



# Characterization and Development of Microsatellite Markers in *Pseudotaxus chienii* (Taxaceae) Based on Transcriptome Sequencing

Ruixiang Xu<sup>1</sup>, Zhen Wang<sup>1</sup>, Yingjuan Su<sup>1,2\*</sup> and Ting Wang<sup>3\*</sup>

<sup>1</sup> School of Life Sciences, Sun Yat-sen University, Guangzhou, China, <sup>2</sup> Research Institute of Sun Yat-sen University in Shenzhen, Shenzhen, China, <sup>3</sup> College of Life Sciences, South China Agricultural University, Guangzhou, China

## OPEN ACCESS

### Edited by:

Annie Machordom,  
National Museum of Natural Sciences  
(MNHN), Spain

### Reviewed by:

Salih Kafkas,  
Çukurova University, Turkey  
Zhiyong Zhang,  
Jiangxi Agricultural University, China  
Mehmet Karaca,  
Akdeniz University, Turkey

### \*Correspondence:

Yingjuan Su  
suyj@mail.sysu.edu.cn  
Ting Wang  
tingwang@scau.edu.cn

### Specialty section:

This article was submitted to  
Evolutionary and Population Genetics,  
a section of the journal  
Frontiers in Genetics

Received: 19 June 2020

Accepted: 21 September 2020

Published: 15 October 2020

### Citation:

Xu RX, Wang Z, Su YJ and  
Wang T (2020) Characterization  
and Development of Microsatellite  
Markers in *Pseudotaxus chienii*  
(Taxaceae) Based on Transcriptome  
Sequencing.  
*Front. Genet.* 11:574304.  
doi: 10.3389/fgene.2020.574304

*Pseudotaxus chienii* (Taxaceae) is an endangered conifer species endemic to China. However, a lack of suitable molecular markers hinders the genomic and genetic studies on this species. Here, we characterized and developed the microsatellite markers from a newly sequenced *P. chienii* transcriptome. A total of 21,835 microsatellite loci were detected from 161,131 non-redundant unigene sequences, and the frequency of SSRs was 13.55%, with an average of one SSR loci per 9.18 kb. Mono-nucleotide, di-nucleotide, and tri-nucleotide were the dominant repeat types, accounting for 50.06, 13.49, and 29.39% of the total SSRs, respectively. In terms of distribution location, the coding regions (CDS) with few microsatellites and mainly consisted of tri-nucleotides. There were significant differences in the length of microsatellite among genic regions and motif types. Functional annotation showed that the unigenes containing microsatellites had a wide range of biological functions, most of which were related to basic metabolism, and a few might be involved in expression regulation of gene and response to environmental stress. In addition, 375 primer pairs were randomly selected and synthesized for the amplification and validation of microsatellite markers. Seventy-seven primer pairs were successfully amplified and 40 primer pairs were found to be polymorphic. Finally, 20 pairs of primers with high polymorphism were selected to assess the genetic diversity in four *P. chienii* populations. In addition, the newly developed microsatellite markers exhibited high transferability (70%) in *Amentotaxus argotaenia*. Our study could enable further genetic diversity analysis and functional gene mining on Taxaceae.

**Keywords:** *Pseudotaxus chienii*, transcriptome, microsatellite markers, genetic diversity, transferability

## INTRODUCTION

*Pseudotaxus chienii*, belonging to monotypic genus *Pseudotaxus* W. C. Cheng (Taxaceae) endemic to China, is narrowly distributed Zhejiang, Jiangxi, Hunan and Guangdong Province, and the Guangxi Zhuang Autonomous Region (Fu et al., 1999). The species is an evergreen shrub or small tree plants, up to 4 meters, characterized by seeds of a white fleshy aril (Fu et al., 1999). Most of its habitats are in the subtropical evergreen broad-leaved forest or deciduous broad-leaved forest at an

elevation of about 800–1500 m (Fu et al., 1999). However, the distribution range and population size of *P. chienii* have been continuously declining due to the low seed-setting rate, the difficult natural reproduction, the strict habitat requirements, and the impact of human activities (Fu and Jin, 1992; Yang et al., 2005). Currently, *P. chienii* has been listed as a second-class nationally protected plant by the Chinese Red Data Book (Fu and Jin, 1992). Thus, it is imperative to develop effective strategies to protect this species. The first step is a comprehensive knowledge of its genetic diversity, which requires a powerful marker resource. However, due to the lack of genomic resources, the knowledge about functional gene analysis and molecular marker development of this species is limited.

Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), consist of short tandem repeated motifs of 1–6 bp units, which are ubiquitous in the genomes of prokaryotic and eukaryotic organisms (Tautz and Renz, 1984; Toth et al., 2000). Microsatellite markers are one of the most efficient types of molecular markers because of their high polymorphism, wide genome coverage, co-dominant inheritance, and good reproducibility (Powell et al., 1996; Gupta and Varshney, 2000; Varshney et al., 2005). Hence, they have been widely used in genetic linkage map, quantitative trait loci (QTL) mapping, marker-assisted selection (MAS) breeding, evolutionary studies, and genetic diversity analysis (Nishio et al., 2008; Garvin et al., 2010; Ahmadi and Fotokian, 2011; Wang et al., 2015; Zhao et al., 2019). Traditional methods for developing microsatellite markers are time-consuming, laborious, and expensive (Zane et al., 2002). With the development of next-generation sequencing technology (NGS), transcriptome sequencing has gradually become a convenient and efficient method for the development of large-scale SSR markers, especially for non-model species without an available reference genome (Taheri et al., 2018). SSRs can be divided into genomic SSRs (g-SSRs) and expressed sequence tag SSRs (EST-SSRs) (Eujayl et al., 2004). EST-SSRs are derived from gene transcription regions and are more conserved than g-SSRs, therefore, are more likely to be transferable across related species (Dutta et al., 2011). To date, there have been few investigations of EST-SSR markers in *P. chienii*, especially based on transcriptome sequencing.

The distribution of microsatellites in genomic regions is non-random (Li et al., 2002). For example, in many species, protein-coding regions contain fewer di-nucleotide and tetra-nucleotide SSRs, but more tri-nucleotide and hexa-nucleotide SSRs (Li et al., 2004). According to a study on SSRs in human and *Arabidopsis thaliana*, the number of SSRs in untranslated regions (UTRs) was more abundant than that in coding regions (CDS) (Zhang et al., 2004). In general, the occurrence frequency of SSRs in coding regions is relatively low, which may be explained by its high mutation rate affecting gene expression (Vieira et al., 2016). Besides, similar distribution patterns were observed in microsatellites located in intergenic regions and introns (Hancock, 1995).

The function of microsatellites is associated with their location in the genome. Compared with other genic regions, microsatellites in coding regions are the most conserved (Xin

et al., 2012), which may be attributed to selection pressure (Li et al., 2004). Microsatellite variations in coding regions can directly affect the protein synthesis process (Gao et al., 2013). SSR variations in 5'-UTRs can be involved in gene expression regulation by affecting transcription and translation (Li et al., 2004; Vieira et al., 2016). SSR expansions in 3'-UTRs can lead to transcription slippage and mRNA sequence expansion, which further disrupt mRNA splicing and other cell functions (Li et al., 2004; Kalia et al., 2011). Intronic SSRs may play a role in regulating gene transcription, mRNA splicing, or export to cytoplasm (Li et al., 2004). SSR variations in intergenic regions may have an impact on gene function by changing the secondary structure of DNA or the source of small RNAs (Gao et al., 2013). However, there has been no report on the function of microsatellites in *P. chienii*, which is not conducive to the discovery of functional genes and the identification of phenotypic variation.

In this study, we mined microsatellite loci from a newly sequenced transcriptome of *P. chienii* using Illumina sequencing technology. The aims of this study were to (i) characterize the frequency, distribution, function, and evolution of microsatellites in *P. chienii* transcriptome; (ii) develop polymorphic microsatellite markers and verified their polymorphism level for *P. chienii*; (iii) test cross-species transferability of the novel microsatellite markers among Taxaceae species. The results of this study will provide new insights into the function and evolution of microsatellites in *P. chienii* transcriptome. Additionally, the newly developed microsatellite markers may facilitate the conservation and management of *P. chienii* genetic resources.

## MATERIALS AND METHODS

### Sample Collection, RNA and DNA Extraction

The plant materials (root, stem, leaf, and strobilus) used for RNA isolation and transcriptome sequencing were collected from a single *Pseudotsaxus chienii* individual growing in Bijiashan, Jiangxi Province, China (26°30'37"N, 114°09'42"E, 1290 m asl). Each tissue sample was immediately stored in RNA Protective Additive after collection and removal of dust. The collected samples were first stored overnight at 4°C and then stored at -20°C the next day. Total RNA was separately extracted from each tissue using the RNeasy Plant Kit (Qiagen, Beijing, China) according to the manufacturer's instructions. RNA degradation and contamination were detected using 1% agarose gels. RNA quality assessments for purity, integrity, and concentration were separately performed using the NanoPhotometer spectrophotometer (IMPLEN, CA, United States), Qubit 2.0 Fluorometer (Life Technologies, CA, United States), and Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, United States).

For SSR characterization, the young leaves of 110 individuals were collected from four geographic localities across China: 31 from Maoshan, Zhejiang (MS), 30 from Yinshan park, Guangxi (YS), 30 from Bijiashan, Jiangxi (BJS), and 19 from

Zhangjiajie, Hunan (ZJJ) (Table 1). Total genomic DNA was extracted using the modified cetyltrimethyl ammonium bromide (CTAB) method (Su et al., 2005). The quality and quantity of DNA were determined by NanoDrop 2000 spectrophotometer (ThermoFisher Scientific, Waltham, MA, United States) and 0.8% agarose gel electrophoresis. Finally, all DNA samples were diluted to 50 ng/ $\mu$ L and stored at  $-20^{\circ}\text{C}$  until use.

## Transcriptome Sequencing and *de novo* Assembly

The qualified RNA samples from four tissues were used for cDNA library construction and transcriptome sequencing. Four cDNA libraries were constructed using NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, United States) according to the manufacturer's instructions. Transcriptome sequencing was performed using Illumina HiSeq 2500 platform (Illumina, San Diego, CA, United States). After sequencing, raw reads were filtered to remove the adaptor sequences and low-quality reads. The remaining high-quality clean reads were *de novo* assembled using Trinity software (Grabherr et al., 2011) with `min_kmer_cov` set to 2 and other parameters set to default. Assembled transcripts were clustered, and removal of the redundant sequences to generate non-redundant unigenes by CD-HIT software (Fu et al., 2012).

## Microsatellite Identification and Characterization

The MISA software (Microsatellite searching tool<sup>1</sup>) was used to identify microsatellite loci in the assembled transcriptome of *P. chienii*. The parameters were set as follows: mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides with minimum repeat numbers of 10, 6, 5, 5, 5, and 5, respectively. For compound SSRs, the maximum interruption between two SSRs was set as 100 bases.

