultivated for its ornamental and edible attributes (Zhao et al., 2020). The flowers of *R. tomentosa* are pink in cluster style and bloom profusely in the spring. The ripe fruit has a purple color and bell-like form with a sweet taste (Lai et al., 2015). Its leaves and fruits were used to make various foods, like tea, wine, and jam (Lai et al., 2015).

Nowadays, due to anthropogenic activities, the habitats of *R. tomentosa* have been fragmented extensively (Xie et al., 2021). There has been a significant reduction in the population size of *R. tomentosa* in China. Comprehensive measures of conservation should be carried out to protect this plant. A better understanding of genetic diversity is essential for plant resource conservation and use (Ramanatha Rao and Hodgkin, 2002). To date, research about the genetic diversity of *R. tomentosa* has been few. A study using ISSR markers revealed that the genetic diversity of *R. tomentosa* in Hong Kong was high (Yao, 2010). The high genetic diversity in the species may provide sufficient sources for this plant to adapt to the new changing environment. In Malaysia, *R. tomentosa* also exhibited high genetic diversity in species but maintained a low diversity at the population level (Hue et al., 2015).

Simple sequence repeats (SSRs), also known as microsatellites, generally refer to repetitive sequences with a motif of 1 to 6 bases, which are widely found in the genomes of prokaryotes and eukaryotes, and are commonly used molecules marker with the characteristics of codominant inheritance, high polymorphism, and good repeatability (Feng et al., 2016;Han et al., 2018). SSR markers have been widely used in species identification, population genetic diversity analysis, gene mapping, and marker-assisted selection (Tuler et al., 2015;Ali et al., 2019;Wu et al., 2020). SSR markers can be developed from either genomic DNA or expressed sequence tag, named gSSR or EST-SSR (Durand et al., 2010). EST-SSRs, derived from transcripts, are more suitable for assessing the functional and phenotypic diversity in plant populations (Varshney et al., 2005a). A comparison between gSSR and EST-SSR in sugarcane showed that the morphological traits related to EST-SSR were more effective in identifying diverse parents (Parthiban et al., 2018). Due to their location in conserved and expressed sequences, EST-SSR is effectively interspecies transferable (Cordeiro et al., 2001;Varshney et al., 2005b;Guo et al., 2014). Transcriptome sequencing provided a cost-efficient and productive method for EST-SSR development, especially for the species without genome sequence, like *Phyllostachys violascens* and *Zingiber officinale* (Cai et al., 2019;Vidya et al., 2021).

In this study, we applied an Illumina sequencing platform to obtain the transcriptome of *R. tomentosa*. The sequences dataset was assembled and annotated. The EST-SSRs markers were also analyzed and characterized. A set of EST-SSRs primers was developed and used to apply genetic relationship analysis in *R. tomentosa*.

Materials And Methods

Plant materials

Sixteen natural plants of *R. tomentosa* were collected across Guangxi province, China (Table 1). The sample NN was selected for transcriptome sequencing. Random samples from the 16 locations were used for genetic diversity analysis. The latitude and altitude of the location of each population were recorded.

RNA isolation, library preparation, Transcriptome sequencing, de novo assembly, and gene annotation

Three kinds of tissues, including leaves, young stem, and flower of sample NN, were obtained for RNA extraction. The total RNA was extracted using the RNAprep Pure Plant Plus Kit (Tiangen, China) according to the manufacturer's instruction. RNA quality and quantity was determined by the 2100 Bioanalyzer (Agilent Technologies, USA), NanoDrop Spectrophotometers (Thermo Fisher Scientific, USA), and Qubit fluorometer (Thermo Fisher Scientific, USA). Equal amount of total RNA of the three tissues was mixed. The RNA library was generated using the NEBNext® Ultra™ RNA Library Prep Kit (NEB, USA), following the kit's protocol. The mRNA, enriched by Oligo(dT) beads, was fragmented and primed with a random primer. The resulting RNA was reverse transcribed into cDNA by ProtoScript II Reverse Transcriptase (NEB, USA). Following the synthesis of second-strand cDNA, end-repair and adaptor ligation was performed. The ligation reaction was size-selected and purified with the AMPure XP system (Beckman Coulter, USA). The purified cDNA was PCR enriched using universal PCR primers and the Index primer. After reaction product purification, the resulting library was assessed using 2100 Bioanalyzer. According to the manufacturer's instructions, the library was sequenced on an Illumina NovaSeq 6000 Sequencing System (Illumina, USA). Quality control was processed to remove adaptor sequence and low-quality reads. The remaining clean data was assembled using the Trinity method to generate a *de novo* transcriptome assembly (Grabherr et al., 2011). The corset program was employed to cluster the assemblies (Davidson and Oshlack, 2014). The obtained longest contig was selected as unigene. The unigenes were annotated based on the following databases: the non-redundant protein database (NR), the nucleotide sequence database (NT), the PFAM database, the Swiss-Prot database, the Gene Ortholog database (GO), Kyoto Encyclopedia of Genes and Genomes Ortholog (KEGG) database, and the Clusters of Eukaryotic Ortholog Groups of proteins database (KOG, NCBI).

EST-SSR identification and primers development

The annotated unigenes were analyzed using the MISA tool to identify the EST-SSRs (Beier et al., 2017). The minimum repeat number of mono-, di- tri-, tetra-, penta-, and hexanucleotide motifs were set as 10, 6, 5, 5, 5, and 5, respectively. The primers were designed using PRIMIER3 from the flanking region of the SSR (Rozen and Skaletsky, 2000). The parameters were set as follows: PCR product size ranges from 100 bp to 280 bp, primer length ranges from 18 bp to 27 bp, and annealing temperature ranges from 57°C to 63°C.

DNA extraction and SSR validation

The *R. tomentosa* genomic DNA was extracted using TaKaRa MiniBEST Plant Genomic DNA Extraction Kit (TaKaRa, Dalian, China). The procedures of extraction followed the manufacturer's instructions. One hundred pairs of primers were selected to validate the developed EST-SSRs. These primers could amplify various kinds of SSR motifs. Four samples, including CZ, WZ-1, BS, and GL, were chosen as amplification templates. The PCR reaction solution included one μl of template DNA (50 ng/μl), 12.5 μl of 2 × Tsingke master mix (Tsingke, China), one μl of each primer (10 μM), and 9.5 μl of ddH_2O. The amplification procedure was performed on a Biometra EasyCycler (Analytik Jena, Germany), applying the program as follows: initial denaturation step at 94°C for 5 min, 35 cycles of the cycling steps (30 sec at 94°C, 30 sec at annealing temperature, 30 sec at 72°C), and final extension step at 72°C for 5 min. The PCR products were analyzed in 8% denaturing polyacrylamide gel electrophoresis (PAGE). The polymorphic primers would be used further to study the genetic diversity of the natural *R. tomentosa*.

Genetic analysis of *R. tomentosa*

A total of 13 polymorphic primers (Table 2) were used to analyze the genetic diversity of the 16 natural samples using capillary electrophoresis. The primers were labeled with FAM fluorescent dye at the 5'-end of the upstream primer. The PCR products were analyzed on an ABI 3730xl DNA Analyzer (Thermo Fisher Scientific, USA) using GeneScan 500 LIZ as size standard (Thermo

Fisher Scientific, USA). Data analysis was conducted in GeneMapper software (Thermo Fisher Scientific, USA) using the Microsatellite Default analysis method. POPGENE version 1.32 (Yeh et al., 1999) was employed to calculate the observed number of alleles (*Na*), number of effective alleles (*Ne*), fixation index (*Fst*) value, and Shannon's information index (*I*). The phylogenetic tree of the 13 natural *R. tomentosa* was constructed by NTSYSpc version 2.2 software using the UPGMA method.

Results

Transcriptome sequencing and assembly

A total of 94,289,248 raw reads were generated from the *R. tomentosa* cDNA library. After the quality control process, 94,152,862 clean reads with 98.6% of Q20 and 95.67% of Q30 were obtained. The total clean data was 14.12 Giga nucleotides bases with 48.73% of GC content (Table 3). The clean data of the *R. tomentosa* transcriptome sequence was deposited NCBI (National Centre for Biotechnology Information) SRA (Sequence Read Archive) database (Accession: PRJNA821925).

