Pak. J. Bot., 55(5): 1841-1850, 2023. DOI: http://dx.doi.org/10.30848/PJB2023-5(18)

TRANSCRIPTOME ANALYSIS AND EST-SSR MARKER DEVELOPMENT OF MEDICINAL PLANT *CYCLOCARYA PALIURUS*

SHUAI MU[1], YU ZHANG[1], JIA XIANG ZHANG[1], MIN ZHENG[1], JIAN ZHONG WANG[1,3], MAN PING DING[2], MAROOF ALI[5*] AND XIAO HONG LI[1,4*]

[1]*College of Life Sciences, Anhui Normal University, Wuhu, 241000, People's Republic of China*
[2]*Anhui Xionglu Nursery Garden, Jixi County, Xuancheng, 242000, P. R. China*
[3]*The Key Laboratory of Biotic Environment and Ecological Safety in Anhui Province, Wuhu, 241000, P. R. China*
[4]*The Key Laboratory of Conservation and Employment of Biological Resources of Anhui, Wuhu, 241000, P. R. China*
[5]*Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Yunnan, 666100, P. R. China*
[*]*Corresponding author's email: lxh79668@ahnu.edu.cn; marufturi059@gmail.com*

Abstract

As people's living standards improve, the Trio H's (Hyperglycemia, Hypertension, Hyperlipidemia) chronic diseases have seriously threatened the health of middle-aged and older people. In recent years, many researchers have become concerned about the development of specific drugs for Trio H's people. Because the leaf and bud teas of *Cyclocarya paliurus* (Juglandaceae) can relieve Trio H's syndrome, it is considered an ideal and potential medicinal tree. Concerning the unclear genetic background of *C. paliurus,* molecular markers applicable to the population level are still in need. In this study, using the BGISEQ-500 platform, we obtained about 108,003 unigenes by mining the leaf cDNA library of *C. paliurus*. Among them, 86,366 unigenes (79.97%) were successfully annotated, referring to the public protein database utilizing BLASTX alignment. 77,441 (Nr), 31,223 (GO), 60,160 (KOG), and 60,580 (KEGG) unigenes were aligned to NCBI databases, respectively. Besides, 27,960 SSRs were excavated and located from 21,517 unigenes, with an average frequency of 0.228 SSR/1 Kb, screened from 60 selected primer pairs, and 13 microsatellite primer pairs were found to be polymorphic and stable for *C. paliurus.* To confirm the validity of these thirteen primer pairs, we further screened 33 individuals derived from three locations (POP-JX, POP-HN, and POP-AH). The result showed that 51 variant loci were detected, with an average of 6.05 polymorphic loci per prime pair. The average genetic diversity index at the population level exhibited higher polymorphic (Ho:0.721; He:0.700; PIC:0.780). The reference transcriptome can facilitate functional genomic research of *C. paliurus*, and the new EST-SSRs will be useful for further population genetics study of *C. paliurus*.

Key words: *Cyclocarya paliurus*, *De novo* assembly, EST-SSR, Transcriptome.

Introduction

With the influence of living habits and the environment, Trio H's diseases, including hyperglycemia, hyperlipidemia, and hypertension, are the pervasive factors that afflict and threaten people's health following cancer and blood clots. Research for effective distinctive drugs for Trio H's diseases has become a hot spot for researchers and medical workers (Yang *et al*., 2016). Plant secondary metabolites, such as polysaccharides, flavones, and saponins, are generally accepted as ideal sources for medicines concerning few side effects (Shinwari *et al*., 2018; Seca *et al*., 2019; Khan *et al*., 2019; Ovais *et al*., 2019; Najeebullah *et al*., 2020; Jan *et al*., 2021). *Cyclocarya paliurus* (Batalin) Iljinskaya, a monotypic species of Juglandaceae, is an ideal resource tree for the treatment of Trio H's diseases syndromes. The National Health and Family Planning Commission of the People's Republic of China have approved *C. paliurus* as a new food resource since 2013 (Xie *et al*., 2016). *C. paliurus* leaf includes numerous beneficial secondary metabolites, like polysaccharides, flavones, and saponins, which can effectively relieve symptoms such as hyperlipidemia and hyperglycemia. Therefore, it is widely popular as a healthy tea in China (Liu *et al*., 2018). In addition, as a tertiary relic plant, *C. paliurus* harbors substantial systematic and evolution research significance (Mao *et al*., 2016). However, the seeds of *C. paliurus* remain deeply dormant for no less than two years in the wild population, which makes the natural regeneration of this tree much more difficult (Fang *et al*., 2006). The shortage of wild resources and the drawbacks of cutting seedlings of *C. paliurus* limit the promotion in the application market. Hence, it is essential to describe its genetic diversity and metabolic mechanism to study its micromorphological traits in detail following the previous work (Ali *et al*., 2020; Ali *et al*., 2021).

