Establishment of novel SSR markers from transcriptome data of Chimonanthus praecox and application in the variety identification

Bin Liu
Southwest University
Huafeng Wu
Southwest University
Yinzhu Cao
Southwest University
Ximeng Yang
Southwest University
Ting Zhu
Southwest University
Shunzhao Sui
sszcq@swu.edu.cn
Southwest University

Research Article

Keywords: Chimonanthus praecox, SSR, RNA-seq, varieties, identification

Posted Date: April 4th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4167345/v1

Additional Declarations: No competing interests reported.

Abstract

Background *Chimonanthus praecox*, belongs to the Calycanthaceae, is a unique traditional famous flower and special economic tree species in China. There are numerous varieties but only a few cultivars were named. At present, EST-SSR markers are widely used to identify different species and varieties, as researchers can identify a large number of microsatellites from transcriptome databases.

Result A total of 162,638 unigenes were assembled by using RNA-seq, and 82,778 unigenes was annotated by Nr, Nt, Swiss-Prot, Pfam, GO, KOG and KEGG databases. A total of 13,556 SSR loci were detected from 11,691 unigenes, with trinucleotide repeat motifs being the most abundant among the six types of repeat motifs. In order to develop markers, 64,440 pairs of SSR primers with polymorphism potential were designed, and 75 pairs of primers were randomly selected for amplification. Among them, seven pairs amplified fragments of the expected size with high polymorphism, and twelve *C.praecox*varieties were clustered into two monophyletic clades by the seven EST-SSR markers.

Conclusion The microsatellites in the transcriptome of *C.preacox* have the advantages of rich types, strong specificity, and great polymorphism potential. These EST-SSR markers can provide molecular technical methods for identifying different varieties of *C.preacox*, and can also explore a large number of candidate genes associated to traits.

Background

Chimonanthus praecox commonly known as wintersweet (2n = 22), which is one of most popular perennial ornamental plants in China because of unique flowering time and long blooming period (from November to March) and has over a thousand-year history of cultivation [1, 2]. It is native to China and widely used for cut flowers and as a garden plant, and also has been introduced to the United States, Japan, South Korea and other countries [3, 4]. In traditional Chinese medicine research, it has been found that *C.praecox* can not only detoxify but also be used to treat diseases such as cough, dizziness, nausea, fever and rheumatoid arthritis [3, 5, 6]. *C.praecox* has multiple cultivated species, many of which are named and identified based on morphological characteristics such as petal color or morphology [7-9]. However, due to the limited number of morphological features and their susceptibility to environmental influences, using morphological features to study genetic relationships and identify phylogenetic relationships may be limited [9].

Molecular markers can reveal genetic relationships at the level of genetic material which have advantages such as being unaffected by the environment, high heritability, and easy detection [10]. A variety of molecular markers have extensively been used in *C.praecox* source conservation and genetic breeding, including random amplified of polymorphic DNA (RAPD) [11, 12], amplified restriction fragment length polymorphism (AFLP) [13, 14], sequence-related amplified polymorphism (SRAP), inter simple sequence repeat (ISSR), and simple sequence repeat (SSR), etc [4, 9, 11-19]. SSR markers belong to co dominant markers, mainly using tandem repeat sequences of 2 to 5 nucleotides as basic repeating units. They can distinguish homozygotes and heterozygotes, detect multiple alleles, and have advantages such as rich polymorphism, easy operation, reliable results, and good repeatability. Therefore, they are usually considered as the preferred marker [20]. SSR markers, also known as microsatellites, can be developed from genomic and transcriptome databases, so they are divided into gSSR and EST-SSR based on the type of data used for development. The cost of gSSR is relatively high, EST-SSR markers are relatively cheap, and EST-SSR has higher cross species transferability due to being developed from more conservative coding regions. [21]. And compared to using traditional methods to develop SSR markers, researchers using high-throughput sequencing technology can develop a large number of microsatellites with lower cost and effort [22]. As a consequence, using the SSR markers will be an efficient way for the identification of each *C.praecox* germplasms from the molecular DNA levels to genotype *C.praecox* cultivars.

In this study, we sequenced the transcriptome of *C.praecox* using BGIseq500 platform and assembled 162,638 unigenes, then identified SSR loci, designed primer pairs based on these data and seven novel EST-SSR markers were developed and characterized in the end. This study developed effective EST-SSR markers from transcriptome sequences to investigate the diversity of different varieties of C. praecox and applied them to variety classification.

Methods

Plant materials and DNA/RNA extraction

The 12 varieties plant materials of *C.praecox* for transcriptome sequencing and identifying polymorphisms were collected from the resource nursery which belong to the Key Laboratory of Agricultural Biosafety and Green Production of Upper Yangtze River (Ministry of Education) of Southwest University in Beibei District, Chongqing in 2023 (Table S1). Fresh leaf tissues were cleaned and immediately preserved in liquid nitrogen until DNA and RNA extraction. Extract total genomic DNA from leaves using CTAB method [23]. Select two varieties for RNA extraction and use the RNArep Pure kit (Tiangen Biotechnology, Beijing, China) to extract total RNA. To ensure the quality and quantity of DNA/RNA, 1% agarose gel electrophoresis was used to observe the DNA/RNA extract, and Nano Drop ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, MA, USA) was used for quantitative detection.

Transcriptome sequencing and assembly

Total RNA samples of acceptable purity and concentration were obtained, and library construction was then performed, mRNA was enriched by oligo (dT) -attached magnetic beads, and the purified mRNA was fragmented. First-strand cDNA was synthesized by using reverse transcriptase, and double-stranded cDNA was synthesized using the first-strand cDNA as a template. Then subjected to end-repair the double-stranded cDNA fragments, and added a single 'A' nucleotide to the 3' ends of the blunt fragments, adaptor ligation is subsequently configured and set up to ligate adaptors with the cDNAs. Obtain total RNA samples with acceptable purity and concentration, then construct the library, enrich messenger RNA with oligo (dT) -attached magnetic beads, and fragment the purified messenger RNA. Using reverse transcriptase method to synthesize first-stranded cDNA, and using first stranded cDNA as a template to synthesize double-stranded cDNA. Amplify the final library using phi29 to create DNA nanoballs (DNBs) with over 300 copies of molecules, and check the quality of library construction. Load DNBs into a patterned nanoarray and generate a counter terminal 100 base reading on the BGIseq500 platform (BGIseq500, Shenzhen, China). Measure in triplicate.

Raw data analysis and function annotation

The raw data was filtered with SOAPnuke (v1.5.2) [24] by Removing reads containing adapters (adapter contamination) first, and then remove reads with an unknown base ('N 'base) ratio greater than 10%, remove reads with a low-quality base ratio (base mass less than or equal to 15) greater than 50%, and obtain clean reads stored in fasta format. Gene Annotation: Annotating assembled unigenes with seven major functional databases including KEGG (Kyoto Encyclopedia of Genes and Genomes), GO (Gene Ontology), NR (National Center for Biotechnology Information nonredundant protein sequences), NT (Nucleotide Sequence Database), SwissProt (Swiss-Prot Sequence Database), Pfam (Protein Families Database), and KOG (EuKaryotic Orthologous Groups of proteins), and predicting the transcription factors [25-31].