Through de novo assembly, 137,256 transcripts were obtained. The total length of the transcripts was 153,737,542 bp, with a mean length of 1,120 bp and a median length of 623 bp. All the transcripts were further clustered into 51,486 unigenes, of which the length ranged from 201 bp to 15,869 bp (Figure 1). The mean and median length of unigenes were 1173 bp and 769 bp, respectively. The total length of unigenes was 153,737,542 bp (Table 3).

Functional annotation of unigenes

To obtain the corresponding annotation information , the unigenes were blasted against the seven databases, including Nr, Nt, PFAM, Swiss-Prot, GO, KEGG, and KOG. Of the 51,486 unigenes (Supplementary Table 1), 36,979 were annotated in at least one database, while 4,571 unigenes were annotated in all seven databases (Table 4). Of the 31,635 unigenes, which were annotated in Nr, 25,351 hits (80.1%) were from *Eucalyptus grandis* (Myrtales, Myrtaceae), while the second most hit species was *Vitis vinifera* (315, 1.0%). The 10,014 hits annotated in the KOG database were assigned to the 25 functional groups (Figure 2a). Posttranslational modification, protein turnover, chaperones is the largest group (1,299, 12.6%), followed by the general function prediction (1,288, 11.5%) and the Signal transduction mechanisms (879, 7.8%). A total of 27,492 unigenes were assigned to the GO database. These unigenes were classified into three major categories: Cellular Component (49,751), Biological Process (71,444), and Molecular Function (33,411), in which cell part (9,149), cellular process (15,991), and binding (15,896) were the largest subcategories, respectively (Figure 2b). 12,223 unigenes were matched in the KEGG pathway database. Among the five main categories, metabolism was the largest category with 5,405 unigenes, followed by genetic information processing (2,549 unigenes) and organismal systems (471 unigenes) (Figure 3).

SSR motifs identification and classification

Using MISA software, a total of 18,879 SSRs were identified in 14,132 unigenes, of which 3,541 unigenes contained more than one SSR. The number of SSR presented in compound microsatellites was 1,187. The occurrence frequency of SSR was 0.37 SSR per unigene and one SSR every 3.20 kb. In summary, mono-nucleotide repeat was the most common repeat type (8,126, 43.04%), followed by di- (5,846, 31.06%) and tri- (4,610, 24.42%) nucleotide repeat (Table 5). In number of repeats, ten-tandem repeat (3,549, 18.80%) was the most abundant type, followed by six-tandem (3,114, 16.49%) and five tandem (2,789, 14.77%) repeat (Table 3). Of all the SSR, only two kinds of motifs were higher than 10% in abundance, which were A/T (7,816, 41.40%) and AG/CT (5,002, 26.50%). The richest trinucleotide repeat was AAG/CTT (934, 4.95%). In the quad-, penta-, or hexanucleotide repeats, AAAT/ATTT (34, 0.18%), AAACT/AGTTT (2, 0.01%) and AAAGCC/CTTTGG (2, 0.01%) were the most dominant motifs, respectively. The distribution of all the 89 kinds of motifs was listed in Supplementary Table 2.

SSRs validation and evaluation

Using Primer 3 software, 11,726 SSRs were eligible for the criteria of designing flanking primers (Supplementary Table 3). Using polyacrylamide gel electrophoresis, we randomly selected one hundred primer pairs for primer efficiency evaluation (Supplementary Table 4). Four *R. tomentosa* individuals collected from four natural populations, which were far distant from each

other, were used as DNA templates. Of the 100 primers, 50 pairs of primers generated clear corresponding PCR products (Supplementary Figure 1). A total of 23 pairs of primers were found polymorphic among the 50 successful primers.

Application of developed SSRs in natural *R. tomentosa* population

The 13 polymorphic primers, with good reproducibility and clear main bands, were selected to assess the 16 natural individual *R. tomentosa* plants (Table 2). A total of 32 alleles were detected, with the mean alleles of 2.462 per locus. The number of alleles ranged from 2 to 3. The observed heterozygosity (*Ho*) ranged from 0.000 to 0.563 with a mean value of 0.240, while the expected heterozygosity (*He*) ranged from 0.121 to 0.667, with a mean value of 0.414. Of the fixation index (*Fst*) value, the range was from 0.105 to 0.800, with a mean of 0.413. The Shannon information index (I) ranged from 0.234 to 1.066 with a mean of 0.641. Similarity based clustering was used to construct a UPGMA dendrogram (Figure 4). The 16 individuals were clustered into five groups, which were found to be related to the geographical locations of plants. The South group included QZ, BH-1, BH-2, BH-3, and BH-4, all collected from the south region of Guangxi province. The West group included two samples (BS and NN) from the two bordering west cities, Baise and Nanning. Samples from Laibin (LB), Guigang (GG), and Guilin (GL) were clustered into the North-Central group due to these three samples were distributed from the central to the north. Two samples collected from Wuzhou (WZ-1 and WZ-2), an east city, were clustered into the East group. The three samples from Pingxiang (PX) and Chongzuo (CZ-1 and CZ-2) were clustered into the out branch of the dendrogram. All the samples were able to distinguish according to their geographical origin, except one sample of the south region, BH-5, which was clustered into the North-Central group.

Discussion

RNA-seq technology has been increasingly applied to analyze the plant's transcriptome and EST-SSR, especially the plant without reference genome (Zhang et al., 2017a;Hina et al., 2020). With the features of codominant, inheritance, high polymorphism, and high repeatability, SSR marker has been widely used in genetic diversity analysis of wild plant populations (Hwang et al., 2008;Hammami et al., 2014;Nandha and Singh, 2014). Based on transcriptome sequences, the EST-SSR has been proved useful in functional varieties detection and gene-associated genetic analysis (Zheng et al., 2013;Yan et al., 2015). In spite of the wide distribution and clinical application (Sianglum et al., 2012;Huang et al., 2019;Wang et al., 2022), there is a deficiency of comprehensive SSR markers in *R. tomentosa*. In this study, we employed an RNA-seq technology to characterize the transcriptome of *R. tomentosa*. Also, we firstly reported the distribution and validation of EST-SSRs in this plant.

Compared to other species, like *Neolitsea sericea* (Chen et al., 2015), *Elymus sibiricus* (Zhou et al., 2016), and *Rhododendron rex* (Zhang et al., 2017a), the 94,152,862 clean reads with 98.6% of Q20 level we obtained in *R. tomentosa* was sufficient to meet the quality of functional annotation and EST-SSR mining. A total of 51,486 unigenes, with a mean length of 1,173 bp and an N50 length of 1783 bp, were acquired in our study (Table 3, Fig. 1), while in a previous transcriptome report of *R. tomentosa* (He et al., 2018), the number of unigenes was 83,175 with the mean length of 888 bp and the N50 length of 1,702 bp (Table 3). The less unigenes number and longer length of our study may be due to the smaller number of samples and the uses of different Illumina sequencing platforms. Nevertheless, the mean length of unigene found in *R. tomentosa* was longer than the transcriptome studies of some plants, like *Magnolia wufengensis* (899 bp), easter lily (628 bp), sugarcane (1147 bp) (Deng et al., 2019;Howlader et al., 2020;Malviya et al., 2020). The longer assembled unigene was advantageous in annotating unigenes and exploring SSR (Patnaik et al., 2016). Overall, this study's transcriptome sequences would complement gene annotation and provide a comprehensive resource for SSR development in *R. tomentosa*.