With regard to molecular research, essential preliminary works of *C. paliurus* have been accumulated. NCBI databases contain a total of 263 DNA/RNA sequences (Up to December 25, 2021, https://www.ncbi.nlm.nih.gov/nuccore/?term=Cyclocarya%20paliurus), including 28 microsatellite sequences (Fan *et al*., 2013). Xu *et al.*, (2016) screened 11,247 putative microsatellite loci by transcriptome sequencing but have not tested them at the population level. Kou *et al.*, (2016) used two chloroplast genes and one nuclear gene to illustrate the importance of *C. paliurus* as a relict plant in inferring its evolutionary history. Considering the potential demands of *C. paliurus*, more effective molecular markers applicable to the population level are still in need.

Next-generation sequencing technology is now widely used in many areas of genetic research. For example, the RNA-seq technology makes it convenient for researchers to develop more microsatellite loci for non-model organisms (Metzker *et al*., 2010; Zhu *et al*., 2022). RNA-seq technology utilizes only part of the gene expression to mine SSRs, also known as EST (expressed

sequence tag)-SSRs. Compared with genomic SSR (gSSR), EST-SSR polymorphism may be lower but may prove superior to gSSR for diversity analysis and portability (Gupta *et al*., 2003). Furthermore, RNA-seq transcriptome sequencing can provide new insights into undeveloped medicinal trees' metabolic pathways and contribute to the discovery of molecules associated with drug development (Bae *et al*., 2018).

In this study, combined with NCBI gene Nr and GO analysis, KOG and KEGG metabolic pathway search, we aimed to obtain a functionally annotated reference transcriptome of *C. paliurus* using the BGISEQ-500 platform (BGI, Shenzhen, China). Due to the abundant exploitable SSR motifs in functional genes, we designed primers and detected polymorphism and availability of these markers in three populations of *C. paliurus.* Our work will contribute to understanding the genetic diversity, population history, and metabolic pathways of *C. paliurus* and provide valuable resources for future functional genomic studies.

Materials and Methods

Plant material collection and gDNA extraction: The sample collection site is in Xionglu Nursery Garden, JiXi County, Anhui Province (30°10' N, 118°87' E). 33 individuals of *C. paliurus* cultivated with seeds from Xiushui County in JiangXi Province (POP-JX, n=11), Yi County in AnHui Province (POP-AH, n=11), and Zhangjiajie National Park in HuNan Province (POP-HN, n=11) were chosen for later population diversity screening. In May 2019, 2-3 young leaves per individual at the stem tip of *C. paliurus* were collected and fast-dried in silica gel. The modified CTAB scheme (Porebski *et al*., 1997) was used to isolate the genomic DNA.

RNA isolation and construction of cDNA library: Fresh leaves from two individuals, derived from POP-AH and POP-HN, were collected and stored in the liquid nitrogen outdoors and then preserved at -80°C refrigerator for RNA isolation. Following the CTAB scheme (Jordon-Thaden *et al*., 2015), the total RNA required for cDNA library construction was extracted from the frozen leaves. DNaseI (Takara, Japan) was added to eliminate the potential DNA pollution during RNA isolation. Determine the RNA segment size and concentration by Agilent 2100 bioanalyzer (Agilent DNA 1000 reagent).

After obtaining total RNA, perform mRNA enrichment and rRNA consumption. Sera-mag Magnetic Oligo (dT) beads enriched poly (A) mRNA. The obtained RNA was further randomly cut into short strains using fragment buffer. Then, reverse transcriptase and random primers amplify the first strand of cDNA. After synthesizing the second-strand cDNA, the double-stranded cDNA was obtained. Finally, the double-stranded DNA was ligated to the adaptor, and the product was amplified by PCR using specific primers. The PCR product was thermally denatured to single-stranded, and then a single-stranded circular DNA library was generated through bridge primers.

Assembly and annotation of unigenes: We use the BGISEQ-500 platform (BGI in Shenzhen, China) to obtain raw reads from the cDNA library. The raw reads have been filtered by deleting low-quality reads containing adaptors and reads with more than 5% unknown nucleotides (N). The original clean reads were used for the *de novo* assembly of the Trinity program. Cluster the assembled transcripts and remove redundancy to obtain unigenes. The raw sequencing data were submitted to the NCBI database (accession numbers: SRR12107134, SRR12107135). We matched the unigenes to NCBI databases using the BLASTX alignment search tool (E-value < 10^{-5}). Finally, we got the functional annotations of unigenes.

EST-SSR mining: SSRs were found and located in unigenes by Microsatellite (MISA) software (http://pgrc.ipk-gatersleben.de/misa) (Zalapa *et al*., 2012). Unigene repeat motifs include mono-, di-, tri-, tetra-, penta-, and hexanucleotide. The minimum number of repeats for these motifs was set to 12, 6, 5, 5, 4, and 4, respectively. The maximum distance between two SSRs of a compound microsatellite was 100bp, and the distance between incomplete SSRs markers was set to 5 bp.