Microsatellites identification and data analysis

Use MISA [32] to detect microsatellite loci according to the following criteria: mono nucleotide repeat motif repeat count greater than or equal to 20, di nucleotide repeat motif repeat count greater than or equal to 10, and other types of repeat motif repeat count greater than or equal to 5. By using Primer3 [33] software to design primers, 75 pairs of primers with target product sizes between 100-300bp were randomly selected.
Twelve *C.praecox* varieties were amplified to survey the polymorphism of the SSR loci. PCR products were visualized via electrophoresis in 8% polyacrylamide gels electrophoresis, and selected SSR to amplify the expected size product to evaluate polymorphism. Put the products into gel in 1×TBE (Tris-borate-EDTA) buffer solution, and run it for 1.5 hours at 200V, while using a 2000bp molecular size ladder (Tiangen Biotech Co., Ltd., Beijing, China), then observe the bands using silver staining.

For SSR data analysis, manually scoring based on allele size follows the principle that if there are no bands, the data is interpreted as "0", and if there are bands, it is interpreted as "1". The genetic information such as the number of alleles (Na), the effective number of alleles (Ne), Shannon's information index (I), and the Fixed index (F) of each locus were calculated in GenALEX 6.5 [34]. UPMGA cluster analysis was conducted using the NTSYSpc program [35].

Results

Transcriptome sequencing and assembly

A total of 114.73 Gb of clean data were obtained, and a total of 162,638 unigenes were assembled. The total length of unigenes is 170,847,856 bp, the average length of unigenes is 1,050 bp, the GC contents were 40.98%, and the N50 of the unigenes was 2,059 bp, indicating a high-quality assembly (Table 1). Among them, there were 80,351 (58.74%) unigenes with a length of 200-1000 bp, 29,042 (21.2%) unigenes with a length of 1-2kb, 16,654 (12.2%) unigenes with a length of 2-3kb, and 5,449 (4.0%) unigenes with a length exceeding 3kb (Fig. 1).

Table 1 Overall data quality and assembly information

Item	Number
Total clean data (Gb)	114.73
Total unigenes	162,638
Total length of unigenes (bp)	170,847,856
Average length of unigenes (bp)	1,050
N50 of unigenes (bp)	2,059
GC contents	40.98%

Functional annotation

In order to annotate the unigenes of *C.praecox*, 162638 single gene sequences were searched in different universal databases. 77,914 (47.91%) unigenes were aligned to sequences in the Nr database (Fig. 3B), 55,460 (34.10%) in the Nt database, 55,465 (34.10%) in the Swiss-Prot database, and 57,638 (35.44%) in the Pfam database. The annotation of 82,778 (50.90%) unigenes was achieved in at least one database, and the annotation

of 24,879 (15.30%) unigenes was achieved in all database (Table 2 and Fig. 2). A total of 62,480 (38.42%) unigenes were aligned to sequences in the GO database, which could be divided to three functional categories: biological processes, cellular components and molecular functions (Fig. 3A). In “biological process”, the largest classes were “cellular process” (41,021, 25.22%) followed by “metabolic process” (32,947, 20.26%) and “biological regulation” (9,619, 5.91%). The largest class in biological processes is "cellular processes"(41,021, 25.22%), followed by "metabolic processes" (32,947, 20.26%) and "biological regulation"(9,619, 5.91%). The categories of “cellular component” only include "cellular anatomical entity"(60,574, 37.24%) and "protein-containing complex"(7,534, 4.63%). Among the molecular functional categories, the largest category is "binding"(46,184, 28.40%), followed by "catalytic activity"(40,653, 25.00%) and "transporter activity"(4,150, 2.55%). A total of 47,180 (29.01%) unigenes were aligned to sequences in the KOG database, which were categorized into 25 functional groups (Fig. 3C), among them, 12,024 (7.39%) were annotated as “general function prediction only”, followed by “signal transduction mechanisms” (5,297, 3.26%), and “posttranslational modification, protein turnover, chaperones” (4,421, 2.72%). A total of 56,185 (34.55%) unigenes were aligned to sequences in the KEGG database, which could be categorized into five groups including cellular processes, environmental information processing, genetic information processing, metabolism, and organismal systems (Fig. 3D). Among the 19 biological pathways, the most frequently observed functional pathways were “Global and overview maps” (14,053, 8.64%), followed by “Carbohydrate metabolism” (5,259, 3.23%), and “Folding, sorting and degradation” (4,330, 2.66%).

Table 2 Unigenes annotation of *C.praecox*

Annotation database	Number of unigenes	Percentage(%)
NR	77,914	47.91%
NT	55,460	34.10%
Swissprot	55,465	34.10%
KEGG	57,525	35.37%
KOG	47,180	29.01%
Pfam	57,638	35.44%
GO	62,480	38.42%
Intersection	24,879	15.30%
Overall	82,778	50.90%
Total	162,638	100.00%

Frequency and distribution of SSRs in the transcriptome

Using the MISA software, 13,556 unigenes which the total length is 170,847,856 bp were selected from 162,638 Unigenes, 1,515 unigenes containing more than 1 SSR, and 11,691 SSR loci were detected in the end (Table 3). Six types of microsatellites were identified from transcriptome data, including mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeat motifs, with significant differences observed among different types of repeat motifs, which the trinucleotide repeats exhibited the highest frequency of occurrence (7,984, 58.90%), followed by dinucleotides (4,613, 34.03%), tetranucleotides (355, 2.62%), hexanucleotides (260, 1.92%), mononucleotides (222, 1.64%), and pentanucleotides (122, 0.90%) (Table 3). The most frequent dinucleotide repeats were AG/CT (4,053), accounting for 29.90% of the total SSRs. Of the trinucleotide repeats, AAG/CTT (3,117, 22.99%) was the most abundant motif, followed by ATC/ATG (1,446, 10.67%) and AGC/CTG (1,021, 7.53%). The most abundant mononucleotide, tetranucleotide, pentanucleotide and hexanucleotide repeats were A/T (222, 1.64%), AAAT/ATTT (91, 0.67%), AAAGG/CCTTT or AGCCC/CTGGG (29,0.21%), and AAGAGG/CCTCTT (33, 0.24%) respectively. The quantities of different types of dinucleotides and trinucleotides are shown in Fig.4.