A total of 36,979 unigenes (71.82%) were annotated in the seven public databases (Nr, Nt, PFAM, Swiss-Prot, GO, KEGG, and KOG) (Table 4). In various reports of transcriptome in Myrtaceae family, the annotation rates of unigenes ranged from 64.61% (*R. tomentosa*) to 77.53% (*Syzygium longifolium*) (He et al., 2018;Soewarto et al., 2019). These results agreed with the 71.82% unigenes annotation rate in our study. The blast hits distribution in the Nr database, which included 31,635 unigenes, showed the most hits with *Eucalyptus grandis* (25,351, 80.1%). Considering the well-annotated genome of *E. grandis* was the first reported genome in Myrtales, no wonder the majority of the blast hits of *R. tomentosa* belonged to *E. grandis* (Myburg et al., 2014). There were 10,014 unigenes in the KOG annotation in present study, which were classified into 25 functional groups (Fig. 2A). The top three groups were the posttranslational modification, protein turnover, chaperones (1,299), general function prediction only (1,288), and signal transduction mechanisms (879), which were partly similar to the reports in other plants (Zhang et al., 2017a;Hina et al.,

2020). Of the GO analysis, 27,492 unigenes were associated with 924 GO terms, classified into 50 subcategories in three main categories (Fig. 2B). The number of subcategories varied in different plants, such as *Menispermum canadense* with 54 subcategories (Hina et al., 2020) and *Pistacia vera* with 56 (Karcı et al., 2020). The KEGG analysis could help understand the high-level function and the utilities of the biological system (Kanehisa and Goto, 2000). He et al. (2018) employed GC-MS and transcriptome to identify the molecular basis of terpenes biosynthesis in *R. tomentosa*, in which KEGG analysis revealed there were 439 unigenes belonging to the terpenoid and polyketide metabolism pathway. Terpenoid was an effective medical metabolite in *R. tomentosa* (Zhang et al., 2017b;Deng et al., 2020). In this study, the terpenoid and polyketide metabolism pathway involved 291 unigenes, while the carbohydrate metabolism pathway contained the largest number of 1084 unigenes (Fig. 3). The overall KEGG analysis results of our study were consistent with the previous reports (He et al., 2018).

EST-SSRs are suitable for assaying the functional and phenotypic diversity in plant populations (Varshney et al., 2005a), identification of diverse parents (Parthiban et al., 2018), structure assessment in species conservation (Chen et al., 2015;Xing et al., 2017). Using the 51,486 unigenes sequences, we identified a total of 18,879 SSRs, with the SSR occurrence frequency of one SSR every 3.20 kb. The SSR occurrence frequency in our study was higher than *Rhododendron rex* (5.65 kb) (Zhang et al., 2017a) and Camellia sinensis (4.99 kb) (Wu et al., 2013), but lower than *Pistacia vera* (2.03 kb) (Karcı et al., 2020) and *Rhododendron latoucheae* (2.87 kb) (Xing et al., 2017). The variant frequency is related to the species differences, different SSR analysis methods, and database size variation (Biswas et al., 2012). The most abundant SSR type was the mononucleotide repeats (43.04%), followed by the dinucleotide (31.06%) and trinucleotide (24.42%) repeats (Table 5). Consistent with the other two Rosids dicotyledonous plants, *Arabidopsis thaliana* (Lawson and Zhang, 2006) and *Pistacia vera* (Karcı et al., 2020), the mono-nucleotide type was the most prevalent type of SSR in *R. tomentosa*. The most abundant SSR motif was A/T (41.4%), followed by AG/CT (26.5%) and AAG/CTT (4.95%) (Supplementary Table 2). The AG/CT motif could represent GAG, AGA, UCU, and CUC codons, which would translate into the amino acids Arg, Glu, Ala, and Leu, respectively (Chen et al., 2015). The two amino acids, Ala and Leu, have a relatively higher frequency than other amino acids (Kantety et al., 2002;Qiu et al., 2010). This could explain the high frequencies of AG/CT motif in transcriptome sequences. According to a previous report, some SSR motifs have been found correlating with gene expression by influencing transcription initiation or methylation of CpG (Li et al., 2004). The length of an SSR influenced the expressional level of *LvIRF*, which shows that the shorter AG/CT repeat led to a stronger expression of *LvIRF* (Yin et al., 2019). Of the tri-nucleotide SSRs, AAG/CTT motif was significantly prominent in dicotyledonous plants, which included cucumber (Cavagnaro et al., 2010), castor bean (Qiu et al., 2010), sesame (Wei et al., 2011), *Neolitsea sericea* (Chen et al., 2015), and *R. rex* (Zhang et al., 2017a).

In addition to SSR development, sequences obtained from transcriptome were appropriate to design specific primers (Wei et al., 2011). In the present study, 100 pairs of primers were randomly selected to evaluate the effectiveness and polymorphism of the 11,726 pairs of primers. Using four *R. tomentosa* DNA, 50 pairs of primers (50%) produced the expected size of PCR products, in which 23 pairs of primers (23%) were found polymorphic. The polymorphic rate of our study was higher than *N. sericea* (9.92%) (Chen et al., 2015) but lower than *E. sibiricus* (30.35%) (Zhou et al., 2016), *Elaeis guineensis* (55.29%) (Zhou et al., 2020). Compared to these reports, the polymorphic rate of *R. tomentosa* in our study was moderate. Nevertheless, the polymorphic level of SSR may be affected by the amount and genetic diversity of selected samples (Wu et al., 2014;Zhou et al., 2016;Zhang et al., 2017a).

In this study, 13 pairs of primers, which produced expected PCR products, were used to detect the 16 natural *R. tomentosa* individuals (Table 1). A total of 32 alleles were obtained (Table 2). The number of alleles per locus ranged from 2 to 3, with a mean number of 2.462. Compared to the ISSR method (Yao, 2010;Hue et al., 2015), the SSR method used in our study produced a smaller number of alleles in *R. tomentosa*. Given the differences between SSR and ISSR methods, it's normal to have lesser alleles in SSR within the same species (Sabreena et al., 2021). Among the 13 markers, the observed heterozygosity (*Ho*) and expected heterozygosity (*He*) varied from 0.000 to 0.563 and 0.121 to 0.667, with a mean value of 0.240 and 0.414, respectively (Table 2). The two heterozygosity value was lower than *R. rex* (*Ho* = 0.323, *He* = 0.372) (Zhang et al., 2017a), *Elymus sibiricus* (Ho = 0.49, He = 0.59) (Zhou et al., 2016) and *Sesamum indicum L.* (*Ho* = 0.84, *He* = 0.76) (Wei et al., 2011). The lower heterozygosity indicated that *R. tomentosa* maintains relatively low genetic variability in Guangxi region. The gene diversity values reported herein can guide selecting the loci that are most likely to be informative in further *R. tomentosa* research.

As mentioned above, *R. tomentosa* is native to southern and southeastern Asia, and widespread throughout the tropical and subtropical regions. In Guangxi province, even from distant habitats, the phenotypic characteristics of wild *R. tomentosa* are similar. To analyze the phylogenetic relationship between the 16 natural *R. tomentosa* individuals, we constructed a UPGMA dendrogram using similarity based clustering (Fig. 4). Five groups were identified within 16 individuals using the 13 markers. The South group and West group were clustered into a major branch of the dendrogram, which is consistent with the geographical connection of these two regions. The North-Central and East group were clustered into another major branch, indicating that these two groups have a closer genetic relationship. The five groups were highly related to their geographical locations. The only exception sample, BH-5, located in the south of Guangxi, was clustered into the North-Central group. Considering the fruit of *R. tomentosa* was one of the food for some birds, the dispersal of this plant could not only be geographical (Aslan and Rejmánek, 2010). In general, the 13 markers were able to distinguish *R. tomentosa* individuals broadly according to their geographical origin. Therefore, these newly developed EST-SSRs primers can be useful for genetic variability analysis in *R. tomentosa*.

Conclusion

In this study, we used transcriptome sequencing to mine the expressed sequences of *R. tomentosa*. A total of 51,486 unigenes were obtained, of which 36,979 unigenes were annotated. Based on the unigene sequences, 18,879 EST-SSRs were identified and classified. A total of 11,726 pairs of EST-SSRs primer were developed. One hundred pairs of primer were randomly selected to validate the efficiency and polymorphism of these markers. 50% of the primers produced were proved effective, and 23 pairs of primer were polymorphic. Using 13 pairs of polymorphic primer, 16 natural *R. tomentosa* individuals were distinguished according to their geographical origin. These results presented a valuable sequence resource and provided a powerful tool for genetic relationship analysis in *R. tomentosa*.