Primer designing and PCR optimization: We synthesized 60 pairs of primers and verified them with a 15 μl PCR reaction system: 1.2 μL dNTP mixture (2.5 mM); 1.2 μL MgCl2 (25 μM); 1.5 μL 10×PCR buffer (Mg2 + Free); 0.1 μL TaKaRa Taq (5U·μL−1); 0.8 μL genomic DNA (2.5 ng); 0.1 μL primer A (10 μM); 0.3 μL primer B (10 μM); 0.3 μL primer C (10 μM); 9.5 μL ddH_2O. Primer A is fluorescently-labeled (6-FAM, HEX, TAMRA) M13-tailed (5'-TGTAAAACGACGGCCAGT-3') primers (Schuelke *et al*., 2000). The primer with the M13 tail at the 5' end is primer B. Primer C refers to the reverse primer. The amplification program consists of two steps. The first step was placed at 94°C for 5 min; followed by 35 cycles of 30 s at 94°C, 40 s at corresponding optimum annealing temperature (54–66°C), and 45 s extension at 72°C. The second step was 8 cycles of 30 s at 94°C, 54°C for 40 s, and 45 s extension at 72°C, with a final extension of 72°C for 10 min. The size of PCR products was determined under a UV lamp after 1% agarose gel electrophoresis.

Polymorphism and availability test: To test the polymorphism and availability of these EST-SSR primers, we collected 33 samples of *C. paliurus* cultivated in JiXi County, Anhui Province (30°10' N, 118°87' E). These sample seeds were collected from Jiangxi Province (POP-JX, n=11), Hunan Province (POP-HN, n=11), Anhui Province (POP-AH, n=11), respectively. The leaves were dried with silica gel, shaded, and stored at a temperature of 4℃. Since leaves of *C. paliurus* are rich in polysaccharides, we used the improved CTAB method to extract genomic DNA. Each genomic DNA was amplified using the above PCR reactions and procedure. The PCR products were sent to Sangon Biotechnology (Shanghai, China) for genotyping.

SSRs data analysis: SSRs raw data were achieved using GeneMarker software (version 2.2.0). GenAlEx 6.0 (Peakall *et al*., 2006) calculated heterozygosity (H_o), expected heterozygosity (H_e), shannon information index (I), and inbreeding coefficient (F_{is}). To check and estimate null alleles according to the "Brookfield 1" method, Micro-Checker version 2.2.3 was used (Excoffier *et al*., 2010). The online tool GENEPOP (http://kimura.univ-montp2.fr/~rousset/Genepop.htm) was used to test the significant deviation of Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD). Use POPGENE software (version 1.32) (Yeh *et al*., 1999) and PIC calculator (https://www.liverpool.ac.uk/~kempsj/pic.html) (Varshney *et al*., 2002) to estimate polymorphic information content (PIC). The sequential Bonferroni procedure (Rice *et al*., 1989) was performed to correct all significance values.

Results and Discussion

Sequencing and *de novo* assembly: *C. paliurus* is an ideal potential medicinal plant. Current research mainly focus on the bioactive extracts of *C. paliurus* leaves (Jiang *et al*., 2006). However, little research has been done on genetics of *C. paliurus*, and until now, the EST sequence resource available for *C. paliurus* is still limited. In comparison to Illumina (San Diego, CA, USA) platforms, some researchers encouraged more attempts to conduct transcriptome analysis using BGISEQ-500 because it has relatively high throughput and generates longer reads than Illumina (Zhu *et al*., 2018).

We obtained a total of 181.48 million raw reads using the BGISEQ-500 platform. The clean reads, generated after removing adaptor contamination, equivocal and inferior quality reads, summed up to 174.75 million which were subjected to *de novo* assembly by Trinity. Using Tgicl clustering and de-redundancy of assembled transcripts were conducted and obtained a total of 108,003 unigenes. The total length, average length, and N50 length of all assembled unigenes were 122,520,960 bp, 1,134 bp, and 1,846 bp, respectively (Table 1). Length distribution of the assembled unigenes was shown in (Fig. 1). The number of unigenes in the range of 200-500 bp, 500-1000 bp, and greater than 1,000 bp were 40,948 (37.91%), 21,280 (19.70%), 45,775 (42.38%), respectively. N50 of *C. paliurus* (1,846 bp) is higher than that of *Salix babylonica* (1315 bp) (Tian *et al*., 2019) and *Litsea cubeba* (1053 bp) (Han *et al*., 2013). Longer reads and larger N50 indicated that the generated transcripts in our study were qualified and could be used for further functional annotation.

Annotation of functional unigenes: Due to the lack of genetic or genomic information of *C. paliurus*, it is difficult to estimate the number of genes and the level of transcript coverage. To identify the putative or possible functions of *C. paliurus*, the BLASTX alignment (E-value < 10^{-5}) was carried out to search the number of unigenes annotated to the public NCBI functional databases including Nr, Go, KOG, KEGG, Swiss-Prot and Pfarn. As a result, 77,441 (Nr), 31,223 (GO), 60,160 (KOG), 60,580 (KEGG), 54,690 (Swiss-Prot) and 56140 (Pfarn) were obtained, respectively (Table 1). A total of 86,366 (79.97%) unigenes obtained functional annotations successfully, which was higher than *Rhododendron* (Xing *et al*., 2017).

Results of Nr annotation of *C. paliurus* were shown in (Fig. 2). 77,441 unigenes were successfully annotated to 503 species in the alignment with Nr protein database in NCBI. *Juglans regia* (75.8%) showed high gene similarity to *C. paliurus*, followed by *Hordeum vulgare* subsp. *Vulgare* (2.05%), *Coccomyxa subellipsoidea* C-169 (1.28%), *Vitis vinifera* (1.04%), *Ziziphus jujuba* (0.73%), indicating closer homology with *J. regia*, owing to their systematic relationships in the same family Juglandaceae.