Table 3 Prediction of SSRs out of the transcript datasets of *C.praecox*

Item	Number
Total number of sequences examined	162,638
Total size of examined sequences (bp)	170,847,856
Total number of identified SSRs	13,556
Number of SSR containing sequences	11,691
Number of sequences containing more than 1 SSR	1,515
Number of SSRs present in compound formation	667
Mono nucleotide	222
Di nucleotide	4,613
Tri nucleotide	7,984
Tetra nucleotide	355
Penta nucleotide	122
Hexa nucleotide	260

Development of polymorphic EST-SSR markers

Seventy-five potential EST-SSR markers primers were selected designed and tested to validate polymorphism of EST-SSR markers in *C. praecox*, but twenty of them were not amplified, while 55 were successfully amplified, producing amplicons of the expected size. Out of 55 EST-SSR markers, seven showed high levels of polymorphism and good transferability in different varieties. The genetic variation analysis of seven loci showed total of 28 alleles, ranging from 2 to 6, with an average of 4 alleles per locus. Ne ranged from 1.492 to 4.235 and total was 20.61 with average was 2.944. I ranged from 0.512 to 1.585 with average was 1.122. Ho ranged from 0.250 to 1.000. He ranged from 0.330 to 0.764 with average was 0.603. The above results indicated that the seven EST-SSR markers had relatively high levels of genetic polymorphism (Table 4).

UPGMA cluster analysis different varieties of *C.praecox* based on the EST-SSR markers

A topology tree based on the unweighted pair-group method analysis (UPMGA) was used to display the relationship between 12 different varieties of *C.praecox* (Fig. 5). The r-value of the matrix correlation of the topology tree is 0.808, and the approximate value of the Mantel t-test is 6.13. According to the UPGMA cluster analysis, 12 varieties of *C.praecox* were clustered into two monophyletic clades, S1, S6, S5, S15, S14, S7, SHA were clustered in clade I, and S12, S17, SX, S16, S24 were clustered in clade II, indicating close genetic relationships.

Table 4 Sequence and genetic diversity information of seven SSR markers

Locus	Motif	Forward primer (5′- 3′)	Reverse primer (5′- 3′)	GenBank accession number	Na	Ne	I	Ho	He
CP14	(CTT)7	CGCTCTCTCCTTAACGCGAT	ACTTCTTGCTTTTGCCGCTG	PP532794	2.000	1.492	0.512	0.417	0.330
CP20	(TC)25	CCATCTGCGACTGTCCCTTT	CGGATCTCTCCCGGATTTCG	PP532795	4.000	3.646	1.332	0.500	0.726
CP22	(CT)18	AGAACATGTCCAATTCCCATGGA	GCATGCTCGCTCTCTCTCTC	PP532796	6.000	4.235	1.585	0.333	0.764
CP33	(AT)10	CAGTCAGGTCCACGTGTTGA	ATCTCGATCTGCTGCCACTG	PP532797	6.000	3.176	1.426	0.444	0.685
CP43	(GA)14	TGCCCAGTTGCCTCTTTTCA	CGACTTCTTCTCCTTCGCCA	PP532798	2.000	1.492	0.512	0.250	0.330
CP44	(TCG)7	CCGGAAGTAGCCATCGGATC	GCATGGAGAGTCCTCGCTAC	PP532799	3.000	2.969	1.093	0.750	0.663
CP67	(AG)22	CACGAAGCCCTCCAGAAAGT	CTTGCAGGGGAGCATGTACA	PP532800	5.000	3.600	1.393	1.000	0.722

Discussion

C.praecox as an ornamental plant has been cultivated for over a thousand years, it originated in China and introduced to South Korea around the 17th century, followed by Japan, Europe, the United States, and Australia [3, 4]. After a long history of cultivation, numerous *C.praecox* varieties were developed but only a few cultivars were named, and there are some homonymy and synonymy exist [8, 36]. During the cultivation, wrong naming of cultivars has led to difficult to distinguish the cultivars [37]. Molecular markers have been proved to be powerful tools for identifying and characterizing varieties, and some molecular markers have been used for variety identification. SSRs also known as microsatellites are one of the most important marker systems for plant genetic analysis, gene mapping, quantitative trait locus (QTL) mapping, and marker-assisted selection (MAS) breeding based on their high rates of mutation, universal distribution and high density in a multitude of genomes [38, 39]. Due to their

homology among related species in DNA coding regions and adequate polymorphisms provide large variations in DNA non-coding regions [38, 40], SSR markers have been specifically applied in a variety of identification procedures that includes the cultivars of a number of plants, such as Prunus persica [41], Morella rubra [42], Punica granatum [43], sympodial bamboo [44] and so more. Traditional SSR development method is a difficult, costly and labour intensively process, but the next-generation sequencing technology can effectively identify a large number of SSRs at a small cost and effort compared to traditional methods. Its main advantage is the ability to generate a large amount of sequence data, from which a large number of whole genome and gene-based SSR loci can be isolated and developed [39, 45]. As NGS techniques have become more advanced, different new methods of SSR marker development have been reported that can be grouped into genomic SSRs (gSSRs) distributed throughout the entire whole genome sequence and expression sequence tags SSRs (EST ssr) embedded in transcriptional sequences [46, 47]. Compared to gSSRs, EST-SSRs are more economical with efficiently amplification and highly transferable among plant species, and less susceptible to invalid alleles [10, 48]. Transcriptome sequencing is updated and efficient which not only helps to discovering new genes, expression pattern identification, but also helps to development of molecular markers [49]. In this study, a total of 162,638 unigenes were assembled，which the average length of unigenes is 1,050 bp, the N50 of the unigenes was 2,059 bp, indicating a high-quality assembly of transcriptome sequencing. The transcriptome data provide abundant resources of the SSR sites, which would be contribute to the identification and characterization of *C.praecox* varieties. And both in *C.praecox* business and research, our newly developed microsatellite markers will be useful in *C.praecox* species and cultivars discrimination and identification.

EST-SSRs have many known or putative functions and may be associated with the targeted trait which are useful for directing allele selection, detecting functional variation and analyzing gene associated genetic [50]. There are numerous evidence indicating that any changes including replication slippage and other mutational mechanisms of SSR repeats may lead to gaining and losing function or gene silencing and inducing novel protein, bacterial pathogenesis or virulence [51]. To obtain comprehensive function classification of unigenes in the *C.praecox* transcriptome data, we performed gene function annotations using the public databases Nr, Nt, Swiss-Prot, Pfam, GO, KOG and KEGG, results indicate that 50.90% unigenes was achieved functionally annotated in at least one database, and 15.30% unigenes was achieved functionally annotated in all database. And 62,480 (38.42%), 47,180 (29.01%), 56,185 (34.55%) unigenes were classified into GO, KOG and KEGG categories, the largest categories in GO, KOG, and KEGG are “cellular process”, “general function prediction only” and “Global and overview maps” which will be useful for development of functional EST–SSR markers. As next-generation DNA sequencing becomes faster and cheaper, a large amount of sequence data from different plant species is being generated exponentially, and more researcher are using transcriptome data to develop EST–SSR markers [52].