Declarations

Acknowledgment

We thank Guangxi Pfomic Bioinformation Co., Ltd. for their help in transcriptome analysis.

Funding

This work was supported by Guangxi Key Laboratory of Special Non-wood Forest Cultivation & Utilization (Grant numbers JA-20-01-02 and JA-20-01-05).

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Lina Sun, Qi Chen, and Mao Lin conceived and designed the experiments. Lina Sun, Jinhua Li, Kaidao Sun, Huaxin Wang, and Kaitai Yang collected the samples. Lina Sun, Jinhua Li, and Qi Chen performed the transcriptome sequencing and data analysis. Lina Sun, Kaidao Sun, Huaxin Wang, Kaitai Yang contributed equally to the EST-SSR validation. Lina Sun wrote the paper. Qi Chen and Mao Lin revised the draft manuscript. Mao Lin supervised the whole project.

Data Availability

The datasets generated in this study are available in BioProject with the accession number of PRJNA821925 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA821925&o=acc_s%3Aa).

References

Ali, A., Pan, Y.-B., Wang, Q.-N., Wang, J.-D., Chen, J.-L., and Gao, S.-J. (2019) Genetic diversity and population structure analysis of Saccharum and Erianthus genera using microsatellite (SSR) markers. *Scientific Reports* 9**:** 395, https://doi.org/10.1038/s41598-

018-36630-7.

Aslan, C.E., and Rejmánek, M. (2010) Avian use of introduced plants: Ornithologist records illuminate interspecific associations and research needs. *Ecological Applications* 20**:** 1005-1020, https://doi.org/10.1890/08-2128.1.

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33**:** 2583-2585, https://doi.org/10.1093/bioinformatics/btx198.

Biswas, M.K., Chai, L., Mayer, C., Xu, Q., Guo, W., and Deng, X. (2012) Exploiting BAC-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in Citrus. *Molecular Biology Reports* 39**:** 5373-5386, https://doi.org/10.1007/s11033-011-1338-5.

Cai, K., Zhu, L., Zhang, K., Li, L., Zhao, Z., Zeng, W., and Lin, X. (2019) Development and Characterization of EST-SSR Markers From RNA-Seq Data in Phyllostachys violascens. *Frontiers in Plant Science* 10**:** https://www.frontiersin.org/article/10.3389/fpls.2019.00050.

Cavagnaro, P.F., Senalik, D.A., Yang, L., Simon, P.W., Harkins, T.T., Kodira, C.D., et al. (2010) Genome-wide characterization of simple sequence repeats in cucumber (Cucumis sativus L.). *BMC Genomics* 11**:** 569, https://doi.org/10.1186/1471-2164-11-569.

Chen, L.-Y., Cao, Y.-N., Yuan, N., Nakamura, K., Wang, G.-M., and Qiu, Y.-X. (2015) Characterization of transcriptome and development of novel EST-SSR makers based on next-generation sequencing technology in Neolitsea sericea (Lauraceae) endemic to East Asian land-bridge islands. *Molecular Breeding* 35**:** 187, https://doi.org/10.1007/s11032-015-0379-1.

Cordeiro, G.M., Casu, R., Mcintyre, C.L., Manners, J.M., and Henry, R.J. (2001) Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci* 160**:** 1115-1123,

Davidson, N.M., and Oshlack, A. (2014) Corset: enabling differential gene expression analysis for de novoassembled transcriptomes. *Genome Biology* 15**:** 410, https://doi.org/10.1186/s13059-014-0410-6.

Deng, S., Ma, J., Zhang, L., Chen, F., Sang, Z., Jia, Z., and Ma, L. (2019) De novo transcriptome sequencing and gene expression profiling of Magnolia wufengensis in response to cold stress. *BMC Plant Biology* 19**:** 321, https://doi.org/10.1186/s12870-019-1933-5.

Deng, X., Wang, X.-R., and Wu, L. (2020) Triketone-terpene meroterpenoids from the leaves of Rhodomyrtus tomentosa. *Fitoterapia* 143**:** 104585, https://www.sciencedirect.com/science/article/pii/S0367326X20301672.

Durand, J., Bodénès, C., Chancerel, E., Frigerio, J.-M., Vendramin, G., Sebastiani, F., et al. (2010) A fast and cost-effective approach to develop and map EST-SSR markers: oak as a case study. *BMC Genomics* 11**:** 570, https://doi.org/10.1186/1471-2164-11-570.

Feng, S., He, R., Lu, J., Jiang, M., Shen, X., Jiang, Y., et al. (2016) Development of SSR Markers and Assessment of Genetic Diversity in Medicinal Chrysanthemum morifolium Cultivars. *Frontiers in Genetics* 7**:** https://www.frontiersin.org/article/10.3389/fgene.2016.00113.

Freire, C.G., Giachini, A.J., Gardin, J.P.P., Rodrigues, A.C., Vieira, R.L., Baratto, C.M., et al. (2018) First record of in vitro formation of ectomycorrhizae in Psidium cattleianum Sabine, a native Myrtaceae of the Brazilian Atlantic Forest. *PLoS One* 13**:** e0196984,

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29**:** 644-652, https://doi.org/10.1038/nbt.1883.

Guo, R., Mao, Y.-R., Cai, J.-R., Wang, J.-Y., Wu, J., and Qiu, Y.-X. (2014) Characterization and cross-species transferability of EST–SSR markers developed from the transcriptome of Dysosma versipellis (Berberidaceae) and their application to population genetic studies. *Molecular Breeding* 34**:** 1733-1746, https://doi.org/10.1007/s11032-014-0134-z.

Hammami, R., Jouve, N., Soler, C., Frieiro, E., and González, J.M. (2014) Genetic diversity of SSR and ISSR markers in wild populations of Brachypodium distachyon and its close relatives B. stacei and B. hybridum (Poaceae). *Plant Systematics and*

Evolution 300: 2029-2040, https://doi.org/10.1007/s00606-014-1021-0.

Han, Z., Ma, X., Wei, M., Zhao, T., Zhan, R., and Chen, W. (2018) SSR marker development and intraspecific genetic divergence exploration of Chrysanthemum indicum based on transcriptome analysis. *BMC Genomics* 19: 291, https://doi.org/10.1186/s12864-018-4702-1.

He, S.-M., Wang, X., Yang, S.-C., Dong, Y., Zhao, Q.-M., Yang, J.-L., et al. (2018) De novo Transcriptome Characterization of Rhodomyrtus tomentosa Leaves and Identification of Genes Involved in α/β-Pinene and β-Caryophyllene Biosynthesis. *Frontiers in Plant Science* 9: https://www.frontiersin.org/article/10.3389/fpls.2018.01231.

Hina, F., Yisilam, G., Wang, S., Li, P., and Fu, C. (2020) De novo Transcriptome Assembly, Gene Annotation and SSR Marker Development in the Moon Seed Genus Menispermum (Menispermaceae). *Frontiers in Genetics* 11: https://www.frontiersin.org/article/10.3389/fgene.2020.00380.

Howlader, J., Robin, A.H.K., Natarajan, S., Biswas, M.K., Sumi, K.R., Song, C.Y., et al. (2020) Transcriptome Analysis by RNA–Seq Reveals Genes Related to Plant Height in Two Sets of Parent-hybrid Combinations in Easter lily (Lilium longiflorum). *Scientific Reports* 10: 9082, https://doi.org/10.1038/s41598-020-65909-x.

Huang, Y., Yang, Z., Huang, S., An, W., Li, J., and Zheng, X. (2019) Comprehensive Analysis of Rhodomyrtus tomentosa Chloroplast Genome. *Plants (Basel, Switzerland)* 8: 89, https://pubmed.ncbi.nlm.nih.gov/30987338

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6524380/.