In order to investigate the distribution of microsatellites in *P. chienii* transcriptome, the coding sequence (CDS) regions of unigenes were predicted based on the BLAST results against the Swiss-Prot database ( $E$ -value  $< 10^{-5}$ ), and then CDS regions that showed no hits in BLAST were predicted using ESTScan (Iseli et al., 1999). The relative position of microsatellites (CDS, 5'-UTR, or 3'-UTR) was inferred from the positions of start and stop codons of the CDS. Chi-square analysis was used to test whether there were significant differences in different genic regions (CDS, 5'-UTR, and 3'-UTR) for the density of microsatellites. Kruskal-Wallis rank sum test (Kruskal and Wallis, 1952) was used to test the influence of genic region, repeat motif size, and repeat motif type on microsatellite length.

## Function Annotation of Unigenes Containing Microsatellites

To identify the possible function of microsatellites, all unigene sequences containing microsatellites were aligned to the public databases by BLASTX program ( $E$ -value  $< 10^{-5}$ ), including

**TABLE 1** | Sampling location information of four *P. chienii* populations.

Population	Location	Sample Size	Latitude (N)	Longitude (E)	Altitude (m a.s.l.)
MS	Maoshan, Zhejiang	31	118°58'23"	28°06'08"	1120
YS	Yinshan park, Guangxi	30	110°14'53"	24°09'15"	1050
BJS	Bijijashan, Jiangxi	30	114°09'41"	26°30'35"	1340
ZJJ	Zhangjiajie, Hunan	19	110°28'56"	29°23'12"	1055

Nr (NCBI non-redundant protein sequences), Nt (NCBI non-redundant nucleotide sequences), Swiss-Prot (A manually annotated and reviewed protein sequence database), KOG (Clusters of eukaryotic Orthologous Groups), and Pfam (Protein family, assigned using the HMMER3.0 package). Besides, for all unigene sequences containing microsatellites, the GO (Gene Ontology) annotation was performed by the Blast2GO software (Conesa et al., 2005) ( $E$ -value  $< 10^{-6}$ ). The metabolic pathway analysis was predicted by KAAS (KEGG Automatic Annotation Server) ( $E$ -value  $< 10^{-10}$ ).

To determine whether microsatellites were significantly enriched in some GO terms or KEGG pathways, GO and KEGG enrichment analysis were performed with the topGO package (Alexa and Rahnenfuhrer, 2010) and clusterProfiler package (Yu et al., 2012), respectively. Over-represented GO terms/ KEGG pathways were estimated by Fisher's exact test with multiple testing correction of FDR ( $Q$ -value  $< 0.05$ ) (Benjamini and Hochberg, 1995).

## Amplification and Validation of Microsatellite Markers

Microsatellite primers were designed using Primer Premier 5.0 (Lalitha, 2000) based on the following criteria: (1) primer length of 18–25 bp, with an optimum length of 20 bp; (2) annealing temperature of 50–62°C with a maximum of 5°C difference in annealing temperature between forward and reverse primers; (3) CG content from 40 to 60%; (4) the expected PCR product sizes of 100–500 bp; (5) avoidance of primer dimers and hairpin structures.

PCR amplifications were conducted in a volume of 25  $\mu$ L containing 50 ng of template DNA, 2.5  $\mu$ L of 10  $\times$  PCR buffer (2.0 mM  $\text{Mg}^{2+}$ ), 1.6  $\mu$ L of dNTPs (10 mM), 0.5  $\mu$ L of each primer (10  $\mu$ M), 1 U of Taq polymerase, and 18.7  $\mu$ L of double-distilled water. The PCR conditions were as follows: initial denaturation at 94°C for 5 min, followed by 35 cycles of 40 s at 94°C, 40 s at annealing temperature, and 30 s at 72°C, and a final extension at 72°C for 10 min. The amplified products were detected by 1.5% (w/v) agarose gel electrophoresis. Then, the successfully amplified primers were selected for polymorphic screening using 6% denaturing polyacrylamide gel electrophoresis. Finally, microsatellite primers with high polymorphism and good repeatability were selected for genotyping on an ABI 3730XL DNA Analyzer

<sup>1</sup><http://pgrc.ipk-gatersleben.de/misa/>

(Applied Biosystems, Foster City, CA, United States) with GeneScan LIZ 500 as an internal reference (Applied Biosystems). Allele sizes were detected with the software GeneMapper v4.0 (Applied Biosystems).

The genetic parameters of microsatellite loci were estimated by GenAlEx v6.5 software (Peakall and Smouse, 2012), including the number of alleles (Na), the number of effective alleles (Ne), observed heterozygosity (Ho), and expected heterozygosity (He). The polymorphic information content (PIC) of each microsatellite marker was calculated using PIC\_CALC v0.6 (Nagy et al., 2012). The null allele frequency for each marker was estimated using MICRO-CHECKER 2.2.3 (Van Oosterhout et al., 2004). Tests for deviations from linkage disequilibrium (LD) and Hardy-Weinberg equilibrium (HWE) were carried out by GENEPOP v4.2 (Raymond and Rousset, 1995). The software BOTTLENECK v1.2.02 (Piry et al., 1999) was used to detect recent population bottlenecks under three different models of microsatellite evolution (Infinite allele model, IAM; Stepwise mutation model, SMM; Two-phased model of mutation, TPM).

### Cross-Species Amplification and Transferability Analysis

For cross-species transferability analysis, the related species of the family Taxaceae, *Amentotaxus argotaenia* (with 58 individuals from four different populations, **Supplementary Table 1**), was chosen to evaluate the transferability of these newly developed microsatellite markers. Genomic DNA isolation and PCR amplification were performed as described above. PCR products were separated by capillary electrophoresis using an ABI 3730XL DNA Analyzer (Applied Biosystems, Foster City, CA, United States), and the sizes of fragments were determined using GeneMapper v4.0 (Applied Biosystems).

## RESULTS

### Illumina Sequencing and *de novo* Assembly

In this study, the Illumina HiSeq 2500 platform was used to sequence the transcriptome of *P. chienii*. A total of 52,633,598, 57,112,536, 62,840,720, and 42,617,842 raw reads were generated from root, stem, leaf, and strobilus, respectively. After trimming adaptors and removing low-quality reads, 51,073,864, 56,419,624, 61,110,258, 41,448,574 high-quality clean reads were obtained for root, stem, leaf, and strobilus (**Table 2**). The total nucleotide number of each tissue was greater than 6G. The Q20 and Q30 ratio of each tissue were over 95 and 91%, respectively. The GC content of each tissue was approximately 45% (**Table 2**). Thus, the sequencing data was of high quality and met the requirement for subsequent analyses. Using the Trinity program (Grabherr et al., 2011), the high-quality clean reads from four tissue were assembled into 156,747, 202,908, 107,675, and 201,124 transcripts, respectively. The assembled transcripts of these tissues were clustered into 66,126, 79,842, 52,207, and 60,391 unigenes (**Table 2**). After removing the redundant sequences, a total of 161,131 non-redundant unigenes were generated.

### Characterization of Microsatellite in *P. chienii* Transcriptome

Using the MISA program, a total of 21,835 potential microsatellite loci were identified from 161,131 non-redundant unigenes, with 2,969 unigenes containing more than one microsatellite loci. The frequency of microsatellite in *P. chienii* transcriptome was 13.55%; an average of one microsatellite loci was found every 9.18 kb (**Table 3**). The most abundant repeat motif types were mono-nucleotide (10,930, 50.06%), which accounted for about half of total SSRs, followed by tri-nucleotide

**TABLE 2** | Summary of *de novo* transcriptome assembly of four tissues from *P. chienii*.

Categories	Items	Root	Stem	Leaf	Strobilus
Raw reads	Total raw reads	52,633,598	57,112,536	62,840,720	42,617,842
Clean reads	Total clean reads	51,073,864	56,419,624	61,110,258	41,448,574
	Total clean bases (Gb)	7.66	8.46	9.17	6.22
	Q20 (%)	97.16	98.03	96.9	97.28
	Q30 (%)	92.16	94.03	91.55	92.34
	GC content (%)	44.25	46.26	45.57	45.23
Transcripts	Total number of transcripts	156,747	202,908	107,675	201,124
	Total length of transcripts (bp)	113,815,935	136,683,283	89,744,115	114,523,821
	Maximum length of transcripts (bp)	13,579	12,558	11,562	12,275
	Minimum length of transcripts (bp)	201	201	201	201
	Mean length of transcripts (bp)	726	674	833	569
	N50 (bp)	1,395	1,224	1,621	1,011
Unigenes	Total number of unigenes	66,126	79,842	52,207	60,391
	Total length of unigenes (bp)	87,790,725	101,369,156	74,169,328	76,205,289
	Maximum length of unigenes (bp)	13,579	12,558	11,562	12,275
	Minimum length of unigenes (bp)	201	201	201	201
	Mean length of unigenes (bp)	1,328	1,270	1,421	1,262
	N50 (bp)	1,843	1,728	1,931	1,754

**TABLE 3** | Analysis results of microsatellite based on the *P. chienii* transcriptome.

Items	Number
Total number of sequences examined	161,131
Total size of examined sequences (bp)	200,495,148
Total number of identified microsatellite loci	21,835
Number of microsatellites containing sequences	17,974
Number of sequences containing more than 1 microsatellite loci	2,969
Number of microsatellites present in compound formation	1,513
Frequency of microsatellite loci	13.55
Distribution density of microsatellite loci (kb)	9.18
Mono-nucleotide	10,930
Di-nucleotide	2,946
Tri-nucleotide	6,418
Tetra-nucleotide	354
Penta-nucleotide	343
Hexa-nucleotide	844

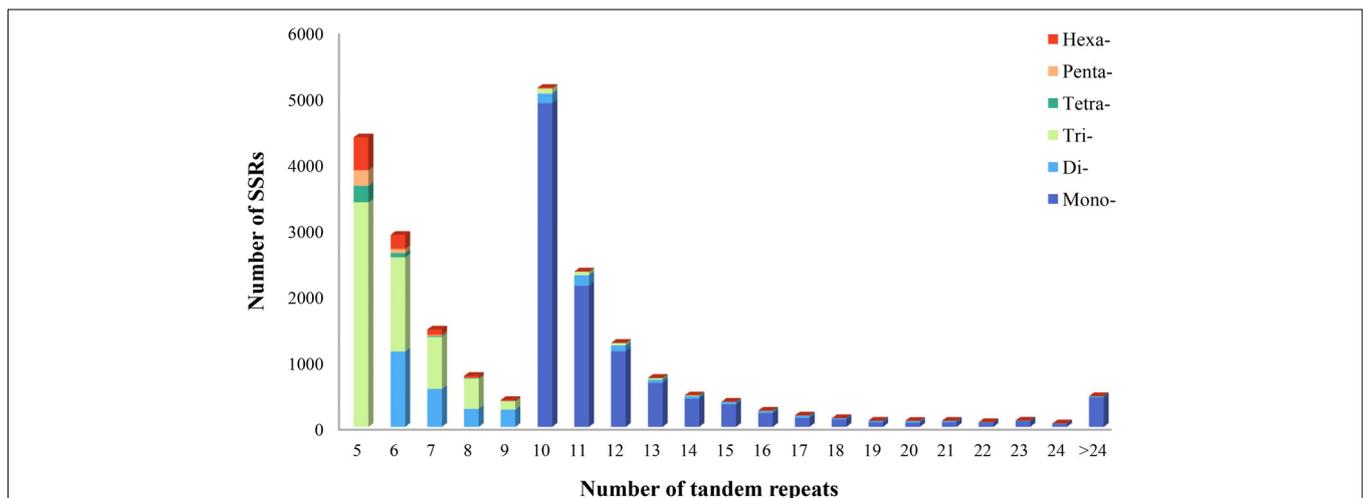
(6,418, 29.39%), di-nucleotide (2,946, 13.49%), hexa-nucleotide (844, 3.87%), tetra-nucleotide (354, 1.62%), and penta-nucleotide (343, 1.57%) (Table 3). The number of tandem repeats of microsatellite motifs ranged from five to 107. Microsatellite with ten tandem repeats (5,141, 23.54%) were the most common, followed by five tandem repeats (4,393, 20.12%), six tandem repeats (2,913, 13.34%), and 11 tandem repeats (2,357, 10.79%). Microsatellite motifs with > 24 tandem repeats only accounted for 2.13% (Figure 1).