Table 1. Summary of transcriptome statistics and functional annotations.

data type	number
Total number of raw reads	181.48 million
Total number of clean reads	174.75 million
Number of non-redundant unigenes	108,003
Number of unigenes between 200 bp and 500 bp in length	40,948 (37.91%)
Number of unigenes between 500 bp and 1,000 bp in length	21,280 (19.70%)
Number of unigenes greater than 1,000 bp in lengt	45,775 (42.38%)
N50 length (bp)	1,846
Total length (bp)	122,520,960
Maximum length (bp)	12,783
Minimum length (bp)	297
Average length (bp)	1,134
Number of unigenes in the NR database	77,441 (71.70%)
Number of unigenes in the KOG database	60160 (55.7%)
Number of unigenes in the Swiss-Prot database	54,690 (50.64%)
Number of unigenes in the Pfam database	56,140 (51.98%)
Number of unigenes in the GO database	31,223 (28.91%)
Number of unigenes in the KEGG database	60,580 (56.09%)
Number of unigenes annotated in at least one database	25,113 (23.25%)

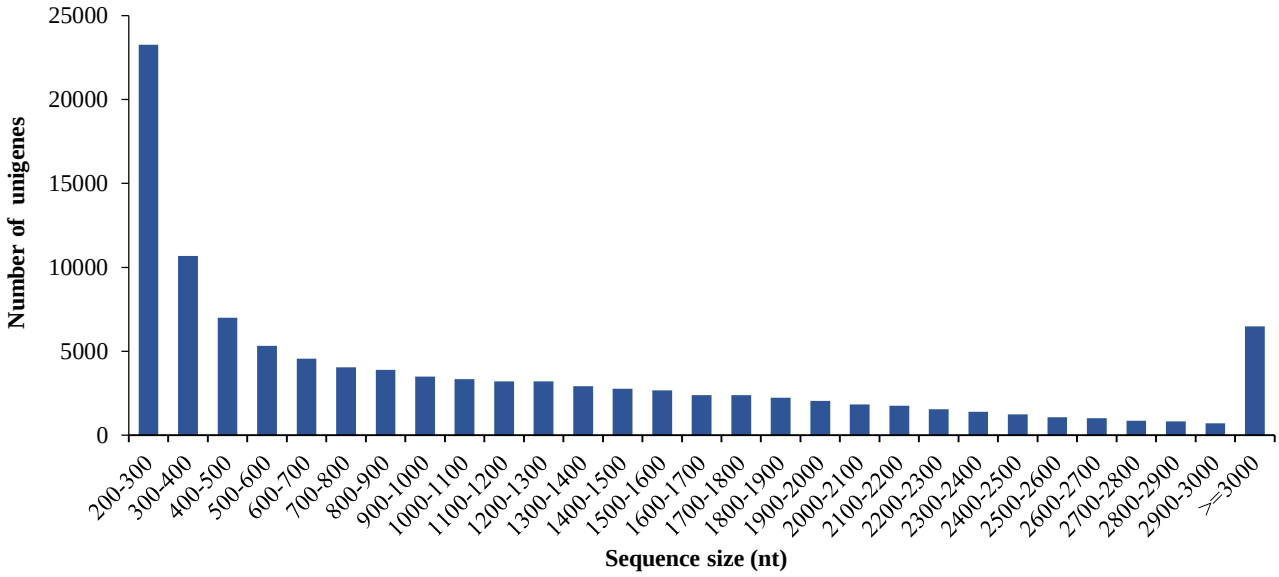

Fig. 1. The number of unigenes of different lengths.

To obtain a broader and deeper understanding of these unigenes' functions, the unigenes annotated successfully to the Nr database were assigned to the GO database. 31,223 unigenes, successfully annotated to the GO database, can be divided into three categories (Fig. 3), namely biological process, cellular component and molecular function. The dominant ontology was the 'Molecular function' (35,326, 44%), followed by the 'Cellular component' (23,714, 29%) and 'Biological process' (21,625, 27%) ontologies. Within the molecular category, the subgroup assignment of GO terms focused on 'binding', 'catalytic activity. For 'Cellular process', 'cell', 'membrane part' and 'organelle' were primarily assigned. In addition, 'cellular process' (12.5%), 'biological regulation' (3.7%), 'metabolic process (3.2%), and localization were the top subgroup in 'Biological processe'. These biological processes may reveal where the bioactive components of *C. paliurus* derive from. A total of 15,021 (18.6%) and 14,894 (18.5%) unigenes were annotated for molecular function, indicating that these genes may regulate biologically active components' production.

As indicated in (Fig. 4), functional prediction and classification of all unigenes were performed using the KOG database. A total of 60, 160 unigenes (55.7%) were annotated in KOG, which were grouped into 25 functional categories. Among those classifications, the category of 'general function prediction only' occupied the top number of unigenes. Other categories, such as 'carbohydrate transport and metabolism', 'secondary metabolites biosynthesis, transport and catabolism', and'function unknown', exhibited valuable mining information in nutritional or medicinal metabolites.