Hue, T.S., Abdullah, T.L., Abdullah, N.A., and Sinniah, U.R. (2015) Genetic variation in Rhodomyrtus tomentosa (Kemunting) populations from Malaysia as revealed by inter-simple sequence repeat markers. *Genet Mol Res* 14: 16827-16839,

Hwang, T.-Y., Nakamoto, Y., Kono, I., Enoki, H., Funatsuki, H., Kitamura, K., and Ishimoto, M. (2008) Genetic diversity of cultivated and wild soybeans including Japanese elite cultivars as revealed by length polymorphism of SSR markers. *Breeding Science* 58: 315-323,

Jeong, D., Yang, W.S., Yang, Y., Nam, G., Kim, J.H., Yoon, D.H., et al. (2013) In vitro and in vivo anti-inflammatory effect of Rhodomyrtus tomentosa methanol extract. *J Ethnopharmacol* 146: 205-213,

Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28: 27-30, https://pubmed.ncbi.nlm.nih.gov/10592173

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/.

Kantety, R.V., La Rota, M., Matthews, D.E., and Sorrells, M.E. (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol* 48: 501-510,

Karcı, H., Paizila, A., Topçu, H., Ilikçioğlu, E., and Kafkas, S. (2020) Transcriptome Sequencing and Development of Novel Genic SSR Markers From Pistacia vera L. *Frontiers in genetics* 11: 1021-1021, https://pubmed.ncbi.nlm.nih.gov/33033493

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7509152/.

Lai, T.N., André, C., Rogez, H., Mignolet, E., Nguyen, T.B., and Larondelle, Y. (2015) Nutritional composition and antioxidant properties of the sim fruit (Rhodomyrtus tomentosa). *Food Chem* 168: 410-416,

Lawson, M.J., and Zhang, L. (2006) Distinct patterns of SSR distribution in the Arabidopsis thalianaand rice genomes. *Genome Biology* 7: R14, https://doi.org/10.1186/gb-2006-7-2-r14.

Li, Y.-C., Korol, A.B., Fahima, T., and Nevo, E. (2004) Microsatellites Within Genes: Structure, Function, and Evolution. *Molecular Biology and Evolution* 21: 991-1007, https://doi.org/10.1093/molbev/msh073.

Limsuwan, S., Hesseling-Meinders, A., Voravuthikunchai, S.P., Van Dijl, J.M., and Kayser, O. (2011) Potential antibiotic and anti-infective effects of rhodomyrtone from Rhodomyrtus tomentosa (Aiton) Hassk. on Streptococcus pyogenes as revealed by proteomics. *Phytomedicine* 18**:** 934-940,

Malviya, M.K., Li, C.-N., Solanki, M.K., Singh, R.K., Htun, R., Singh, P., et al. (2020) Comparative analysis of sugarcane root transcriptome in response to the plant growth-promoting Burkholderia anthina MYSP113. *PLOS ONE* 15**:** e0231206, https://doi.org/10.1371/journal.pone.0231206.

Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., et al. (2014) The genome of Eucalyptus grandis. *Nature* 510**:** 356-362, https://doi.org/10.1038/nature13308.

Nandha, P.S., and Singh, J. (2014) Comparative assessment of genetic diversity between wild and cultivated barley using gSSR and EST-SSR markers. *Plant Breeding* 133**:** 28-35, https://onlinelibrary.wiley.com/doi/abs/10.1111/pbr.12118.

Parthiban, S., Govindaraj, P., and Senthilkumar, S. (2018) Comparison of relative efficiency of genomic SSR and EST-SSR markers in estimating genetic diversity in sugarcane. *3 Biotech* 8**:** 144, https://doi.org/10.1007/s13205-018-1172-8.

Patnaik, B.B., Wang, T.H., Kang, S.W., Hwang, H.-J., Park, S.Y., Park, E.B., et al. (2016) Sequencing, De Novo Assembly, and Annotation of the Transcriptome of the Endangered Freshwater Pearl Bivalve, Cristaria plicata, Provides Novel Insights into Functional Genes and Marker Discovery. *PLOS ONE* 11**:** e0148622, https://doi.org/10.1371/journal.pone.0148622.

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

Ramanatha Rao, V., and Hodgkin, T. (2002) Genetic diversity and conservation and utilization of plant genetic resources. *Plant Cell, Tissue and Organ Culture* 68**:** 1-19, https://doi.org/10.1023/A:1013359015812.

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

Sabreena, Nazir, M., Mahajan, R., Hashim, M.J., Iqbal, J., Alyemeni, M.N., et al. (2021) Deciphering allelic variability and population structure in buckwheat: An analogy between the efficiency of ISSR and SSR markers. *Saudi Journal of Biological Sciences* 28**:** 6050-6056, https://www.sciencedirect.com/science/article/pii/S1319562X21006434.

Sianglum, W., Srimanote, P., Taylor, P.W., Rosado, H., and Voravuthikunchai, S.P. (2012) Transcriptome Analysis of Responses to Rhodomyrtone in Methicillin-Resistant Staphylococcus aureus. *PLOS ONE* 7**:** e45744, https://doi.org/10.1371/journal.pone.0045744.

Soewarto, J., Hamelin, C., Bocs, S., Mournet, P., Vignes, H., Berger, A., et al. (2019) Transcriptome data from three endemic Myrtaceae species from New Caledonia displaying contrasting responses to myrtle rust (Austropuccinia psidii). *Data in Brief* 22**:** 794-811, https://www.sciencedirect.com/science/article/pii/S2352340918316202.

Tayeh, M., Nilwarangoon, S., Mahabusarakum, W., and Watanapokasin, R. (2017) Anti-metastatic effect of rhodomyrtone from Rhodomyrtus tomentosa on human skin cancer cells. *Int J Oncol* 50**:** 1035-1043,

Tuler, A.C., Carrijo, T.T., Nóia, L.R., Ferreira, A., Peixoto, A.L., and Da Silva Ferreira, M.F. (2015) SSR markers: a tool for species identification in Psidium (Myrtaceae). *Mol Biol Rep* 42**:** 1501-1513,

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

Varshney, R.K., Sigmund, R., Börner, A., Korzun, V., Stein, N., Sorrells, M.E., et al. (2005b) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Science* 168**:** 195-202, https://www.sciencedirect.com/science/article/pii/S0168945204003656.

Vidya, V., Prasath, D., Snigdha, M., Gobu, R., Sona, C., and Maiti, C.S. (2021) Development of EST-SSR markers based on transcriptome and its validation in ginger (Zingiber officinale Rosc.). *PLOS ONE* 16**:** e0259146, https://doi.org/10.1371/journal.pone.0259146.

Wang, R., Yao, L., Lin, X., Hu, X., and Wang, L. (2022) Exploring the potential mechanism of Rhodomyrtus tomentosa (Ait.) Hassk fruit phenolic rich extract on ameliorating nonalcoholic fatty liver disease by integration of transcriptomics and metabolomics profiling. *Food Research International* 151**:** 110824, https://www.sciencedirect.com/science/article/pii/S0963996921007249.

Wei, W., Qi, X., Wang, L., Zhang, Y., Hua, W., Li, D., et al. (2011) Characterization of the sesame (Sesamum indicum L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics* 12**:** 451, https://doi.org/10.1186/1471-2164-12-451.

Wu, H., Chen, D., Li, J., Yu, B., Qiao, X., Huang, H., and He, Y. (2013) De Novo Characterization of Leaf Transcriptome Using 454 Sequencing and Development of EST-SSR Markers in Tea (Camellia sinensis). *Plant Molecular Biology Reporter* 31**:** 524-538, https://doi.org/10.1007/s11105-012-0519-2.

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

Wu, P., Ma, G., Li, N., Deng, Q., Yin, Y., and Huang, R. (2015) Investigation of in vitro and in vivo antioxidant activities of flavonoids rich extract from the berries of Rhodomyrtus tomentosa(Ait.) Hassk. *Food Chem* 173**:** 194-202,

Wu, Q., Zang, F., Ma, Y., Zheng, Y., and Zang, D. (2020) Analysis of genetic diversity and population structure in endangered Populus wulianensis based on 18 newly developed EST-SSR markers. *Global Ecology and Conservation* 24**:** e01329, https://www.sciencedirect.com/science/article/pii/S2351989420308702.