Of the 21,835 microsatellite loci, 239 different repeat motifs were detected. For the mono-nucleotide motif, A/T was the most abundant type with a frequency of 49.13%. For the di-nucleotide repeat motifs, AT/AT (6.31%) and AG/CT (5.32%) were the dominant types. The two most frequent types in the tri-nucleotide were AAG/CTT (6.38%) and AGG/CCT (6.16%). Moreover, a low percentage (7.06%) of tetra-, penta- and hexa-nucleotide repeat motifs were observed in all identified microsatellite motifs (Figure 2).

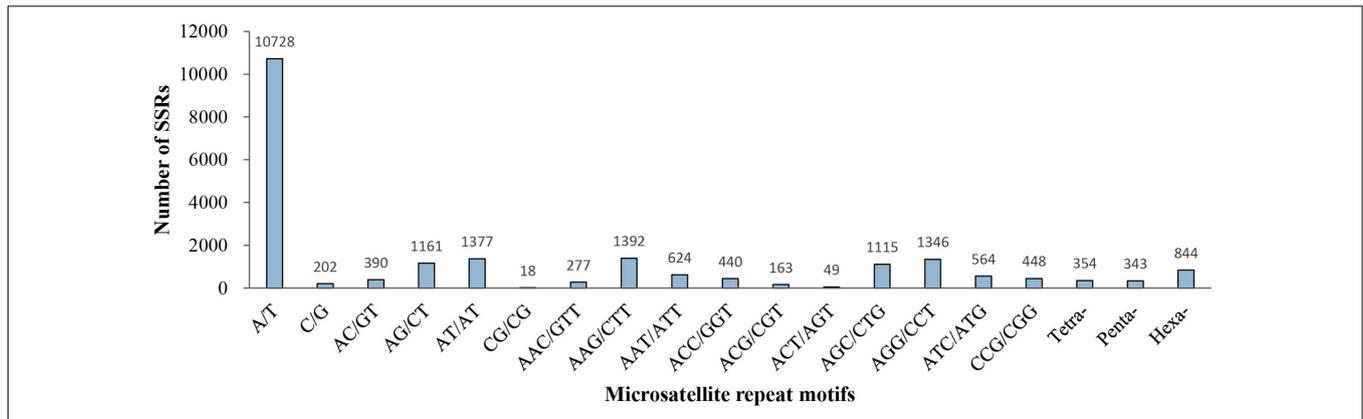
Microsatellite length ranged from 10 to 107 bp, with an average length of 16.23 bp. In general, the length of microsatellites differed significantly between the motif size classes (Kruskal–Wallis rank sum test,  $P < 2.2e^{-16}$ ), and their average length decreased with the increase of the motif size (Nemenyi test,  $P < 2.2e^{-16}$ ). However, there was no significant difference in length between tetra- and penta-nucleotide according to the pairwise length comparisons by the Nemenyi test (Supplementary Figure 1).

### Microsatellite Distribution in Different Genic Regions

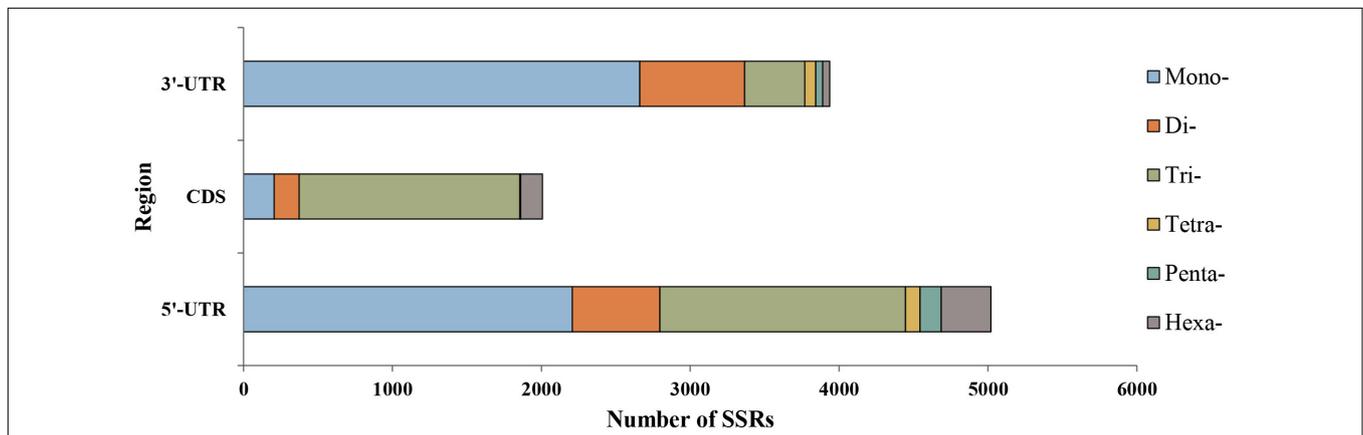
We explored the distribution characteristics of microsatellite loci in *P. chienii* transcriptome. Of the 21,835 microsatellite loci, 2,007 and 8,958 were located in the coding sequence regions (CDS) and untranslated regions (UTRs), respectively (Figure 3). The remaining 10,870 microsatellites were excluded from the analysis because there was insufficient information to determine their distribution. Microsatellites in different genic regions (CDS, 5'-UTRs, and 3'-UTRs) showed distinct patterns of distribution ( $\chi^2 = 2993.2$ ,  $P < 2.2e^{-16}$ ). CDS contained fewer microsatellites and was dominated by trinucleotides (1,481, 73.79%), whereas the mono- and di-nucleotide microsatellites in the UTRs (4,869, 54.35%) were abundant. Smaller motif types (mono-, di-, and trinucleotide microsatellites) were also observed to be more common in the transcriptome (Figure 3). Furthermore, significant differences in the length of microsatellite among three regions (CDS, 5'-UTR, and 3'-UTR) (Kruskal–Wallis rank sum test,  $P < 2.2e^{-16}$ , Supplementary Figure 2). The average length of microsatellites located in the CDS regions (17.84 bp) was slightly longer than in the UTRs (16.43 bp). The length of the microsatellite was also affected by the interaction of the regions and motif types according to the Kruskal–Wallis rank sum test ( $P < 2.2e^{-16}$ , Supplementary Figure 3).



**FIGURE 1** | Distribution of microsatellite motif types and tandem repeat numbers in *P. chienii* transcriptome.



**FIGURE 2 |** Distribution of microsatellite repeat motifs in *P. chienii* transcriptome.



**FIGURE 3 |** Distribution of six microsatellite repeat types in different genic regions of *P. chienii*.

### Functional Annotation and Classification of Unigenes Containing Microsatellites

To predict the potential functions of microsatellites, all microsatellite-containing unigenes were blasted against the public databases using the BLASTX program. As a result, 12,652 (70.39%) unigenes were annotated in Nr, 6,660 (37.05%) in Nt, 10,622 (59.10%) in Swiss-Prot, 4,274 (23.78%) in KOG, 11,267 (62.68%) in Pfam, 11,334 (63.06%) in GO, and 5,417 (30.14%) in KEGG. A total of 2,045 (11.38%) unigenes were annotated in all seven databases, and 14,264 (79.36%) unigenes were annotated in at least one database (Table 4).

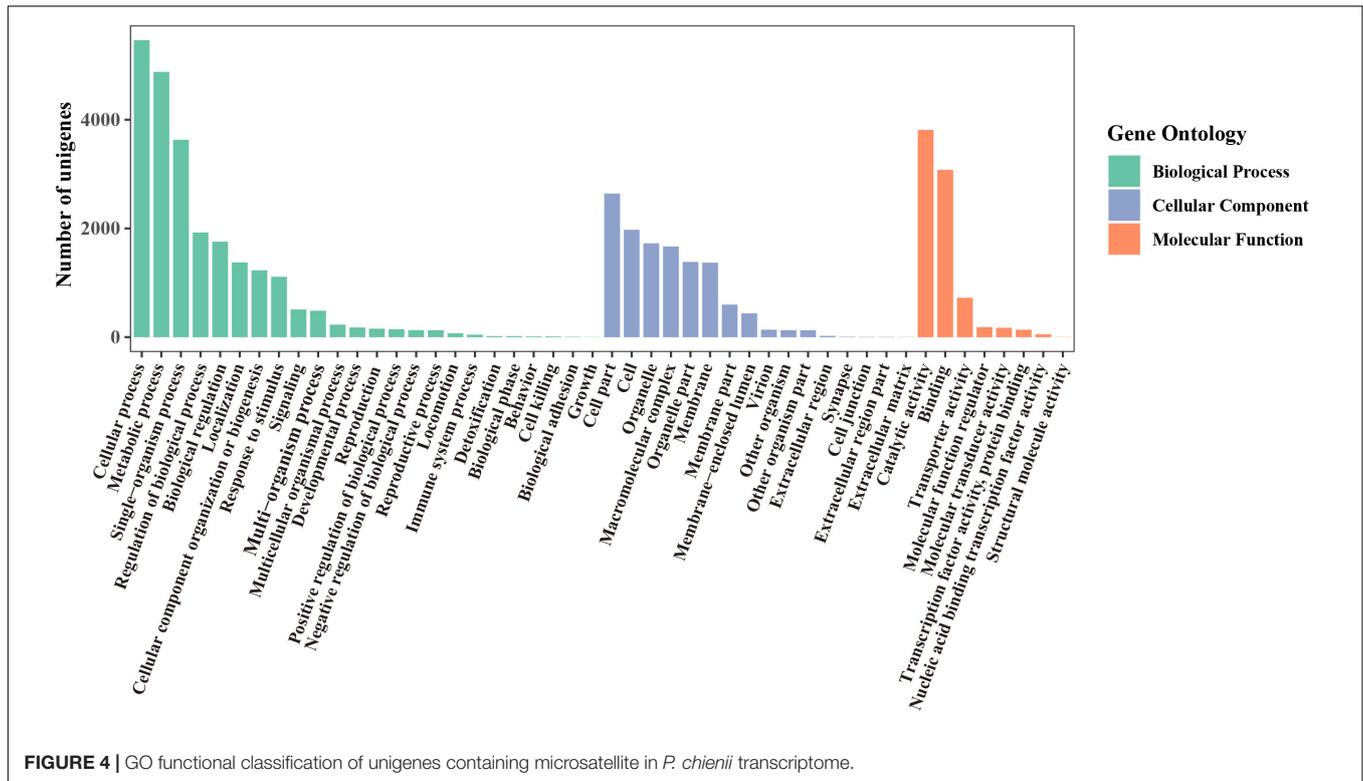
Using the Blast2GO Program, the microsatellite-containing unigenes were divided into three categories: biological process (BP), cellular component (CC), and molecular function (MF), including 48 sub-categories (Figure 4). In the BP category, “cellular process” (5,464, 48.21%) and “metabolic process” (4,879, 43.05%) were represented prominently. Within the CC category, “cell” (1,975, 17.43%) and “cell part” (2,642, 23.31%) were the most abundant terms, whereas only a few unigenes were assigned to “extracellular region part” (3), “cell junction” (3), and “extracellular matrix” (2). Among eight different MF categories, “catalytic activity” (3,813, 33.64%) and “binding” (3,078, 27.16%)

**TABLE 4 |** Summary of functional annotation results of unigenes containing microsatellite in *P. chienii* transcriptome.