In this study, EST–SSR markers of *C.praecox* were developed using NGS technology, and we detected 13,556 EST-SSR loci distributed among 11,691 of a total of 162,638 unigenes. Among these EST-SSR loci, trinucleotide repeat motifs were the most abundant type, followed by dinucleotide repeat motifs, is different from the results of previous research [9, 53], which reported that the dinucleotide repeat motifs were the most abundant type, followed by the trinucleotide repeat motifs. But the situation of different results appear in *C.praecox* is same as in *Allium sativum*. Li et al. [21] reported the dinucleotide repeat motifs was the most abundant type was different from the results of Liu et al. [54] who reported that the trinucleotide repeat motifs were the most abundant type. And in many other plant studies have been found the same patterns, which the trinucleotide repeat is the most common motif of SSR motif abundance, like in *Elymus sibiricus* [55], *Pueraria thomsonii* [56]，*Dolichos bean* [57], *Elymus breviaristatus* [47] and 14 tree species [58].The differences in the previous finding may have been due to the SSR search criteria, the size of the dataset, and the database-mining tools [21].

Microsatellite markers have been widely used in species and cultivars identification, to check the effectiveness of newly developed EST-SSR markers, 75 pairs of SSR primers were randomly selected to genotype 12 *C.praecox* varieties, in which 66.7% markers successfully amplified target bands with 9.3% markers show highly polymorphic. And 33.7% markers did not amplify any fragments, possibly because the primers were designed across splice sites or large introns were present within the target amplicon [59]. Using cluster analysis, 12 varieties of *C.praecox* were clustered into two monophyletic clades, 7 varieties were clustered in clade I, and 5 varieties were clustered in clade II. In most cases the *C.praecox* cultivars are divided into three groups base on the color of inner tepals, including the Patens Group , the Intermedius Group, and the Concolor Group [2, 7, 60]. Two varieties of Intermedius Group and two varieties of Patens Group both could be classified into one group, the relationship between two concolor varieties is also relatively close in clade I. Two concolor varieties and one Intermedius varieties classified into one group in clade II, indicated a close relationship with them. The results show that EST-SSR markers distinguishes different varieties by inner tepals color to some extent. It means that these EST-SSR markers may be related to phenotype of flower color. Previous research has shown that SSR may be related to targeted traits and play important roles in development, gene regulation, evolution, and other aspects [61]. The results of this study demonstrated that phylogenetic analysis based on EST-SSR markers not only can provide valuable reference used in varieties identification, but also revealed a potential connection with the color of inner tepals, and this provides a theoretical basis for the breeding of new varieties of *C.praecox*.

Conclusions

A large number of SSR loci were identified using transcriptome data, and highly polymorphic microsatellite markers were developed and applied in the differentiation of *C.praecox* varieties, 12 varieties were detected into two monophyletic clades. The molecular markers developed in this study will contribute to the identification of *C. praecox* varieties, and these results also provide theoretical basis for conducting functional genomics, population genetics, and phylogenetic analysis of *C.praecox*.

Declarations

Supplementary Information

Table S1.

Acknowledgements

The authors are grateful to the Key Laboratory of Agricultural Biosafety and Green Production of Upper Yangtze River (Ministry of Education) for providing an experimental platform.

Authors' contributions

BL and SZS conceived the study and designed the experiments. BL, HFW and YZC performed the experiments. XMY and TZ helped with data analysis. The manuscript was written and revised by BL and HFW. All authors read and approved the final version of the manuscript.

Funding

This research was funded by the Special Key Project for Technological Innovation and Application Development in Chongqing (Grant No. CSTB2023TIAD-KPX0039).

Availability of data and materials

The clean data of RNA-seq generated in this study have been submitted to the BioProject database of the National Center for Biotechnology Information (PRJNA1091468).

Ethics approval and consent to participate

The authors complied with all relevant institutional, national and international guidelines.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Sui SZ, Luo JG, Ma J, Zhu QL, Lei X, Li MY. Generation and analysis of expressed sequence tags from *Chimonanthus praecox* (Wintersweet) flowers for discovering stress-responsive and floral development-related genes. International Journal of Genomics. 2012;2012.
2. Shen ZG, Li WY, Li YL, Liu ML, Cao HP, Provart N, et al. The red flower wintersweet genome provides insights into the evolution of magnoliids and the molecular mechanism for tepal color development. The Plant Journal. 2021;108(6):1662-78.
3. Shang J, Tian J, Cheng H, Yan Q, Chen L. The chromosome-level wintersweet (*Chimonanthus praecox*) genome provides insights into floral scent biosynthesis and flowering in winter. Genome Biology. 2020;21:200.
4. Fu XM, Yang N, Du YQ, Kamran HM, Wang HB, Chen SY, et al. Development of SSR molecular markers and genetic diversity analysis of TPS gene family in *Chimonanthus praecox*. Agriculture. 2023;13(4):893.
5. Kitagawa N, Ninomiya K, Okugawa S, Motai C, Nakanishi Y, Yoshikawa M, et al. Quantitative determination of principal alkaloid and flavonoid constituents in wintersweet, the flower buds of *Chimonanthus praecox*. Natural Product Communications. 2016;11(7):953-6.
6. Wu HF, Wang X, Cao YZ, Zhang HY, Hua R, Liu HM, et al. *CpBBX19*, a B-box transcription factor gene of *Chimonanthus praecox*, improves salt and drought tolerance in *Arabidopsis*. Genes. 2021;12(9):1456.
7. Yao CH, Wang CY. Three basic problems in the classification of *Chimonanthus praecox* varieties. Journal of Beijing Forestry University. 1995;17(S1):164-7.
8. Zhu T, Feng YY, Dong XX, Yang XM, Liu B, Yuan PY, et al. Optimizing DUS testing for *Chimonanthus praecox* using feature selection based on a genetic algorithm. Frontiers in Plant Science. 2024;14:1328603.
9. Liu DF, Ma J, Yang JF, Nguyen TV, Liu HM, Huang RW, et al. Mining simple sequence repeat and single nucleotide polymorphism markers in a transcriptomic database of wintersweet (*Chimonanthus praecox*). HortScience. 2014;49(11):1360-4.
10. Yang WL, Bai ZY, Wang FQ, Zou MZ, Wang XR, Xie JK, et al. Analysis of the genetic diversity and population structure of *Monochasma savatieri* Franch. ex Maxim using novel EST-SSR markers. BMC genomics. 2022;23:597.