Xie, C., Huang, B., Jim, C.Y., Han, W., and Liu, D. (2021) Predicting differential habitat suitability of Rhodomyrtus tomentosa under current and future climate scenarios in China. *Forest Ecology and Management* 501**:** 119696, https://www.sciencedirect.com/science/article/pii/S0378112721007866.

Xing, W., Liao, J., Cai, M., Xia, Q., Liu, Y., Zeng, W., and Jin, X. (2017) De novo assembly of transcriptome from Rhododendron latoucheae Franch. using Illumina sequencing and development of new EST-SSR markers for genetic diversity analysis in Rhododendron. *Tree Genetics & Genomes* 13**:** 53, https://doi.org/10.1007/s11295-017-1135-y.

Yan, X., Zhang, X., Lu, M., He, Y., and An, H. (2015) De novo sequencing analysis of the Rosa roxburghii fruit transcriptome reveals putative ascorbate biosynthetic genes and EST-SSR markers. *Gene* 561**:** 54-62, https://www.sciencedirect.com/science/article/pii/S0378111915002048.

Yao, X. (2010). *Mating system and genetic diversity of Rhodomyrtus tomentosa (Myrtaceae) detected by ISSR markers.* Master of Philosophy, The University of Hong Kong.

Yeh, F., Yang, R., and Boyle, T. (1999) POPGENE version 1.32: Microsoft Windows–based freeware for population genetic analysis, quick user guide. *Center for International Forestry Research, University of Alberta, Edmonton, Alberta, Canada* 1-29,

Yin, B., Wang, H., Zhu, P., Weng, S., He, J., and Li, C. (2019) A Polymorphic (CT)n-SSR Influences the Activity of the Litopenaeus vannamei IRF Gene Implicated in Viral Resistance. *Frontiers in Genetics* 10**:** https://www.frontiersin.org/article/10.3389/fgene.2019.01257.

Zhang, Y., Zhang, X., Wang, Y.-H., and Shen, S.-K. (2017a) De Novo Assembly of Transcriptome and Development of Novel EST-SSR Markers in Rhododendron rex Lévl. through Illumina Sequencing. *Frontiers in Plant Science* 8**:** https://www.frontiersin.org/article/10.3389/fpls.2017.01664.

Zhang, Y.-L., Zhou, X.-W., Wu, L., Wang, X.-B., Yang, M.-H., Luo, J., et al. (2017b) Isolation, Structure Elucidation, and Absolute Configuration of Syncarpic Acid-Conjugated Terpenoids from Rhodomyrtus tomentosa. *Journal of Natural Products* 80**:** 989-

998, https://doi.org/10.1021/acs.jnatprod.6b01005.

Zhao, Z., Wu, L., Xie, J., Feng, Y., Tian, J., He, X., et al. (2020) Rhodomyrtus tomentosa (Aiton.): A review of phytochemistry, pharmacology and industrial applications research progress. *Food Chem* 309**:** 125715,

Zheng, X., Pan, C., Diao, Y., You, Y., Yang, C., and Hu, Z. (2013) Development of microsatellite markers by transcriptome sequencing in two species of Amorphophallus (Araceae). *BMC Genomics* 14**:** 490, https://doi.org/10.1186/1471-2164-14-490.

Zhou, L., Yarra, R., Zhao, Z., Jin, L., and Cao, H. (2020) Development of SSR markers based on transcriptome data and association mapping analysis for fruit shell thickness associated traits in oil palm (Elaeis guineensis Jacq.). *3 Biotech* 10**:** 280-280, https://pubmed.ncbi.nlm.nih.gov/32537380

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7261811/.

Zhou, Q., Luo, D., Ma, L., Xie, W., Wang, Y., Wang, Y., and Liu, Z. (2016) Development and cross-species transferability of EST-SSR markers in Siberian wildrye (Elymus sibiricus L.) using Illumina sequencing. *Scientific Reports* 6**:** 20549, https://doi.org/10.1038/srep20549.

Tables

Table 1. Description of *Rhodomyrtus tomentosa* samples used for genetic analysis with EST-SSRs

Sample codes	Locations	Geographical coordinates		Region orientation
		Longitude	Latitude	
PX	Shangsi, Pinxiang	106.77	22.038	Southwest
CZ-1	Ningming, Chongzuo	106.95	21.93	Southwest
CZ-2	Ningming, Chongzuo	106.98	21.96	Southwest
BH-1	Shankou, Beihai	109.73	21.63	South
BH-2	Guantou hill, Beihai	109.09	21.38	South
BH-3	Guantou hill,Beihai	109.06	21.45	South
BH-4	Donghuan, Beihai	109.26	21.62	South
BH-5	Donghuan, Beihai	109.22	21.64	South
QZ	Nali, Qinzhou	108.84	21.84	South
BS	Forest Institute, Basise	106.62	23.97	West
NN	Wumin, Nanning	107.99	23.25	West
LB	Jinxiu, Laibin	110.20	24.14	Central
GL	Lingchuan, Guilin	110.43	25.52	North
GG	Pingnan, Guigang	110.18	23.69	Central
WZ-1	Changzhou, Wuzhou	111.20	23.50	East
WZ-2	Cangwu, Wuzhou	111.07	23.56	East

Table 2 The polymorphic primers used to analyze the natural *R. tomentosa* samples

Locus	SSR Motif	Primer sequence	Tm	Size (bp)	Na[a]	Ho[b]	He[c]	Fst[d]	I[e]
c4073_g1_i0	(GTTCG)5	GTCATCCCCTCCTCCTCCTT	60.03	246-253	2	0.188	0.417	0.800	0.594
		GCCGACTCGATGCTCATACA	59.97						
c4296_g0_i0	(CTCA)6	GCAACGCAGAAACAAGGAGG	60.04	246-266	2	0.438	0.498	0.316	0.676
		GTGATAGTGACGGCGAGAGG	59.97						
c7589_g0_i0	(CAAGCA)5	AGGGAAATACGTGAAGCGGG	60.11	197-209	2	0.000	0.508	0.630	0.685
		TTTCCTCATTCCTCGGCGAC	60.11						
c7023_g1_i0	(TCGGC)5	GCGGACAATCTGATTTGGGC	59.90	192-217	3	0.188	0.417	0.586	0.728
		CTCTCACGTCCCTATGTGGC	59.90						
c7600_g0_i0	(CTGGGT)5	GGAATCAGCCGAGATGGAGG	59.97	169-187	3	0.375	0.613	0.331	0.974
		TACGCGATTCCAAAACCCCT	59.67						
c11558_g0_i2	(AGTTG)5	AGATGTGAGCAGCTGGAACC	60.04	126-141	2	0.438	0.466	0.390	0.644
		TGACCTCCACCACTTCATGC	59.96						
c13120_g0_i0	(TTAGT)5	GCATCAGGGAGCGATTCTCA	59.89	278-293	2	0.188	0.353	0.716	0.525
		TCAGTCAGTACAAGCGTGGC	60.32						
c23286_g0_i0	(GGA)7	GCAAGAGGCGATCTCGTACA	59.90	213-222	3	0.563	0.667	0.145	1.066
		ACCACTATGAGCGCGTTCTC	60.18						
c23457_g0_i1	(CCAACG)6	GGACTGAAAAAGGACGCTGC	59.76	232-250	3	0.250	0.325	0.312	0.567
		TGGCACCGGAATCTTCTGAC	60.04						
c26367_g14_i0	(TTC)7	CCTCCTCTCAGTCTCTGCCT	60.03	144-153	3	0.125	0.179	0.164	0.371
		GGTTGATCTAGAGCGTCCGG	59.97						
c36092_g0_i0	(TCA)7	AGCAAACAACAAGTGTCGTGT	59.19	157-169	3	0.250	0.331	0.105	0.602
		CGAGCTTCATTGCCTGCATC	59.97						
c38567_g0_i2	(CGC)7	GTCCTCTCTCGCTCTGCATC	59.97	248-257	2	0.125	0.484	0.520	0.662
		GAGCGATGTTCTTCACGTGC	59.91						
c43257_g0_i0	(ATAC)6	TGGTGCTTCTCTCAAAATGGT	57.77	256-260	2	0.000	0.121	0.138	0.234
		TGTACTTCTTGGTCGTCGCC	60.04						
Mean					2.462	0.240	0.414	0.413	0.641

a, Na, number of alleles; b, Ho, observed heterozygosity; c, He, expected heterozygosity; d, I, Shannon's diversity index; e, Fst, fixation index.