Annotated databases	Number of unigenes	Percentage (%)
Nr	12,652	70.39
Nt	6,660	37.05
Swiss-Prot	10,622	59.10
KOG	4,274	23.78
Pfam	11,267	62.68
GO	11,334	63.06
KEGG	5,417	30.14
Annotated in all databases	2,045	11.38
Annotated in at least one database	14,264	79.36
Total	17,974	100

were the two most frequent classes. These results indicated that the SSRs-containing unigenes were mainly involved in the basal metabolism of cells.

Furthermore, KEGG analysis was used to determine the biological pathways and functions of the unigenes containing microsatellites. In total, 5,125 (28.51%) unigenes were mapped in 124 different pathways and classified into 5 main categories.



The largest category was “metabolism” (3,086, 56.97%), followed by “genetic information processing” (1,326, 24.48%), “cellular processes” (284, 5.24%), “environmental information processing” (228, 4.21%), and “organismal systems” (201, 3.71%) (Figure 5). Among these 124 pathways, the most represented pathways including “spliceosome,” “ribosome,” “carbon metabolism,” “plant–pathogen interaction,” “plant hormone signal transduction,” “biosynthesis of amino acids,” “RNA transport,” “protein processing in endoplasmic reticulum,” “starch and sucrose metabolism,” and “ubiquitin mediated proteolysis” (Supplementary Table 2).

Finally, all microsatellite-containing unigenes were searched against the KOG database to classify and functionally annotation the orthologous gene products. The results showed that a total of 4,274 unigenes were classified into 24 functional groups. Among them, the three largest groups were “posttranslational modification, protein turnover, chaperones” (642, 15.02%), “general function prediction only” (511, 11.96%), and “translation, ribosomal structure and biogenesis” (484, 11.32%), respectively. However, the smallest categories were “extracellular structures” and “nuclear structure,” with two and 18 genes annotated, respectively (Figure 6). These results reflected that the unigenes containing microsatellites had extensive basic biological functions.

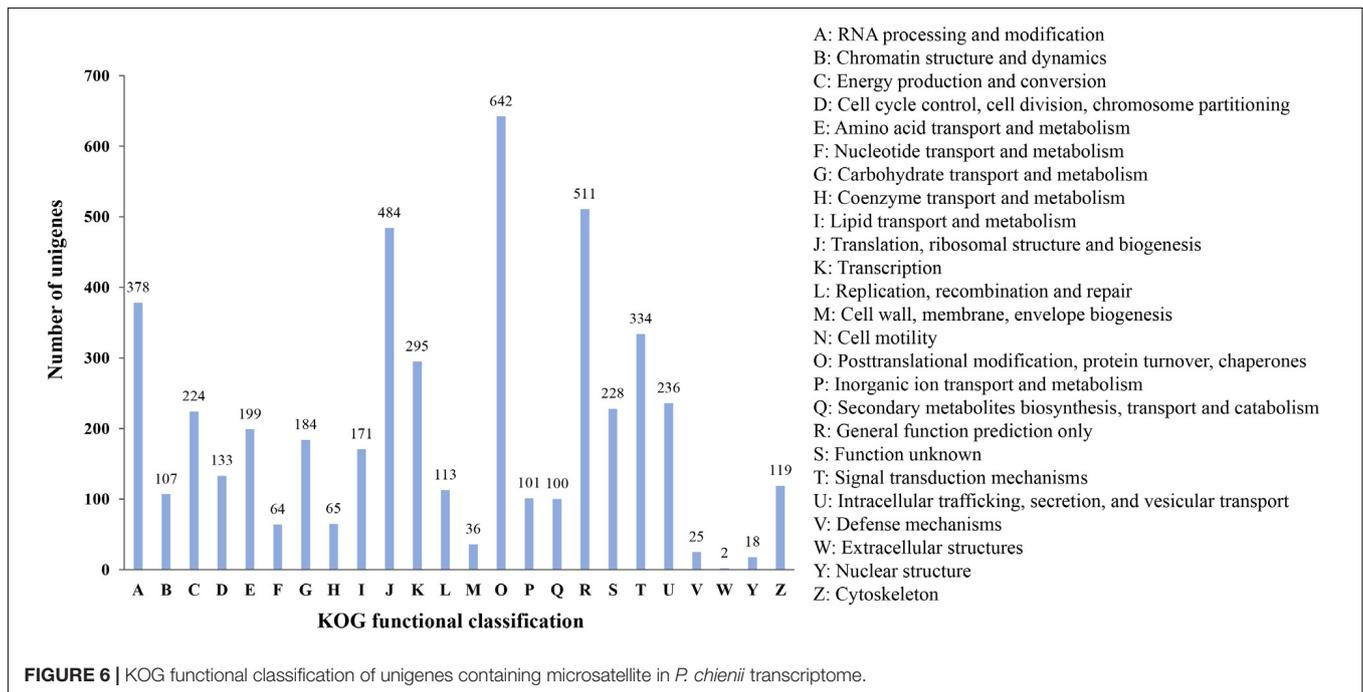
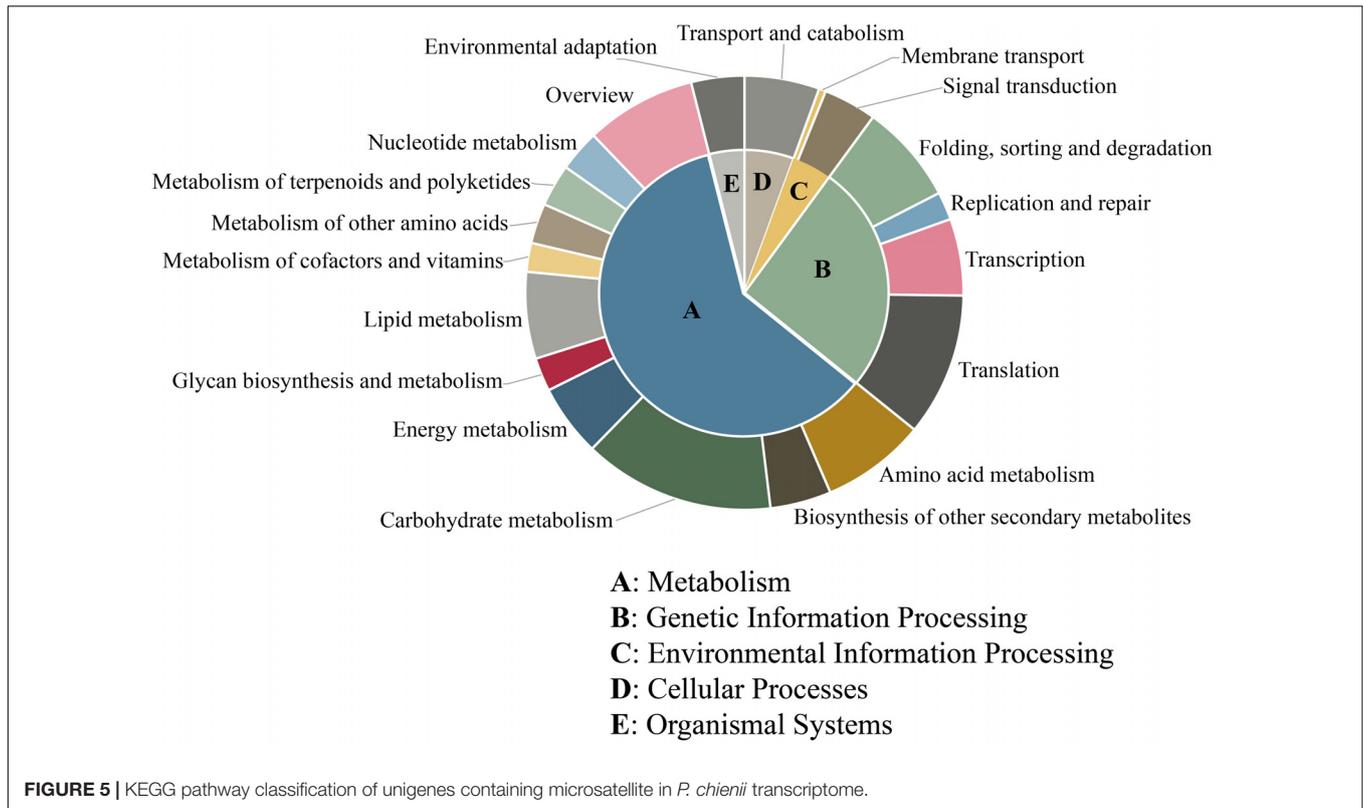
### GO and KEGG Enrichment Analysis of the Microsatellite-Containing Unigenes

In order to further identify molecular and biological functions of the unigenes containing microsatellites, Fisher’s exact test

was used for GO and KEGG enrichment analysis, respectively. There was significant enrichment of microsatellites in a total of 211 GO terms and most were mapped in transcription, regulation of transcription, and binding (Q-value < 0.05, Supplementary Table 3). KEGG pathway enrichment analysis indicated that the microsatellite-containing unigenes showed significantly enriched in 13 pathways (Q-value < 0.05), and Nine of them are related to basal metabolism and biosynthesis of cells. Notably, the remaining four pathways involved in gene expression regulation, and might play important roles in the response to stress (Figure 7).