KEGG assignments were also conducted to classify the functions of the predicted *C. paliurus* genes, which is widely used as reference canonical database. A total of 132 KEGG pathways were identified, involving 60,580 unigenes (Fig. 5). KEGG pathways were grouped into five main categories. The biggest category was metabolism (38,332; 63%), followed by genetic information processing (13,600; 22%), environmental information processing (3,594; 6%), cellular processes (2,964; 5%), and organismal systems (2,218; 4%). We found 18 pathways involved in metabolite synthesis (Fig. 6). Most of the unigenes were rich in the phenylpropanoid biosynthesis pathway (ko00940, 671 unigenes). Terpenoid backbone biosynthesis pathways (ko00900, 90 unigenes), monoterpenoid biosynthesis pathways (ko00902, 15 unigenes), diterpenoid biosynthesis pathways (ko00904, 80 unigenes), and sesquiterpenoid and triterpenoid biosynthesis pathways (ko00909, 39 unigenes) were all related to terpenoids synthesis. This annotated information would be useful for metabolic engineering.

Based on the above results, a lot of useful unigenes of *C. paliurus* were successfully matched to the known proteins in the public NCBI databases, implying that the Illumina-based sequencing technique' convinces and advantages for the species without the known genome information. With regard to the annotated number, our data yielded an extensive and large proportion of the diverse genes expressed in *C. paliurus*. Because detailed functional information is essential to an overall understanding of the gene expression profiles, these unigenes were assigned a putative gene or protein name descriptions and categorized with GO terms and other metabolic pathways, which could provide a valuable resource for investigating specific processes, functions and pathways and will contribute to the identification of novel genes involved in the pathways of secondary metabolite biosynthesis.

Characteristics of the SSRs: We identified SSRs from 108,003 unigenes using Microsatellite software (MISA, http//:www.pgrc.ipk-gatersleben.de/misa). 27,960 SSRs were detected in 21,517 unigenes. The frequency of SSR was 0.228 SSR/1 Kb, meaning one SSR can be found every 4.39 Kb on average, i.e., 19.9% EST sequences possessing SSR. The frequencies of microsatellites observed in this context were higher than 2.65%-16.82%, a range reported earlier for 49 dicotyledonous species (Kumpatla *et al*., 2005). More than 1 SSR was found in 4805 unigenes, and 2,256 SSRs were compounded. Among the identified SSRs, there were 229 motifs. Dinucleotide repeat sequences were the most dominant motif (14,393, 51.48% of all SSRs), followed by

mono- (7,822, 27.98%), tri- (4,358, 15.59%), hexa- (555, 1.98%), penta- (468, 1.67%), and tetra- (364, 1.30%) nucleotide motifs (Fig. 7). As shown in (Fig. 8) the most abundant motifs of mono-, tri-, hexa-, penta-, and tetra-nucleotide were A/T (7,046, 25.20% of all SSRs), AG/CT (11,410, 44.81%), AAG/CTT (1,394, 4.99%), AAAT/ATTT (71, 0.25%), AAAAG/CTTTT (94, 0.34%), AGAGGG/CCCTCT (40, 0.14%), respectively. Most repeat types of dinucleotide and trinucleotide are AG/CT, AAG/CTT, respectively. Furthermore, the length range of SSRs was 12-78 bp (Fig. 9). The most common lengths were 12-16 bp. Lengths greater than 36 bp were rare. The number of microsatellites decreases as the length of SSRs increases. The CG/CG repeat sequences were also rare in this context, which may be related to the CpG sequences' methylation (Blanca *et al*., 2011). The previous reports showed that the trinucleotide repeats were the most common type (Varshney *et al*., 2005), but this was inconsistent with our result and Xu's study (2016). The most abundant repeats were dinucleotide, followed by trinucleotide.

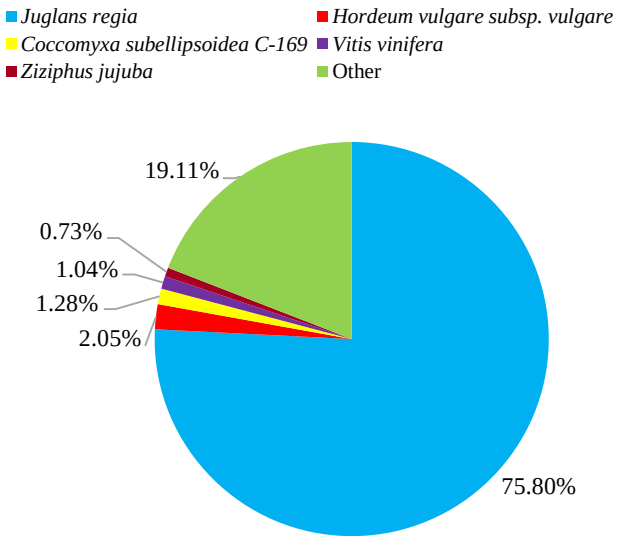

Fig. 2. Nr annotation of *C. paliurus*.