Table 3 The summary of transcriptome sequencing

Category	Parameter	Value
Raw data	No. of total raw reads	94,289,248
	No. of total clean reads	94,152,862
	No. of nucleotides of clean data	14.12 G
	Q20 percentage	98.6%
	Q30 percentage	95.67%
	GC content	48.73%
Assemblies	Number of transcripts	137,256
	Total length of transcripts (bp)	153,737,542
	Mean length of transcripts (bp)	1,120
	Median length of transcripts (bp)	623
	Max length of transcripts (bp)	15,869
	N50 value of transcripts	2,040
Unigenes	Number of unigenes	51,486
	Total length of unigenes (bp)	60,393,577
	Mean length of unigenes (bp)	1,173
	Median length of unigenes (bp)	769
	Max length of unigenes (bp)	15,869
	N50 value of unigenes	1,783

Table 4 The summary of functional annotation of unigenes

Database	Number of annotated genes	Percentage of annotated genes (%)
Nr	31,635	61.44
Nt	27,273	52.97
PFAM	23,896	46.41
Swiss-Prot	25,254	49.05
GO	27,492	53.39
KEGG	12,223	23.74
KOG	10,014	19.44
Annotated in all databases	4,571	8.87
Annotated in at least one database	36,979	71.82
Total unigenes	51,486	100

Table 5 The distribution of the SSRs based on repeat and tandem number

Number of repeats	Mono-	Di-	Tri-	Quad-	Penta-	Hexa-	Total	Percentage (%)
5	-	-	2556	189	37	7	2789	14.77%
6	-	1713	1361	32	1	7	3114	16.49%
7	-	1235	661	1	-	3	1900	10.06%
8	-	1079	28	-	-	1	1108	5.87%
9	-	1131	1	-	-	-	1132	6.00%
10	2943	605	1	-	-	-	3549	18.80%
11	1434	99	1	-	-	-	1534	8.13%
12	942	1	1	-	-	1	945	5.01%
13	683	-	-	-	-	-	683	3.62%
14	594	-	-	-	-	-	594	3.15%
15	483	-	-	-	-	-	483	2.56%
16	375	-	-	-	-	-	375	1.99%
17	237	-	-	-	-	-	237	1.26%
18	170	-	-	-	-	-	170	0.90%
19	112	-	-	-	-	-	112	0.59%
20	86	-	-	-	-	-	86	0.46%
21	44	-	-	-	-	-	44	0.23%
22	18	-	-	-	-	-	18	0.10%
23	5	-	-	-	-	-	5	0.03%
26	-	1	-	-	-	-	1	0.01%
Total	8126	5864	4610	222	38	19	18879	
Percentage (%)	43.04%	31.06%	24.42%	1.18%	0.20%	0.10%		

Figures

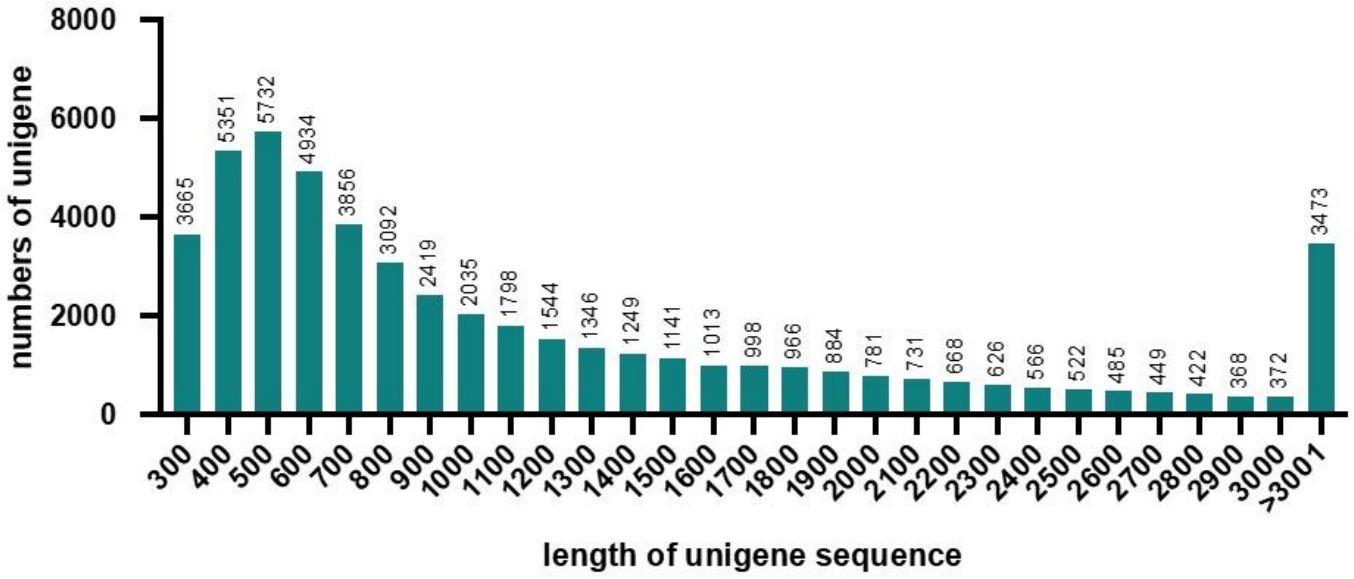

Figure 1

Length distribution of all unigene sequences in *R. tomentosa*.

The x-axis represents the length of unigenes, and the y-axis represents the number of unigenes with a certain length.

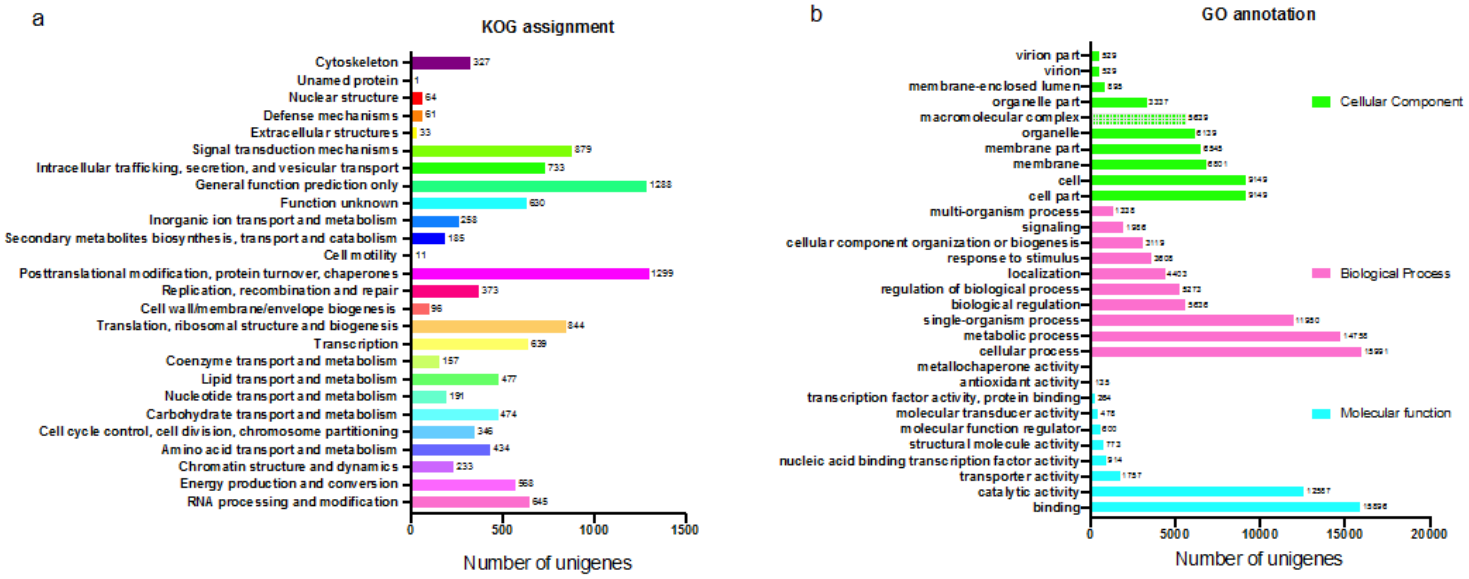

Figure 2

KOG assignment and GO annotation of the unigene sequences of *R. tomentosa*.

a The functional classification of the *R. tomentosa* unigenes according to KOG criteria. **b** The distribution of *R. tomentosa* unigenes among the GO functional annotation classes. The x-axis indicates the number of unigenes in a specific functional cluster. The y-axis indicates the function class.