### Development and Validation of Microsatellite Markers

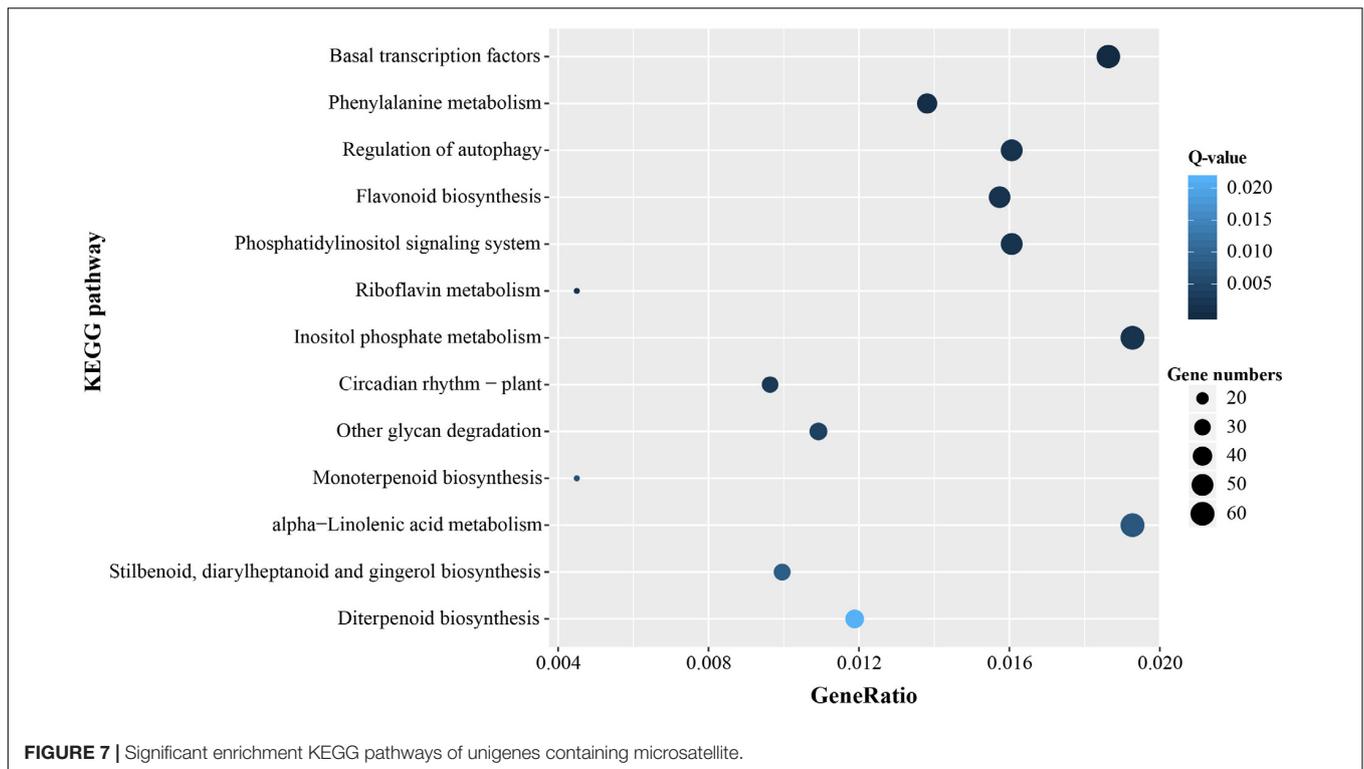
Based on the SSR-containing unigene sequences, 375 primer pairs were randomly selected and synthesized for the amplification and validation of microsatellite markers. After screening, 77 primer pairs (20.59%) successfully amplified and produced clear bands with expected size, 40 pairs of primers were identified as polymorphic, and 37 pairs were monomorphic (Supplementary Table 4). Finally, 20 primers with high polymorphism were selected and used for genetic diversity analysis (Table 5). In total, 128 alleles were detected at the 20 polymorphic microsatellite loci. The number of amplified alleles per locus (Na) ranged from 2 (*P. chienii*-198) to 16 (*P. chienii*-152), with an average of 6.4 alleles per locus. The observed (Ho) and expected (He) heterozygosity ranged from 0.000 to 1.000 and from 0.129 to 0.887, with an average of 0.349 and 0.527, respectively. The PIC values ranged from 0.125 to 0.877, with an average of



0.483. Besides, a high frequency of null alleles was detected at six loci (*P. chieni*-17, *P. chieni*-25, *P. chieni*-86, *P. chieni*-162, *P. chieni*-267, and *P. chieni*-288), and one pair of loci (*P. chieni*-28 and *P. chieni*-162) showed LD. HWE test revealed that 13 of 20 microsatellite loci showed significant deviations

from the equilibrium after Bonferroni correction ( $P < 0.05$ ) (Table 5).

At the population level, MS population had the highest level of genetic diversity ( $N_a = 3.450$ ,  $N_e = 1.997$ ,  $H_o = 0.384$ , and  $H_e = 0.389$ ), whereas the lowest diversity levels were in



ZJJ population ( $N_a = 2.850$ ,  $N_e = 1.759$ ,  $H_o = 0.272$ , and  $H_e = 0.340$ ). All of the inbreeding coefficient ( $F_{IS}$ ) values were positive for four populations, which indicated that there was a certain degree of inbreeding in *P. chienii* populations (Table 6). Besides, bottleneck analysis showed that YS and BJS populations showed heterozygous excess and might experience a significant genetic bottleneck under the SMM models ( $P < 0.05$ ) (Supplementary Table 5).

### Transferability of Microsatellite Markers

To evaluate the transferability of the microsatellite markers from the *P. chienii* transcriptome, we performed cross-species amplification analysis for these 20 novel microsatellite markers in *A. argotaenia*, 14 primers were amplified products. Therefore, the overall transferability rates were 70%. Among the 14 primer pairs, 12 showed polymorphisms in 58 *A. argotaenia* samples. For the 12 polymorphic loci, there were 2–7 alleles at each locus, with a total of 50 alleles. The  $H_o$ ,  $H_e$ ,  $I$ , and PIC per locus ranged from 0 to 1.000, 0.133 to 0.741, 0.257 to 1.539, and 0.155 to 0.664, respectively (Supplementary Table 6). The microsatellite loci were moderately polymorphic.

## DISCUSSION

### Characterization of Microsatellites in *P. chienii* Transcriptome

Recently, next-generation sequencing (NGS) technologies, such as transcriptome sequencing, have made it possible to develop large-scale microsatellite markers (Taheri et al., 2018). In this

study, a total of 21,835 microsatellites were identified from 161,131 unigenes of the transcriptome. The frequency of microsatellites was 13.55%, and the distribution density was one SSR per 9.18 kb, which was much higher than previous reports in *Pinus dabeshanensis* (1/23.08 kb) (Xiang et al., 2015) and *Larix principis-rupprechtii* Mayr (1/26.8 kb) (Dong et al., 2018), but lower than *Cephalotaxus hainanensis* (1/1.652 kb) (Qiao et al., 2014) and *Glyptostrobus pensilis* (1/7.59 kb) (Li et al., 2019). This result showed that SSR sites were abundant in *P. chienii* transcriptome. The frequency and density of SSRs varied from species to species, which could be attributed to the size of unigene assembly dataset, the search criteria, mining tools for SSR locus, and genome organization (Morgante et al., 2002; Varshney et al., 2005; Taheri et al., 2019).

In our study, six different repeat motifs were identified. Excluding the mono-nucleotide repeats, tri-nucleotide (29.39%) repeats were the most abundant type, followed by di-nucleotide repeats (13.49%). The AAG/CTT (1,382, 6.38%) motif was the most abundant of the tri-nucleotide repeats, followed by AGG/CCT (1,346, 6.16%). This result might be common in conifer species, including *Glyptostrobus pensilis* (Li et al., 2019), *Larix principis-rupprechtii* Mayr (Dong et al., 2018), and *Torreya grandis* (Zeng et al., 2018). Among the di-nucleotide repeats, AT/AT (1377, 6.31%) was the most dominant motif, which was consistent with studies in *Cephalotaxus hainanensis* (Qiao et al., 2014) and *Pinus koraiensis* (Li et al., 2020), but not in *Pennisetum purpureum* (Zhou et al., 2018), *Liquidambar formosana* (Chen et al., 2020), and *Rhododendron arboretum* (Sharma et al., 2020). By contrast, the lowest frequency of CG/CG motifs (18, 0.08%) of the

**TABLE 5 |** Characteristics of 20 newly developed polymorphic microsatellite markers in *P. chienii*.

Locus	Primer sequence (5'-3')	Repeat motif	Size range (bp)	Annealing temperature (°C)	Na	Ho	He	PIC*	Null	GenBank accession no.
<i>P. chienii</i> -17	F: GGTCGAGTACGTGGTGGTTT R: GCCTGCGCTGTCATAAACTG	(AGG)5	158–161	57	3.000	0.145	0.257	0.237**	0.2703 (Y)	MT563347
<i>P. chienii</i> -20	F: TGTGCCAGTACTGCTACTGC R: TGAATGCGTGCGGAAACAAG	(ACC)6	183–195	57	5.000	0.209	0.585	0.517	–0.0267 (N)	MT563350
<i>P. chienii</i> -25	F: GATGCCGCTGGTTTCAATCC R: GCCGTACCGATTGGGATCAT	(GGA)8	198–214	57	5.000	0.309	0.397	0.366*	0.315 (Y)	MT563351
<i>P. chienii</i> -28	F: GAGTGGGAGACGAAGAGTGC R: CGAAGTGGGCTGCAACAATG	(CTC)8	261–266	57	6.000	0.600	0.670	0.626	0.0196 (N)	MT563352
<i>P. chienii</i> -29	F: AGCTGCAAGGCTACACAGAG R: CAATCCCGGGCCTGTAGAA	(GAA)5	235–239	57	5.000	0.318	0.563	0.500**	–0.1413 (N)	MT563353
<i>P. chienii</i> -36	F: GGGCCATCCTCTTCTCAAC R: CTCGACACTGCTCCACATCT	(TCC)8	240–247	57	3.000	0.055	0.129	0.125**	0.0299 (N)	MT563358
<i>P. chienii</i> -75	F: CGCTCCAACGAATCCAACC R: TAATGCCATCCGCACAACC	(CAGAAG)5	241–277	57	14.000	0.743	0.844	0.826	–0.1586 (N)	MT563368
<i>P. chienii</i> -86	F: GAATTTGAAGCACGGCCTCA R: GAGTGCCTGCTTTCTGGAT	(GGCACC)5	242–277	57	7.000	0.156	0.595	0.549**	0.3533 (Y)	MT563369
<i>P. chienii</i> -134	F: ACGCCAGTTAGGACACAAT R: CCTAGATCAAGAGCGCCTG	(CTT)6	238–319	57	6.000	0.227	0.389	0.364	0.0983 (N)	MT563374
<i>P. chienii</i> -141	F: CTGTCAACAAGCGGCTTTCC R: AGAGCCGGGGAAAATTGAG	(CGG)7	232–319	57	8.000	0.473	0.694	0.640**	0.0296 (N)	MT563377
<i>P. chienii</i> -152	F: CCCATCTGAACCCACGCTAA R: AAAGCGCTCATGCCAAAAC	(GGC)7	239–300	57	16.000	1.000	0.887	0.877**	–0.2134 (N)	MT563378
<i>P. chienii</i> -162	F: ACCTATCACCTCCTCGACCC R: CCGTTCATCACTGTGGACA	(CCACCG)6	203–229	55	13.000	0.573	0.807	0.784**	0.3871 (Y)	MT563379
<i>P. chienii</i> -198	F: GAGGGATACAGAAGCACAG R: TATGACAAACCCAAACGAG	(ATA)5	276–288	56	2.000	0.591	0.416	0.330**	0 (N)	MT563329
<i>P. chienii</i> -214	F: GACAACGGCAAAGGAGGAAT R: GCGATAGCCACCAAAGACAT	(ATA)6	319–322	58	3.000	0.000	0.329	0.290**	0 (N)	MT563334
<i>P. chienii</i> -216	F: TGCGGTTTCAGTAACAGTCCCTC R: TCCCCACCTCTTCCAG	(CTCCTG)5	439–452	58	6.000	0.657	0.513	0.408	–0.314 (N)	MT563335
<i>P. chienii</i> -233	F: TGTGTGAAAGGACAAGGCGT R: GCACCCTATTCACCCGAGAT	(ATGCAG)7	225–259	56	5.000	0.318	0.474	0.431	–0.0267 (N)	MT563337
<i>P. chienii</i> -267	F: CCCCTCATTGACAGTTTC R: AAGATAGTCGGGACACCAAG	(CTG)5	309–320	56	6.000	0.036	0.336	0.306**	0.3232 (Y)	MT563381
<i>P. chienii</i> -288	F: CACGCCACCATAGTTGT R: GGAGGAAGATGTCGTTGAAG	(AAGG)5	263–270	58	3.000	0.045	0.523	0.449**	0.2884 (Y)	MT563386
<i>P. chienii</i> -341	F: GACCTCTTACCAGCTGCGAG R: ACCACCGTTTCAGTTTCGT	(CCT)11	208–227	62	9.000	0.200	0.765	0.726**	0 (N)	MT563390
<i>P. chienii</i> -358	F: TAAGTGGCTGCTGCATCACA R: TACAGCAGCAGCAGAGCTTT	(TCC)5	249–251	58	3.000	0.318	0.360	0.302	0.1828 (N)	MT563392