11. Chen LQ, Chen JY, Zhen YL, Lu DF. Detection of Genetic Variation Within and Between Natural Populations of *Chimonanthus praecox*(L.)Link Using RAPD Markers. Journal of Beijing Forestry University. 1999;21(02):86-90.
12. Wang Q, Yao QJ, Xu ZL, Hu JG, Yang CS. Genetic diversityof four populations of Calycanthus chinensis based on ISR and RAPD markers. Guihaia. 2013;33(01):30-4.
13. Chen DW, Chen LQ. The first intraspecific genetic linkage maps of wintersweet [*Chimonanthus praecox* (L.) Link] based on AFLP and ISSR markers. Scientia Horticulturae. 2010;124(1):88-94.
14. Zhao MX. Establishment of in vitro plant regeneration and ssr and aflp reaction system of *chimonanthus praecox* Link. var. concolor [master]2012.
15. Zuo DD, Zhao HT, Liu C, Mu D, Wang XW, Min J. Genetic Diversity in Natural Populations of *Chimonanthus praecox* (L.) Link Revealed by SRAP Markers. Acta Horticulturae Sinica. 2009;36(08):1197-202.
16. Bu HF, Gu ZH, Zhang WD, Li D. Establishment of SRAP-PCR reaction system for *chimonanthus praecox*. Agriculture and Technology. 2022;42(16):129-32.
17. Hu H, Chai N, Zhu Hx, Li R, Huang Rw, Wang X, et al. Factors affecting vegetative propagation of wintersweet (*Chimonanthus praecox*) by softwood cuttings. HortScience. 2020;55(11):1853-60.
18. Zhao B, Zhang QX. Genetic diversity of germplasm resources of *Chimonanthus praecox*(L.)Link based on AFLP marker. Acta Ecologica Sinica. 2007;27(11):4452-9.
19. WANG Xj, ZHAO Yb, WANG J, LI Zh, ZHANG Jt, LI Qw. Genetic diversity analysis and fingerprinting of 175 *Chimonanthus praecox* germplasm based on SSR molecular marker. Chinese Journal of Biotechnology. 2024;40(01):252-68.
20. Yang Qw, Jiang Yj, Wang Yp, Han Rl, Liang Zs, He Ql, et al. SSR loci analysis in transcriptome and molecular Marker development in *Polygonatum sibiricum*. BioMed Research International. 2022.
21. Li Xx, Qiao Lj, Chen Br, Zheng Yj, Zhi Cc, Zhang Sy, et al. SSR markers development and their application in genetic diversity evaluation of garlic (*Allium sativum*) germplasm. Plant Diversity. 2022;44(5):481-91.
22. Ning HJ, Gui FF, Tian EW, Yang LY. The novel developed microsatellite markers revealed potential hybridization among Cymbidium species and the interspecies sub-division of *C. goeringii* and *C. ensifolium*. BMC Plant Biology. 2023;23:492.
23. Siegel C, Stevenson F, Zimmer E. Evaluation and comparison of FTA card and CTAB DNA extraction methods for non-agricultural taxa. Applications in plant sciences. 2017;5(2):1600109.
24. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience. 2018(1).
25. Hiroyuki O, Susumu G, Kazushige S, Wataru F, Hidemasa B, Minoru K. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 1999;27(1):29-34.
26. Michael A, Catherine AB, Judith AB, David B, Heather B, J MC, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 2000;25(1):25-9.
27. Deng Y, Li J, Wu S, Zhu Y, Chen YW, He F. Integrated nr database in protein annotation system and its localization. 2006.
28. Tamara K, Ruth A, Philippe A, Nicola A, Mikael A, Alastair B, et al. EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Research. 2007;35(Database issue):16-20.
29. Amos B, Rolf A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research. 2000;28(1):45-8.
30. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, et al. The Pfam protein families database. Nucleic Acids Research. 2008;32(Database issue):138-41.
31. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome biology. 2004;5:1-28.
32. Sebastian B, Thomas T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics (Oxford, England). 2017(3):53.
33. Rozen S. Primer3 : a software component for picking PCR primer. 1996.
34. Peakall R, Smouse PE. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics. 2012;28(19):2537-9.
35. Rohlf FJ. NTSYS-pc 2.0. Numerical Taxonomy and Multivariate Analysis System. NTSYS-Pc : Numerical Taxonomy and Multivariate Analysis System. 2000;2.1.
36. Zhao K-G, Zhou M-Q, Chen L-Q, Zhang D, Robert GW. Genetic diversity and discrimination of *Chimonanthus praecox* (L.) Link germplasm using ISSR and RAPD markers. HortScience. 2007;42(5):1144-8.
37. Li D, Long C, Pang X, Ning D, Wu T, Dong M, et al. The newly developed genomic-SSR markers uncover the genetic characteristics and relationships of olive accessions. PeerJ. 2020;8:e8573.