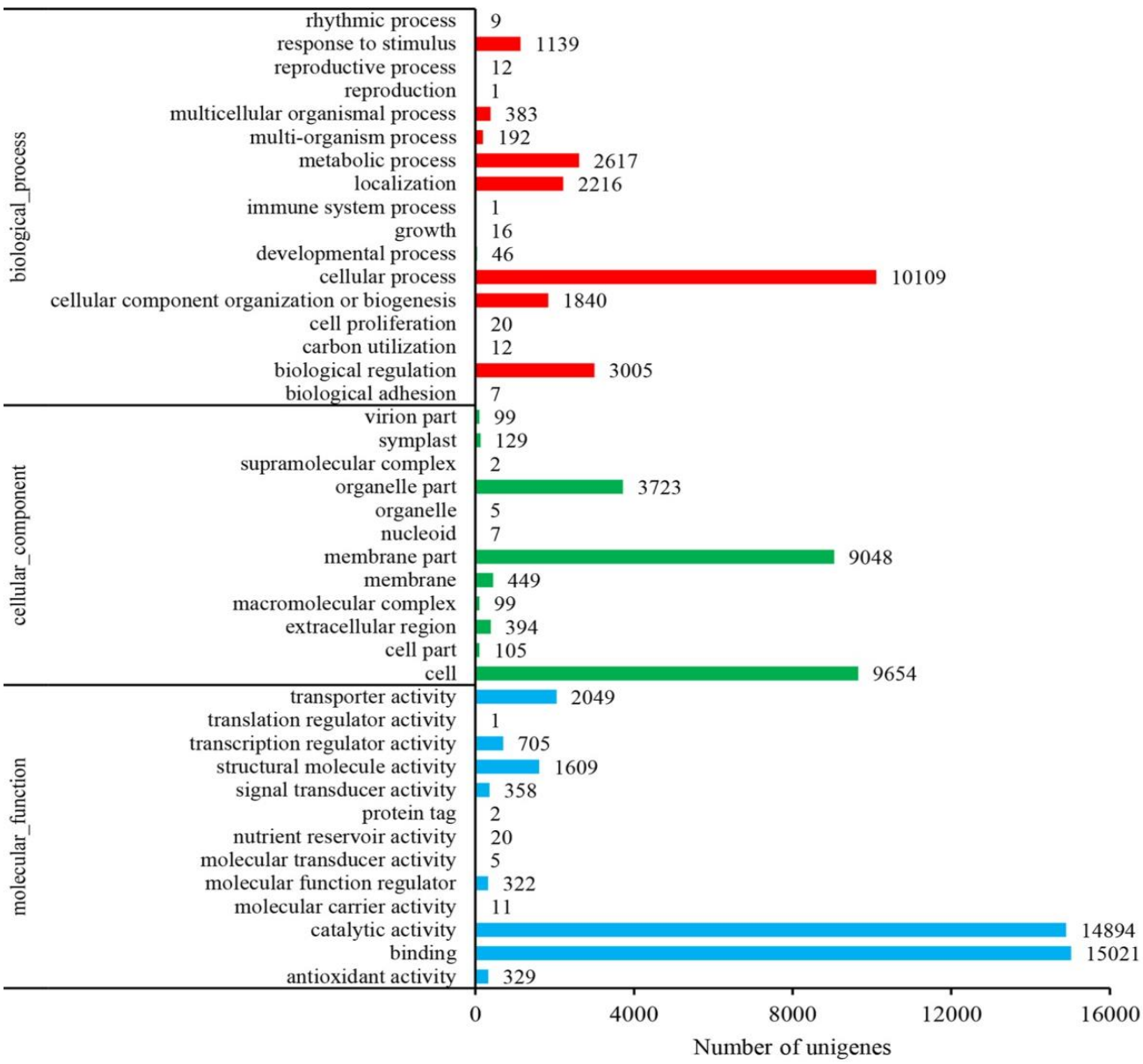

Fig. 3. GO annotation of *C. paliurus*.