Locus, locus name; Na, number of alleles; Ho, observed heterozygosity; He, expected heterozygosity; PIC, Polymorphism information content (\* $P < 0.05$ , significantly departures from Hardy–Weinberg equilibrium; \*\* $P < 0.01$ , extremely significantly departures from Hardy–Weinberg equilibrium); Null, the frequency of null allele (Y represents that the SSR locus may have a null allele, N represents that the SSR locus may not have null allele).

**TABLE 6 |** Genetic diversity parameters of four populations of *P. chienii*.

Population	N	Na	Ne	I	Ho	He	PPB (%)	$F_{IS}$
MS	31	3.450	1.997	0.702	0.384	0.389	90	0.031 <sup>ns</sup>
YS	30	3.500	1.918	0.670	0.360	0.357	85	0.008 <sup>ns</sup>
BJS	30	3.650	2.033	0.716	0.350	0.381	100	0.099**
ZJJ	19	2.850	1.759	0.602	0.272	0.340	85	0.227**
Species level	110	6.400	2.814	1.071	0.349	0.527	100	0.342

N, numbers of individuals; Na, number of alleles; Ne, number of effective alleles; Ho, observed heterozygosity; He, expected heterozygosity; I, Shannon's information index; PPB, percentage of polymorphic loci;  $F_{IS}$ , inbreeding coefficient. <sup>ns</sup> $P > 0.05$ , \*\* $P < 0.01$ .

di-nucleotide repeats, which might be explained by cytosine methylation inhibited gene transcription (Chen et al., 2015; Xing et al., 2017).

The length of the microsatellite was one of the main factors affecting its polymorphism. According to Temnykh et al. (2001), SSR polymorphism can be considered low, medium or high if length < 12 bp,  $12 \leq$  length < 20 bp, and length > 20 bp. In this study, SSR length between 12 and 20 bp (10638, 48.72%) accounted for the largest proportion. Therefore, it can be speculated that the microsatellites in *P. chienii* transcriptome may have a moderate level of polymorphism. In addition, the mutation rate of microsatellite is positively correlated with its length (Calabrese and Sainudiin, 2005). In *P. chienii*, six microsatellite motifs were found to have significant differences in length (Kruskal–Wallis rank sum test,  $P < 2.2e^{-16}$ ), and decreased with increasing motif size (Nemenyi test,  $P < 2.2e^{-16}$ ). Longer microsatellites were expected to have a high mutation rate because of more chances of replication slippage (Calabrese and Sainudiin, 2005). Therefore, tetra-, penta-, and hexa-nucleotide microsatellites might have higher mutation rates than those of the mono-, di-, and tri-nucleotide microsatellites.

## Microsatellite Distribution in Different Genic Regions

Studies have shown that microsatellites are non-randomly distributed across protein-coding regions, UTRs, and introns (Li et al., 2004). In this study, the number of microsatellites in UTRs was much higher than that in CDS regions. One possible reason was that microsatellites located in UTRs were subject to less selection pressure and evolutionary constraints, thus, the microsatellites located in this region was more prone to expansion. In contrast, due to the high selection pressure of the microsatellite in coding regions, the structure and function of the gene will be seriously damaged if “indel” mutation occurs in this region (Levinson and Gutman, 1987; Li et al., 2004). We also found that the microsatellites located in CDS regions were dominated by tri-nucleotides, while other SSR motifs accounted for a small proportion. This was common in many plants, such as *Phoenix dactylifera* L. (Zhao et al., 2013), *Tetraena mongolica* (Dang et al., 2020), and *Paeonia lactiflora* (Wan et al., 2020). Compared to other repeat types, tri-nucleotide is less likely to frame-shift mutations in coding regions (Metzgar et al., 2000).

Although some studies have shown that the microsatellites in coding regions were subject to higher evolutionary constraints and thus exhibited a shorter length. For example, Liu et al. (2016) found in a study on *Sargassum thunbergii* that microsatellites in UTR regions were much longer than those in CDS regions. However, our results showed that the microsatellites located in UTR regions had relatively short average length, which could be attributed to the large number of mono- and di-nucleotide microsatellites located in these regions.

## Potential Function of Unigenes Containing Microsatellites

Microsatellites derived from the transcribed sequence might be directly related to gene function and thus might play

an adaptive and evolutionary role in affecting gene products, inducing phenotypic changes, and regulating gene expression (Li et al., 2004). In order to reveal the potential functions of microsatellites-containing unigene sequences, functional annotation and classification of these unigenes were carried out. GO functional annotations showed that a large number of unigenes containing microsatellites were assigned to “cellular process,” “metabolic process,” “cell,” “cell part,” “catalytic activity,” and “binding” terms, suggesting they might be related to the basal metabolism and life activities of *P. chienii*. Similarly, KEGG and KOG analysis indicated that SSRs-containing unigenes had a wide range of biological functions and were involved in various aspects of *P. chienii* growth and development. In addition, GO enrichment analysis showed that unigenes containing microsatellites were significantly enriched in terms related to “transcription,” similar to the reports in *Onobrychis viciifolia* (Shen et al., 2019) and Siberian wildrye (Zhou et al., 2016). Additionally, we found that some unigenes containing microsatellites were associated with transcription factors. TFs (Transcription Factors) regulate a lot of biological processes, including plant defense, response to a range of abiotic and biotic stresses, and tolerance to disease and environmental stress (Singh et al., 2002). As an ancient relic plant, *P. chienii* might be disturbed by various adverse factors during the long evolutionary process. However, it still survived to the present day and has strong adaptability to the environment. From the point of view of molecular biology, the microsatellites identified in this study that are related to transcription factors might serve as important regulatory buttons for gene expression in the adaptive evolution of *P. chienii*. Likewise, KEGG enrichment analysis revealed that four pathways that were associated with gene expression regulation (ko03022, ko04140, ko04070, and ko04712) showed significant enrichment, indicating that they played an important role in the response to environmental stress.

Additionally, there may be some differences in the functions of microsatellites in different tissues. However, we only analyzed the distribution and function of microsatellites in *P. chienii* transcriptome as a whole. Ince et al. (2010b) studied microsatellite-containing gene expression profiles in 16 different tissues and at different developmental stages in pepper (*Capsicum annuum* L.), and found that tissue-specific genes contained more dinucleotide microsatellites, and housekeeping genes contained more trinucleotide microsatellites. Therefore, we can further compare the expression patterns of microsatellite-containing genes in four tissues of *P. chienii* in the future, so as to deepen the understanding of tissue-specific expression of microsatellites.

## Validation of Microsatellite Markers

The 375 pairs of primers were randomly selected for validating the amplification quality and polymorphisms of the microsatellites in the present study. Among of them, only 77 (20.53%) produced clear bands with expected size. This low amplification success rate (20.53%) was comparable to that reported in some conifer species, such as *Abies alba* (24%) (Postolache et al., 2014), *Torreya grandis* (29.25%) (Zeng et al., 2018), and *Pinus koraiensis* (32%) (Li et al., 2020). There were two main reasons for the low success rate of amplification in *P. chienii*.

On the one hand, there were a large number of pseudogenes in conifer genomes, which may hinder the amplification of microsatellite markers. On the other hand, the highly repetitive sequences in conifer genomes, such as transposons, might affect the amplification efficiency of microsatellites (Kovach et al., 2010). Besides, the presence of large introns or indels, assembly errors in the unigenes, and low specificity might cause the primer amplification failure (Bazzo et al., 2018; Niu et al., 2019).

## Genetic Diversity of *P. chienii* Populations

Genetic diversity is the product of long-term evolution of a species or population and is a necessary condition for the survival and sustainable development of populations (Zhang and Zhou, 2013). We validated the polymorphism of the 20 polymorphic microsatellite markers using four populations of *P. chienii*. A total of 128 alleles were detected at the 20 loci. Among of them, loci *P. chienii*-152 (with 16 alleles) and *P. chienii*-75 (with 14 alleles) showed a high level of polymorphism. Loci *P. chienii*-152 was highly homologous to an unknown protein of *Picea sitchensis*. Loci *P. chienii*-75 was associated with a gene encoding R2R3-MYB transcription factor, therefore, played an important regulatory role in *P. chienii* growth and development, as well as in response to abiotic and biotic stress. At the species level, there was a moderate level of genetic diversity ( $N_a = 6.4$ ,  $H_o = 0.349$ ,  $H_e = 0.527$ ,  $I = 1.071$ ) for *P. chienii*, which was similar to the previous reports of Kou et al. (2020) using nuclear loci ( $H_d = 0.5400$ ,  $\pi = 0.00265$ ), but higher than that of Su et al. (2009) by ISSR markers ( $h = 0.2118$ ,  $I = 0.2390$ ). In addition, compared with other conifer species such as *Larix principis-rupprechtii* Mayr ( $N_a = 3.850$ ,  $H_o = 0.487$ ,  $H_e = 0.490$ ) (Dong et al., 2018) and *Pinus koraiensis* ( $N_a = 6.45$ ,  $H_o = 0.299$ ,  $H_e = 0.311$ ) (Li et al., 2020), *P. chienii* has a slightly higher level of genetic diversity ( $N_a = 6.4$ ,  $H_o = 0.349$ ,  $H_e = 0.527$ ). The abundance of genetic variation at the species level can be attributed to internal and historical factors. As an ancient tertiary relic plant, *P. chienii* has accumulated a large number of genetic variations during the long-term evolution, so it is expected to maintain a high level of genetic diversity. Besides, the Central and Nanling regions of China were less affected by the Quaternary glacial and became the possible refugia of many relic plants (Chen et al., 2011). We speculated that these refugia might be beneficial to the preservation of genetic variation in *P. chienii*. At the population level, all the four populations of *P. chienii* showed a low level of genetic diversity, which could be explained by its population size and distribution characteristics. The wild populations of *P. chienii* were few and scattered distributions, which limited the gene flow among populations and reduced genetic diversity. In addition, the small population size might result in the bias of genetic diversity in this study. We will sample more individuals of *P. chienii* to further explore its genetic variation in the future.