38. Yang J, Zhang J, Han R, Zhang F, Mao A, Luo J, et al. Target SSR-Seq: A Novel SSR Genotyping Technology Associate With Perfect SSRs in Genetic Analysis of *Cucumber* Varieties. other. 2019;10.
39. Wu F, Zhang S, Gao Q, Liu F, Wang J, Wang X. Genetic diversity and population structure analysis in a large collection of *Vicia amoena* in China with newly developed SSR markers. BMC Plant Biology. 2021;21:544.
40. Chai M, Ye H, Wang Z, Sun G, Wang Y. Genetic divergence and relationship among Opisthopappus species identified by development of EST-SSR markers. Frontiers in Genetics. 2020;11.
41. Yin JW, Han BB, Ma YX, Wu XY, Xu ZJ, Jiang JF, et al. Analysis of genetic diversity of 154 peach cultivars based on SSR markers. Jiangsu Agricultural Sciences. 2023;51(16):18-26.
42. Ju PJ, Zhu YF, Zhao L, Wang GY, Zhou CC, Yan LJ, et al. Construction of Specific Fluorescent-labeled SSR Marker Database of Chinese Bayberry (*Morella rubra*) Varieties. Journal of Agricultural Biotechnology. 2023;31(10):2209-20.
43. Wang R, Zhong Yh, Hong Wj, Luo H, Li Dl, Zhao L, et al. Genetic diversity evaluation and core collection construction of pomegranate (*Punica granatum* L.) using genomic SSR markers. Scientia Horticulturae. 2023;319:112192.
44. Yuan Jl, Ma Jx, Zhong Yb, Yue Jj. SSR-based hybrid identification, genetic analyses and fingerprint development of hybridization progenies from sympodial bamboo (Bambusoideae, Poaceae). Journal of Nanjing Forestry University (Natural Sciences Edition). 2021;45(05):10-8.
45. Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, et al. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American Journal of Botany. 2012;99(2):193-208.
46. Wang LY, Li SS, Wang TY, He CY, Luo HM, Zhang JG, et al. Genomic SSR and EST-SSR markers for phylogenetic and pedigree reconstructions-A comparison in sea buckthorn. Plant Breeding. 2021;140:167-83.
47. Li J, Zhang Cb, Chen Sy, Jiang Kk, Guan H, Liu Wh. Characterization and application of EST-SSR markers developed from transcriptome sequences in *Elymus breviaristatus* (Poaceae: Triticeae). Genes. 2023;14:302.
48. Wu J, Cai C, Cheng F, Cui H, Zhou H. Characterisation and development of EST-SSR markers in tree peony using transcriptome sequences. Molecular Breeding. 2014;34(4):1853-66.
49. Tulsani NJ, Hamid R, Jacob F, Umretiya NG, Nandha AK, Tomar RS, et al. Transcriptome landscaping for gene mining and SSR marker development in Coriander (*Coriandrum sativum* L.). Genomics. 2020;112(2):1545-53.
50. Zheng X, Pan C, Diao Y, You Y, Yang C, Hu Z. Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). BMC genomics. 2013;14:490.
51. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Molecular biology and evolution. 2004;21(6):991-1007.
52. Bárbara RB, Carvalho LMD, Carazzolle MF, Pereira GAG, Colombo CA. Development of novel EST-SSR markers in the macaúba palm (*Acrocomia aculeata*) using transcriptome sequencing and cross-species transferability in Arecaceae species. BMC Plant Biology. 2018;18:276.
53. Li X, Yang N, Zhao Kg, Chen Yx, Tang Rj, Chen Lq. Development and primer selection of EST-SSR molecular markers based on transcriptome sequencing of *Chimonanthus praecox*. Journal of Beijing Forestry University. 2013;35(S1):25-32.
54. Liu Tm, Zeng Lb, Zhu Sy, Chen Xj, Tang Qm. Large-scale development of expressed sequence tag-derived simple sequence repeat markers by deep transcriptome sequencing in garlic (*Allium sativum* L.). Molecular Breeding. 2015;35:204.
55. Zhang Z, Xie W, Zhao Y, Zhang J, Wang Y. EST-SSR marker development based on RNA-sequencing of E. sibiricus and its application for phylogenetic relationships analysis of seventeen *Elymus* species. BMC Plant Biology. 2019;19:235.
56. Liang X, Xiaohong S, Sheng C, Xxiangyu X, Wwendan Z, huabing Y. Utilization of Simple Sequence Repeat(SSR)Markers Developed from ade novo Transcriptome Asembly in *Pueraria thomsoni* Benth. Acta Botanica Boreali-Occidentalia Sinica. 2019;39(01):59-67.
57. Shweta K, Shraddha U, Nita S, Sunil A, Dhammaprakash PW. Development of Genic Simple Sequence Repeat Markers as Novel Genomic Resources in Dolichos Bean (*Lablab purpureus* L.). Indian Journal of Plant Genetic Resources. 2022;35(1):80-4.
58. Song X, Li N, Guo Y, Bai Y, Wu T, Yu T, et al. Comprehensive identification and characterization of simple sequence repeats based on the whole-genome sequences of 14 forest and fruit trees. Forestry Research. 2021;1:7.
59. Liu T, Zhu S, Fu L, Tang Q, Yu Y, Chen P, et al. Development and Characterization of 1,827 Expressed Sequence Tag-Derived Simple Sequence Repeat Markers for Ramie (Boehmeria nivea L. Gaud). Plos One. 2013;8(4):e60346.
60. Chen LQ, Lu DF. Cultivar classification system of *Chimonanthus praecox* (L.) Link. Journal of Beijing Forestry University. 2001;23(S1):107-8.
61. Gao Xy, Su Qx, Yao Bl, Yang Wj, Ma Ws, Yang B, et al. Development of EST-SSR markers related to polyphyllin biosynthesis reveals genetic diversity and population structure in *Paris polyphylla*. Diversity. 2022;14(8):589.