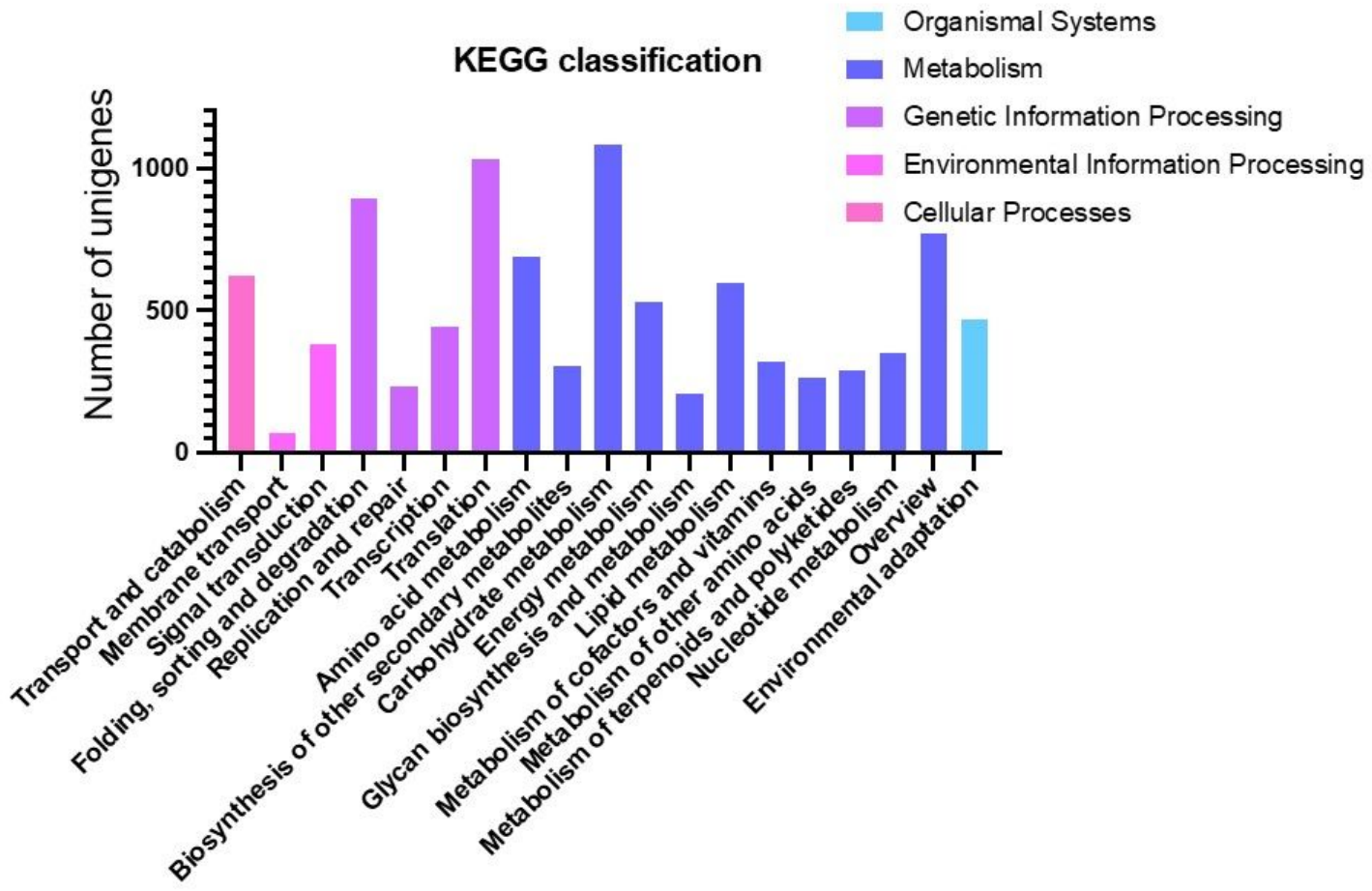

Figure 3

KEGG classification of the unigene sequences of *R. tomentosa*.

The x-axis represents the KEGG pathway classifications of unigenes. The five main categories were appeared in different colors. The y-axis represents the number of unigenes with a certain pathway.

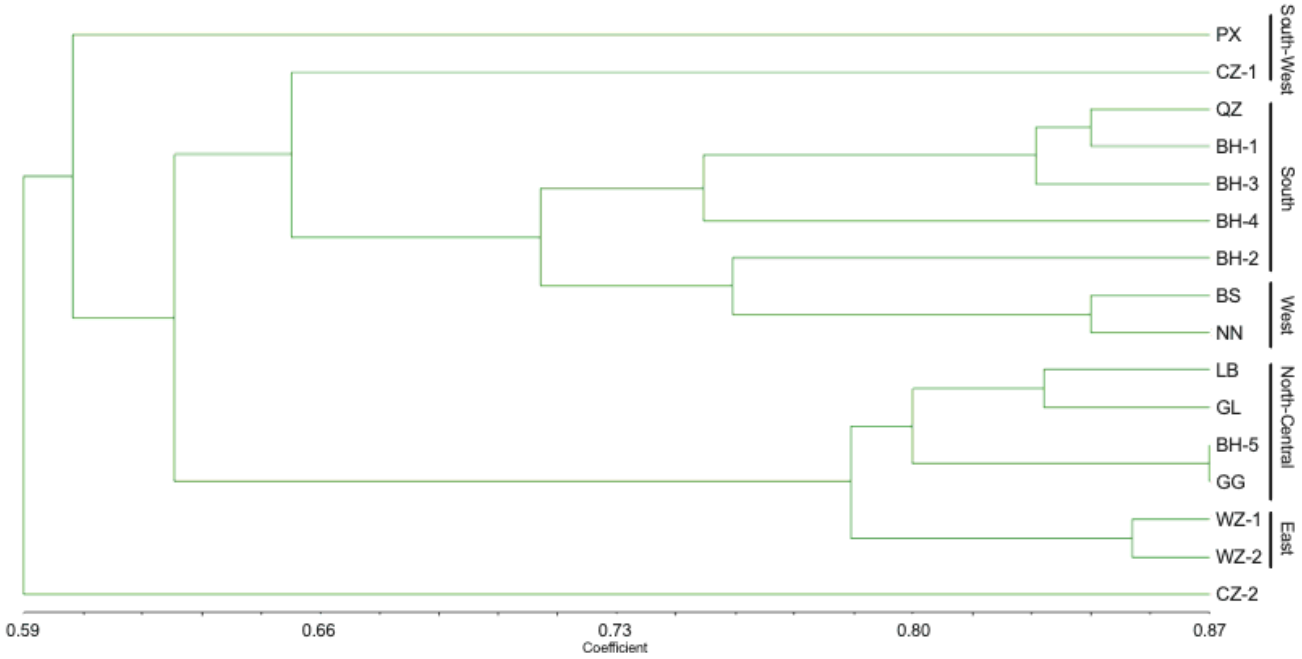

Figure 4

UPGMA dendrogram of 16 *R. tomentosa* individuals based on similarity coefficient using 13 SSR loci.

Supplementary Files

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

- SupplementaryFigure1.docx
- SupplementaryTable1.xlsx
- Supplementarytable2.docx
- Supplementarytable3.xlsx
- Supplementarytable4.xlsx