## Transferability of Microsatellite Markers

In general, microsatellite markers developed from transcriptome data/ESTs showed a higher level of transferability in related species (Ince et al., 2010a, 2014; Yan et al., 2017). In the present

study, 14 microsatellite markers could successfully amplify in *A. argotaenia*, and 12 of them were polymorphic in all *A. argotaenia* individuals. The transferability ratio was 70%, which was higher than the reported in *Capsicum* (32.11%) (Ince et al., 2010c), and *Elymus* (49.11%) (Zhou et al., 2016), indicating that the markers developed in the present study had good potential for cross-species amplification and could be used in future genetic studies of other Taxaceae species.

## CONCLUSION

In this study, we identified and characterized the frequency, distribution, and function of microsatellites based on transcriptome data of *P. chienii*. The unigenes containing microsatellites had a wide range of biological functions, most of which were associated with basic metabolism, and a few might be involved in expression regulation of gene and response to environmental stress. These results provided new insights into the role of microsatellites in the transcriptome. In addition, twenty microsatellite markers were developed for *P. chienii* with high polymorphism and good transferability, which will facilitate the genetic study in *P. chienii* and other related species.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://www.ncbi.nlm.nih.gov/genbank/>, SRR11715801; <https://www.ncbi.nlm.nih.gov/genbank/>, SRR11715800; <https://www.ncbi.nlm.nih.gov/genbank/>, SRR11715799; <https://www.ncbi.nlm.nih.gov/genbank/>, SRR11715798. The EST sequences in this study have been submitted to GenBank with accession numbers MT563329–MT563392.

## AUTHOR CONTRIBUTIONS

YS and TW conceived and designed the study and revised the manuscript. RX and ZW performed the experiments, analyzed the data, and wrote the manuscript. All the authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the National Natural Science Foundation of China (32071781, 31670200, 31770587, and 31872670), the Natural Science Foundation of Guangdong Province, China (2016A030313320 and 2017A030313122), Science and Technology Planning Project of Guangdong Province, China (2017A030303007), Project of Department of Science and Technology of Shenzhen City, Guangdong, China (JCYJ20160425165447211, JCYJ20170413155402977, JCYJ20170818155249053, and JCYJ20190813172001780), and Science and Technology Planning Project of Guangzhou City, China (201804010389).

## ACKNOWLEDGMENTS

The authors thank Dr. Q. Fan of School of Life Sciences, Sun Yat-sen University for assistance with the collection of plant materials.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.574304/full#supplementary-material>

**Supplementary Figure 1** | Box plots of the length distribution of six microsatellite repeat motifs.

**Supplementary Figure 2** | Box plots of the length distribution of microsatellite loci located in different genic regions.

**Supplementary Figure 3** | Box plots of the length distribution of six microsatellite repeat motifs in different genic regions.

**Supplementary Table 1** | Sampling location information of four populations of *A. argotaenia*.

**Supplementary Table 2** | 124 KEGG pathways annotated for the unigenes containing microsatellites.

**Supplementary Table 3** | The results GO enrichment analysis for unigenes containing microsatellites.

**Supplementary Table 4** | The 77 primers pairs of *P. chienii* screened in this study.

**Supplementary Table 5** | Bottleneck analysis for four populations of *P. chienii*.

**Supplementary Table 6** | PCR amplification results of 12 pairs of polymorphic SSR primers in *Amentotaxus argotaenia*.

## REFERENCES

- Ahmadi, J., and Fotokian, M.-H. (2011). Identification and mapping of quantitative trait loci associated with salinity tolerance in rice (*Oryza Sativa*) using SSR markers. *Iran. J. Biotechnol.* 9, 21–30. doi: 10.5897/AJB2018.16661
- Alexa, A., and Rahnenfuhrer, J. (2010). TopGO: enrichment analysis for gene ontology. *R Pack. Vers.* 2:2010.
- Bazzo, B. R., de Carvalho, L. M., Carazzolle, M. F., Pereira, G. A. G., and Colombo, C. A. (2018). Development of novel EST-SSR markers in the macaúba palm (*Acrocomia aculeata*) using transcriptome sequencing and cross-species transferability in Arecaceae species. *BMC Plant Biol.* 18:276. doi: 10.1186/s12870-018-1509-9
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* 57, 289–300. doi: 10.2307/2346101
- Calabrese, P., and Sainudiin, R. (2005). “Models of Microsatellite Evolution”, In: *Statistical Methods in Molecular Evolution*. New York, NY: Springer, 289–305.
- Chen, D. M., Kang, H., and Liu, C. J. (2011). An overview on the potential quaternary glacial refugia of plants in China mainland. *Bull. Bot. Res.* 31, 623–632. doi: 10.7525/j.issn.1673-5102.2011.05.019
- 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 markers based on next-generation sequencing technology in *Neolitsea sericea* (Lauraceae) endemic to East Asian land-bridge islands. *Mol. Breed.* 35:187. doi: 10.1007/s11032-015-0379-1
- Chen, S. Y., Dong, M. L., Zhang, Y., Qi, S. Z., Liu, X. Z., Zhang, J. F., et al. (2020). Development and characterization of simple sequence repeat markers for, and genetic diversity analysis of *Liquidambar formosana*. *Forests* 11:203. doi: 10.3390/f11020203
- Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610
- Dang, Z. H., Huang, L., Jia, Y. Y., Lockhart, P. J., Fong, Y., and Tian, Y. Y. (2020). Identification of genic SSRs provide a perspective for studying environmental adaptation in the endemic shrub *Tetraena mongolica*. *Genes* 11:322. doi: 10.3390/genes11030322
- Dong, M. L., Wang, Z. W., He, Q. W., Zhao, J., Fan, Z. R., and Zhang, J. F. (2018). Development of EST-SSR markers in *Larix principis-rupprechtii* Mayr and evaluation of their polymorphism and cross-species amplification. *Trees* 32, 1559–1571. doi: 10.1007/s00468-018-1733-9
- Dutta, S., Kumawat, G., Singh, B. P., Gupta, D. K., Singh, S., Dogra, V., et al. (2011). Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea *Cajanus cajan* (L.) Millsp. *Bmc Plant Biol.* 11:17. doi: 10.1186/1471-2229-11-17
- Eujayl, I., Sledge, M. K., Wang, L., May, G. D., Chekhovskiy, K., Zwonitzer, J. C., et al. (2004). Medicago truncatula EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theor. Appl. Genet.* 108, 414–422. doi: 10.1007/s00122-003-1450-6
- Fu, L. G., and Jin, J. M. (1992). *Red List of Endangered Plants in China*, Vol. 1. Science Press.
- Fu, L. G., Li, N., and Mill, R. R. (1999). “Taxaceae,” in *Flora of China*, eds Z. Y. Wu and P. H. Raven (Beijing: Science Press), 89–98.
- Fu, L. M., Niu, B. F., Zhu, Z. W., Wu, S. T., and Li, W. Z. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565
- Gao, C. H., Ren, X. D., Mason, A. S., Li, J., Wang, W., Xiao, M. L., et al. (2013). Revisiting an important component of plant genomes: microsatellites. *Funct. Plant Biol.* 40, 645–661. doi: 10.1071/12325
- Garvin, D. F., McKenzie, N., Vogel, J. P., Mockler, T. C., Blankenheim, Z. J., Wright, J., et al. (2010). An SSR-based genetic linkage map of the model grass *Brachypodium distachyon*. *Genome* 53, 1–13. doi: 10.1139/g09-079
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, L., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883
- Gupta, P. K., and Varshney, R. K. (2000). The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113, 163–185. doi: 10.1023/A:1003910819967
- Hancock, J. M. (1995). The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* 41, doi: 1038-1047. 10.1007/BF00173185
- Ince, A. G., Karaca, M., and Elmasulu, S. Y. (2014). New microsatellite and CAPS-microsatellite markers for clarifying taxonomic and phylogenetic relationships within *Origanum* L. *Mol. Breed.* 34, 643–654. doi: 10.1007/s11032-014-0064-9
- Ince, A. G., Karaca, M., and Onus, A. N. (2010a). CAPS-microsatellites: use of CAPS method to convert non-polymorphic microsatellites into useful markers. *Mol. Breed.* 25, 491–499. doi: 10.1007/s11032-009-9347-y
- Ince, A. G., Karaca, M., and Onus, A. N. (2010b). Differential expression patterns of genes containing microsatellites in *Capsicum annum* L. *Mol. Breed.* 25, 645–658. doi: 10.1007/s11032-009-9362-z
- Ince, A. G., Karaca, M., and Onus, A. N. (2010c). Polymorphic microsatellite markers transferable across *Capsicum* species. *Plant Mol. Biol. Rep.* 28, 285–291. doi: 10.1007/s11105-009-0151-y
- Iseli, C., Jongeneel, C. V., and Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *ISMB* 1999, 138–148. doi: 10.0000/PMID10786296
- Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., and Dhawan, A. K. (2011). Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177, 309–334. doi: 10.1007/s10681-010-0286-9
- Kou, Y. X., Zhang, L., Fan, D. M., Cheng, S. M., Li, D. Z., Hodel, R. G., et al. (2020). Evolutionary history of a relict conifer, *Pseudotsuga chienii* (Taxaceae), in south-east China during the late Neogene: old lineage, young populations. *Ann. Bot.* 125, 105–117. doi: 10.1093/aob/mcz153