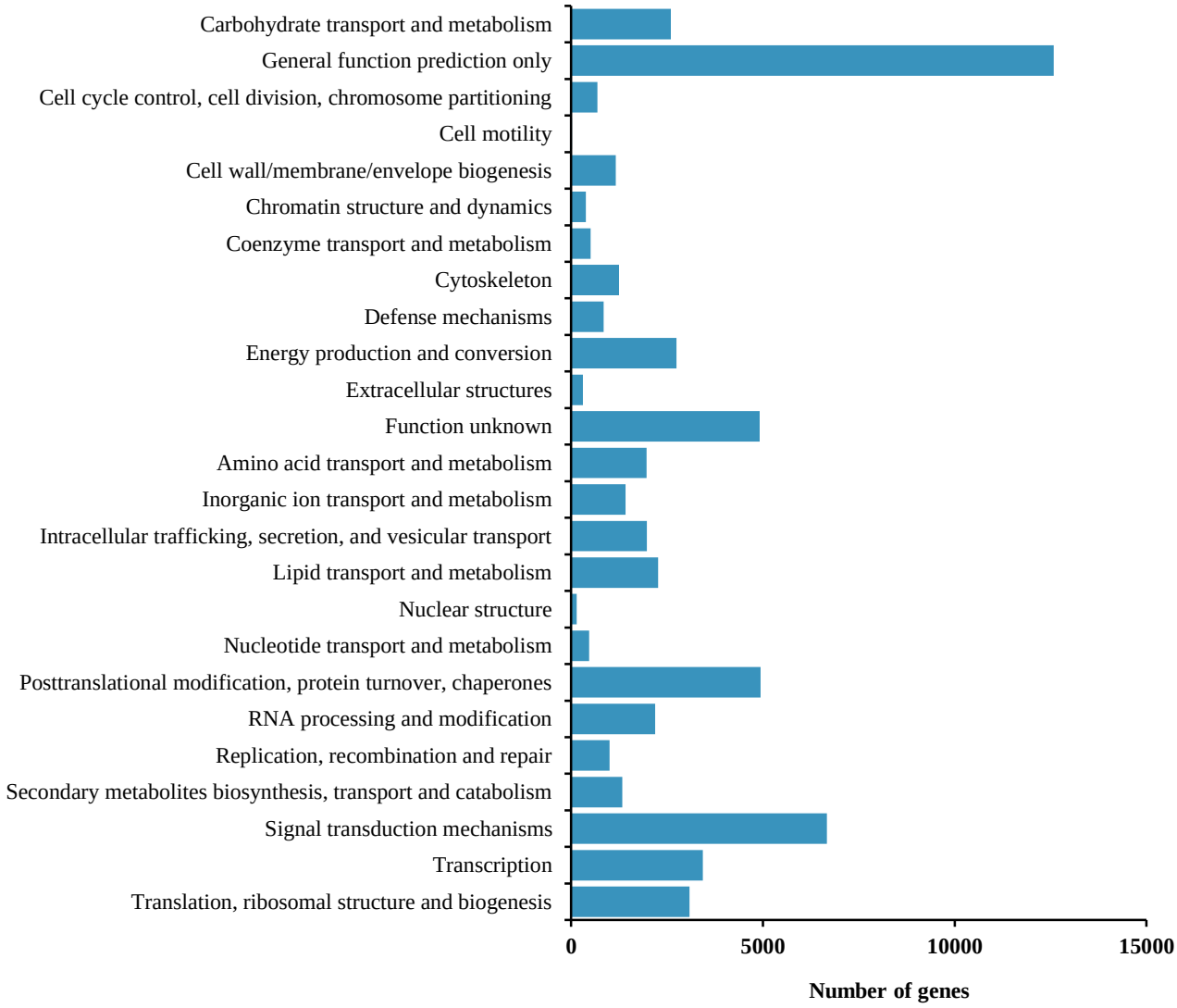

Fig. 4. KOG pathways of *C. paliurus.*

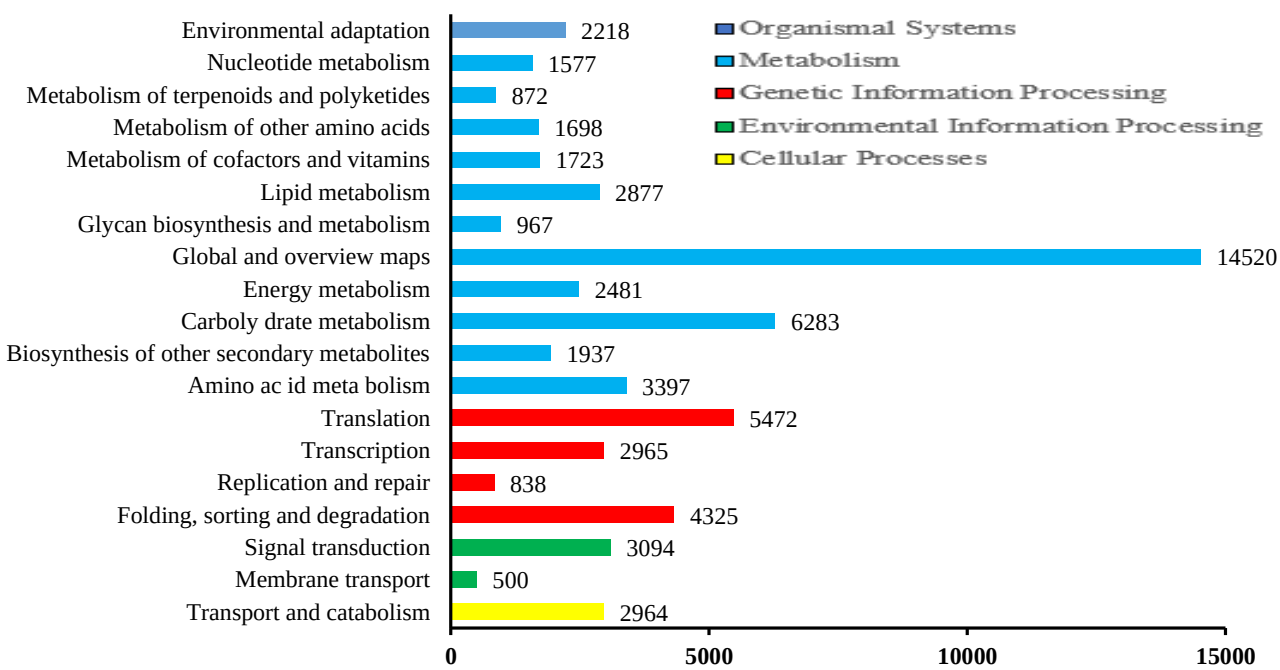

Fig. 5. KEGG pathways of *C. paliurus*.