Figures

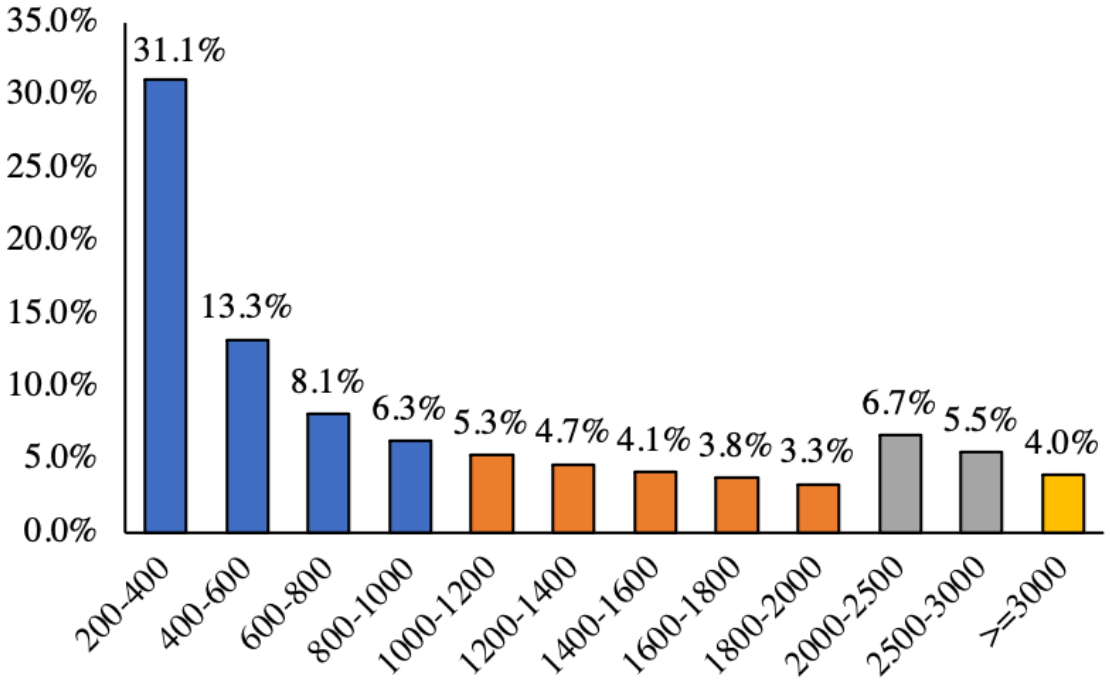

Figure 1

Information of unigenes length distribution

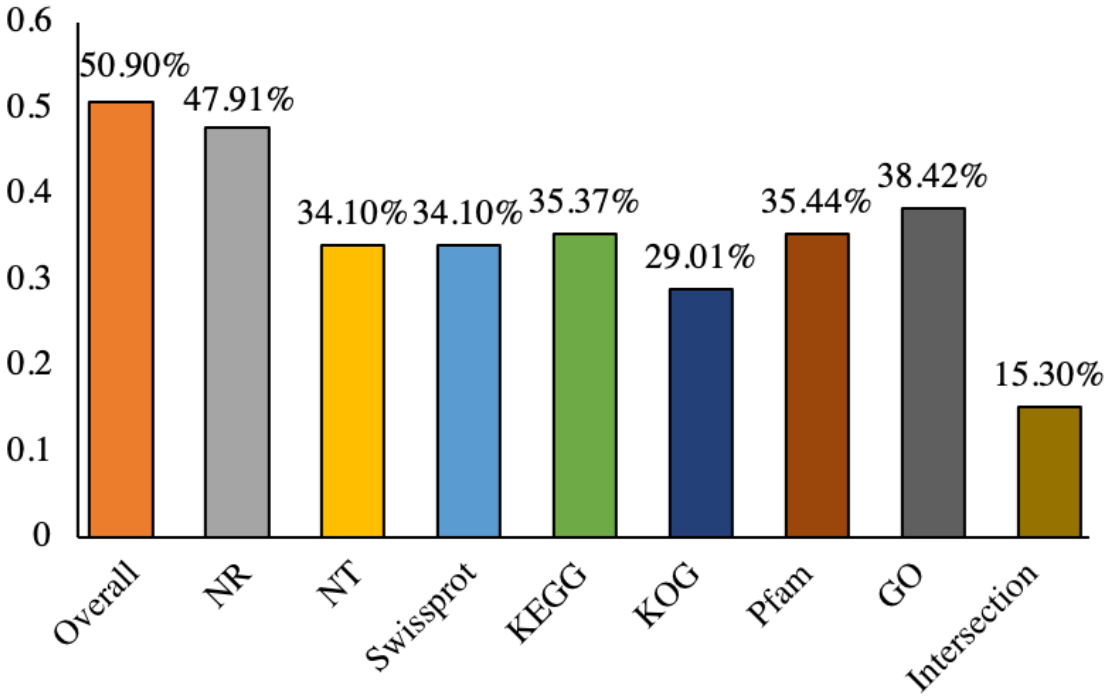

Figure 2

Unigenes annotation of *C.praecox*

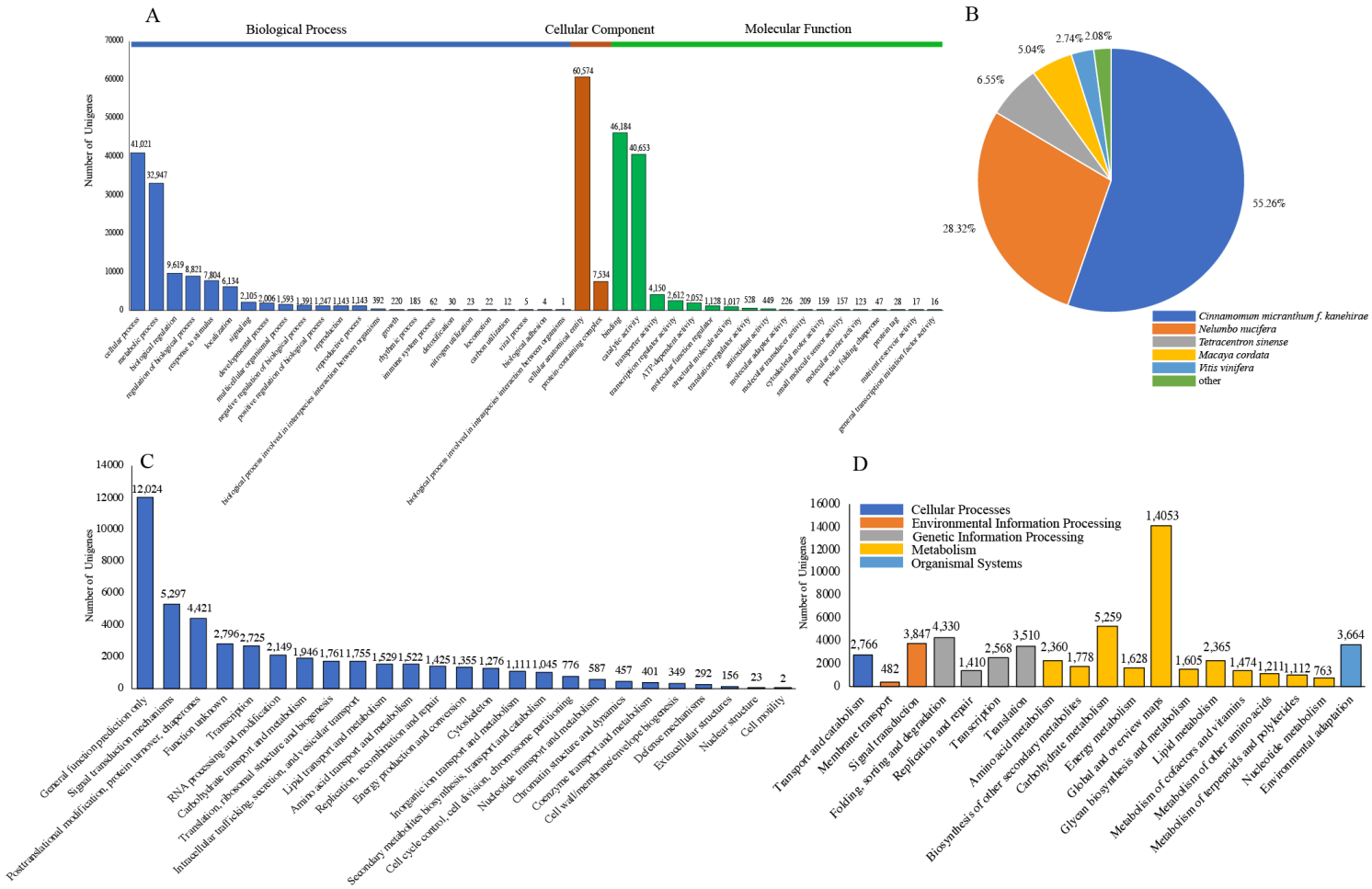

Figure 3

Unigenes annotation based on GO, NR, KOG, and KEGG databases. A. GO annotations of *C.praecox*. B. NR annotations of *C.praecox*. C. KOG annotations of *C.praecox*. D. KEGG annotations of *C.praecox*.