- Kovach, A., Wegrzyn, J. L., Parra, G., Holt, C., Bruening, G. E., Loopstra, C. A., et al. (2010). The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genom.* 11:420. doi: 10.1186/1471-2164-11-420
- Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion analysis of variance. *J. Am. Stat. 47*, 583–621. doi: 10.1080/01621459.1952.10483441
- Lalitha, S. (2000). Primer premier 5. *Biotech. Softw. Intern. Rep.* 1, 270–272. doi: 10.1089/152791600459894
- Levinson, G., and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4, 203–221. doi: 10.1093/oxfordjournals.molbev.a040442
- Li, X., Liu, X. T., Wei, J. T., Li, Y., Tigabu, M., and Zhao, X. Y. (2020). Development and transferability of EST-SSR markers for *Pinus koraiensis* from cold-stressed transcriptome through Illumina Sequencing. *Genes* 11:500. doi: 10.3390/genes11050500
- Li, X. Y., Lin, X. Y., Ruhsam, M., Chen, L., Wu, X. T., Wang, M. Q., et al. (2019). Development of microsatellite markers for the critically endangered conifer *Glyptostrobus pensilis* (Cupressaceae) using transcriptome data. *Silvae Genet.* 68, 41–44. doi: 10.2478/sg-2019-0007
- Li, Y. C., Korol, A. B., Fahima, T., Beiles, A., and Nevo, E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11, 2453–2465. doi: 10.1046/j.1365-294X.2002.01643.x
- Li, Y. C., Korol, A. B., Fahima, T., and Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 21, 991–1007. doi: 10.1093/molbev/msh073
- Liu, F. L., Hu, Z. M., Liu, W. H., Li, J. J., Wang, W. J., Liang, Z. R., et al. (2016). Distribution, function and evolution characterization of microsatellite in *Sargassum thunbergii* (Fucales, Phaeophyta) transcriptome and their application in marker development. *Sci. Rep.* 6:18947. doi: 10.1038/srep18947
- Metzgar, D., Bytof, J., and Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* 10, 72–80. doi: 10.1101/gr.10.1.72
- Morgante, M., Hanafey, M., and Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30, 194–200. doi: 10.1038/ng822
- Nagy, S., Poczai, P., Cernak, I., Gorji, A. M., Hegedus, G., and Taller, J. (2012). PICcalc: an online program to calculate polymorphic information content for molecular genetic studies. *Biochem. Genet.* 50, 670–672. doi: 10.1007/s10528-012-9509-1
- Nishio, Z., Ito, M., Tanio, M., Tabiki, T., and Yamauchi, H. (2008). Microsatellite marker-assisted selection of fusarium head blight resistance in backcrossed Japanese winter wheat lines. *Breed. Sci.* 58, 183–185. doi: 10.1270/jsbbs.58.183
- Niu, J., Wang, Y. J., Shi, Y. L., Wang, X. F., Sun, Z. M., Huang, K. Y., et al. (2019). Development of SSR markers via de novo transcriptome assembly in *Akebia trifoliata* (Thunb.) Koidz. *Genome* 62, 817–831. doi: 10.1139/gen-2019-0068
- Peakall, R., and Smouse, P. E. (2012). GenALEX 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460
- Piry, S., Luikart, G., and Cornuet, J. M. (1999). BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. *J. Hered.* 90, 502–503. doi: 10.1093/jhered/90.4.502
- Postolache, D., Leonarduzzi, C., Piotti, A., Spanu, I., Roig, A., Fady, B., et al. (2014). Transcriptome versus genomic microsatellite markers: highly informative multiplexes for genotyping *Abies alba* Mill. and congeneric species. *Plant Mol. Biol. Rep.* 32, 750–760. doi: 10.1007/s11105-013-0688-7
- Powell, W., Machray, G. C., and Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.* 1, 215–222. doi: 10.1016/1360-1385(96)86898-1
- Qiao, F., Cong, H. Q., Jiang, X. F., Wang, R. X., Yin, J. M., Qian, D., et al. (2014). De novo characterization of a *Cephalotaxus hainanensis* transcriptome and genes related to paclitaxel biosynthesis. *PLoS One* 9:e106900. doi: 10.1371/journal.pone.0106900
- Raymond, M., and Rousset, F. (1995). GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *J. Hered.* 86, 248–249. doi: 10.1093/oxfordjournals.jhered.a111573
- Sharma, H., Kumar, P., Singh, A., Aggarwal, K., Roy, J., Sharma, V., et al. (2020). Development of polymorphic EST-SSR markers and their applicability in genetic diversity evaluation in *Rhododendron arboreum*. *Mol. Biol. Rep.* 47, 2447–2457. doi: 10.1007/s11033-020-05300-1
- Shen, S. H., Chai, X. T., Zhou, Q., Luo, D., Wang, Y. R., and Liu, Z. P. (2019). Development of polymorphic EST-SSR markers and characterization of the autotetraploid genome of sainfoin (*Onobrychis viciifolia*). *Peer J.* 7:e6542. doi: 10.7717/peerj.6542
- Singh, K. B., Foley, R. C., and Onate-Sanchez, L. (2002). Transcription factors in plant defense and stress responses. *Curr. Opin. Plant Biol.* 5, 430–436. doi: 10.1016/s1369-5266(02)00289-3
- Su, Y. J., Wang, T., and Ouyang, P. Y. (2009). High genetic differentiation and variation as revealed by ISSR marker in *Pseudotsaxus chienii* (Taxaceae), an old rare conifer endemic to China. *Biochem. Syst. Ecol.* 37, 579–588. doi: 10.1016/j.bse.2009.10.005
- Su, Y. J., Wang, T., Zheng, B., Jiang, Y., Chen, G. P., Ouyang, P. Y., et al. (2005). Genetic differentiation of relictual populations of *Alsophila spinulosa* in southern China inferred from cpDNA *trnL-F* noncoding sequences. *Mol. Phylogenet. Evol.* 34, 323–333. doi: 10.1016/j.ympev.2004.10.016
- Taheri, S., Abdullah, T. L., Rafii, M., Harikrishna, J. A., Werbrouck, S. P., Teo, C. H., et al. (2019). De novo assembly of transcriptomes, mining, and development of novel EST-SSR markers in *Curcuma alismatifolia* (Zingiberaceae family) through Illumina sequencing. *Sci. Rep.* 9:3047. doi: 10.1038/s41598-019-39944-2
- Taheri, S., Abdullah, T. L., Yusop, M. R., Hanafi, M. M., Sahebi, M., Azizi, P., et al. (2018). Mining and Development of Novel SSR Markers Using Next Generation Sequencing (NGS) Data in Plants. *Molecules* 23:399. doi: 10.3390/molecules23020399
- Tautz, D., and Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucl. Acids Res.* 12, 4127–4138. doi: 10.1093/nar/12.10.4127
- Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and McCouch, S. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* 11, 1441–1452. doi: 10.1016/j.ces.2004.03.045
- Toth, G., Gaspari, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10, 967–981. doi: 10.1101/gr.10.7.967
- Van Oosterhout, C., Hutchinson, W. F., Wills, D. P. M., and Shipley, P. (2004). Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. *Mol. Ecol. Notes* 4, 535–538. doi: 10.1111/j.1471-8286.2004.00684.x
- Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* 23, 48–55. doi: 10.1016/j.tibtech.2004.11.005
- Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.* 39, 312–328. doi: 10.1590/1678-4685-gmb-2016-0027
- Wan, Y. L., Zhang, M., Hong, A. Y., Zhang, Y. X., and Liu, Y. (2020). Characteristics of microsatellites mined from transcriptome data and the development of novel markers in *Paeonia lactiflora*. *Genes* 11:214. doi: 10.3390/genes11020214
- Wang, Q. B., Zhang, L., and Zheng, P. J. (2015). Genetic diversity and evolutionary relationship analyses within and among *Raphanus* species using EST-SSR markers. *Mol. Breed.* 35:62. doi: 10.1007/s11032-015-0261-1
- Xiang, X. Y., Zhang, Z. X., Wang, Z. G., Zhang, X. P., and Wu, G. L. (2015). Transcriptome sequencing and development of EST-SSR markers in *Pinus dabeshanensis*, an endangered conifer endemic to China. *Mol. Breed.* 35:158. doi: 10.1007/s11032-015-0351-0
- Xin, D. W., Sun, J. Y., Wang, J. L., Jiang, H. W., Hu, G. H., Liu, C. Y., et al. (2012). Identification and characterization of SSRs from soybean (*Glycine max*) ESTs. *Mol. Biol. Rep.* 39, 9047–9057. doi: 10.1007/s11033-012-1776-8
- Xing, W., Liao, J. Y., Cai, M. Y., Xia, Q. F., Liu, Y., Zeng, W., et al. (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 Genet. Genom.* 13:53. doi: 10.1007/s11295-017-1135-y
- Yan, Z. Z., Wu, F., Luo, K., Zhao, Y. F., Yan, Q., Zhang, Y. F., et al. (2017). Cross-species transferability of EST-SSR markers developed from the transcriptome of *Melilotus* and their application to population genetics research. *Sci. Rep.* 7:17959. doi: 10.1038/s41598-017-18049-8

- Yang, X., Yu, M. J., Ding, B. Y., Xu, S. X., and Ye, L. X. (2005). Population structure and community characteristics of *Pseudotaxus chienii* in Fengyangshan National Natural Reserve. *Chin. J. Appl. Ecol.* 16, 1189–1194. doi: 10.3321/j.issn:1001-9332.2005.07.002
- Yu, G. C., Wang, L. G., Han, Y. Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omic* 16, 284–287. doi: 10.1089/omi.2011.0118
- Zane, L., Bargelloni, L., and Patarnello, T. (2002). Strategies for microsatellite isolation: a review. *Mol. Ecol.* 11, 1–16. doi: 10.1046/j.0962-1083.2001.01418.x
- Zeng, J., Chen, J., Kou, Y. X., and Wang, Y. J. (2018). Application of EST-SSR markers developed from the transcriptome of *Torreya grandis* (Taxaceae), a threatened nut-yielding conifer tree. *Peer J.* 6:e5606. doi: 10.7717/peerj.5606
- Zhang, D. Q., and Zhou, N. (2013). Genetic diversity and population structure of the endangered conifer *Taxus wallichiana* var. *mairei* (Taxaceae) revealed by Simple Sequence Repeat (SSR) markers. *Biochem. Syst. Ecol.* 49, 107–114. doi: 10.1016/j.bse.2013.03.030
- Zhang, L. D., Yuan, D. J., Yu, S. W., Li, Z. G., Cao, Y. F., Miao, Z. Q., et al. (2004). Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. *Bioinformatics* 20, 1081–1086. doi: 10.1093/bioinformatics/bth043
- Zhao, Y. L., Williams, R., Prakash, C. S., and He, G. H. (2013). Identification and characterization of gene-based SSR markers in date palm (*Phoenix dactylifera* L.). *BMC Plant Biol.* 12:237. doi: 10.1186/1471-2229-12-237
- Zhao, Y. Y., Zhu, X. C., Li, Z., Xu, W. B., Dong, J., Wei, H., et al. (2019). Genetic diversity and structure of Chinese grass shrimp, *Palaemonetes sinensis*, inferred from transcriptome-derived microsatellite markers. *BMC Genet.* 20:75. doi: 10.1186/s12863-019-0779-z
- Zhou, Q., Luo, D., Ma, L. C., Xie, W. G., Wang, Y., Wang, Y. R., et al. (2016). Development and cross-species transferability of EST-SSR markers in Siberian wildrye (*Elymus sibiricus* L.) using *Illumina* sequencing. *Sci. Rep.* 6:20549. doi: 10.1038/srep20549
- Zhou, S. F., Wang, C. R., Frazier, T. P., Yan, H. D., Chen, P. L., Chen, Z. H., et al. (2018). The first *Illumina*-based de novo transcriptome analysis and molecular marker development in Napier grass (*Pennisetum purpureum*). *Mol. Breed.* 38:95. doi: 10.1007/s11032-018-0852-8

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Xu, Wang, Su and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.