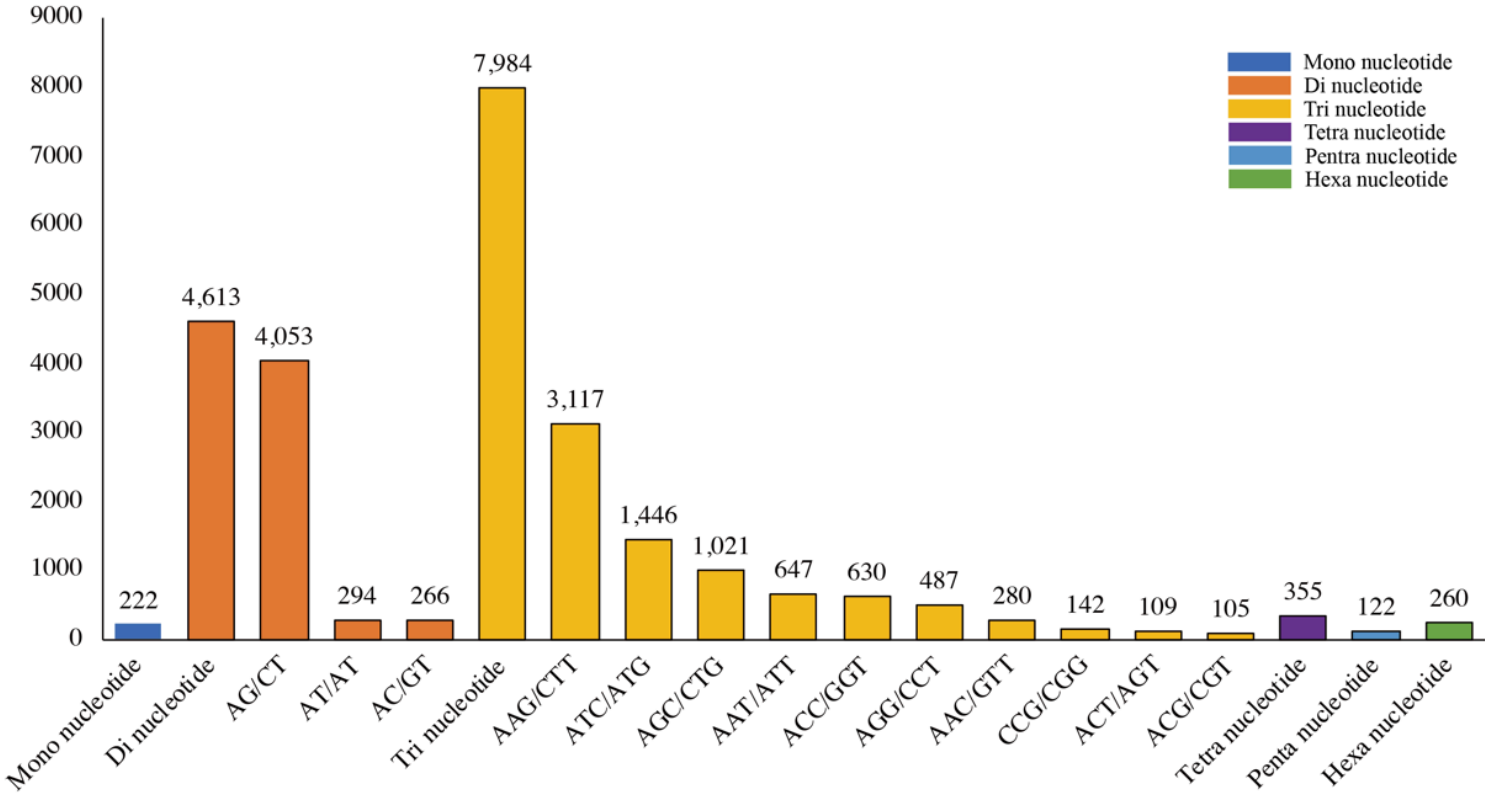

Figure 4

Microsatellite loci distribution in the transcriptome data of *C.praecox*

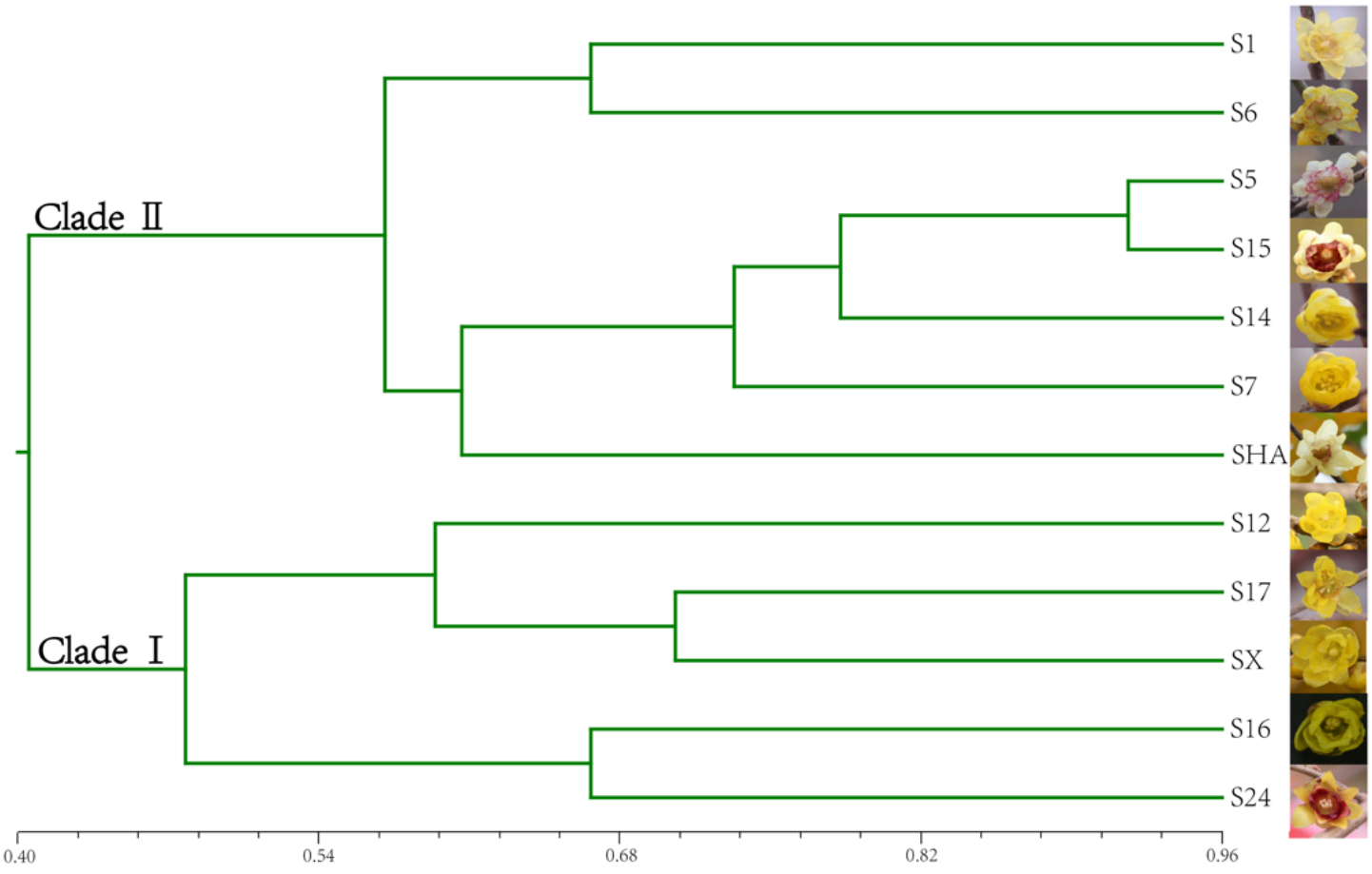

Figure 5

Phylogenetic tree of 12 *C.praecox*varieties

Supplementary Files

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

- TableS1.xlsx