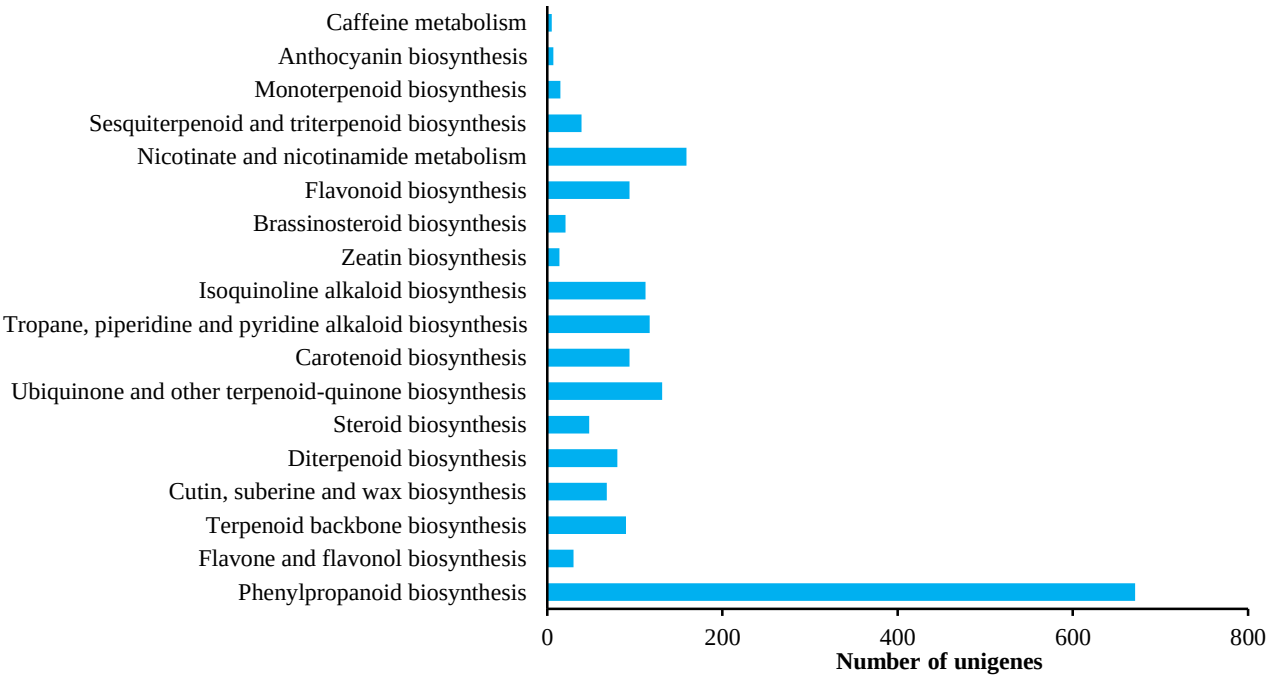

Fig. 6. Number of unigenes of *C. paliurus* involved in metabolite synthesis.

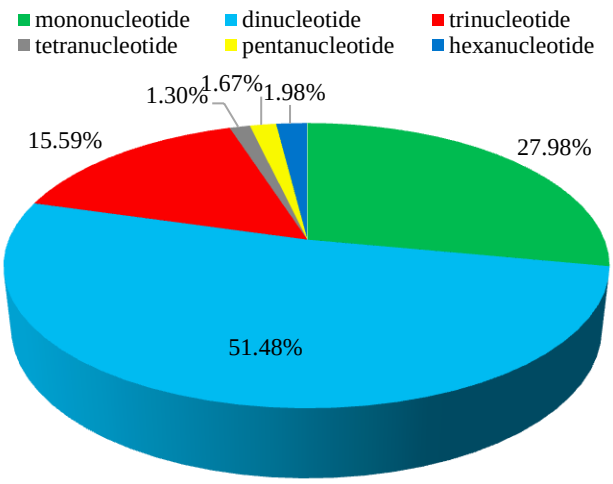

Fig. 7. Proportion of SSR different motifs of *C. paliurus.*

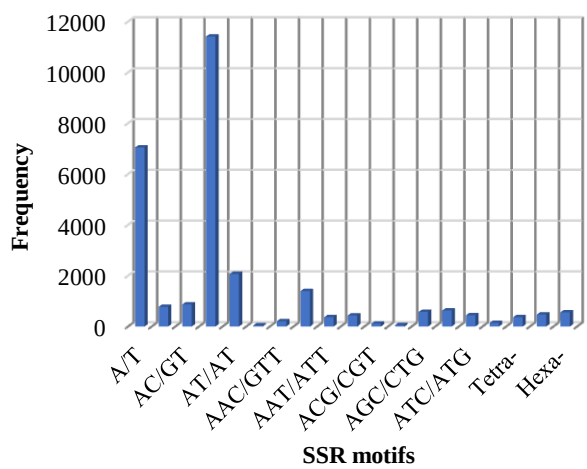

Fig.8. The frequency of different motifs of *C. paliurus.*

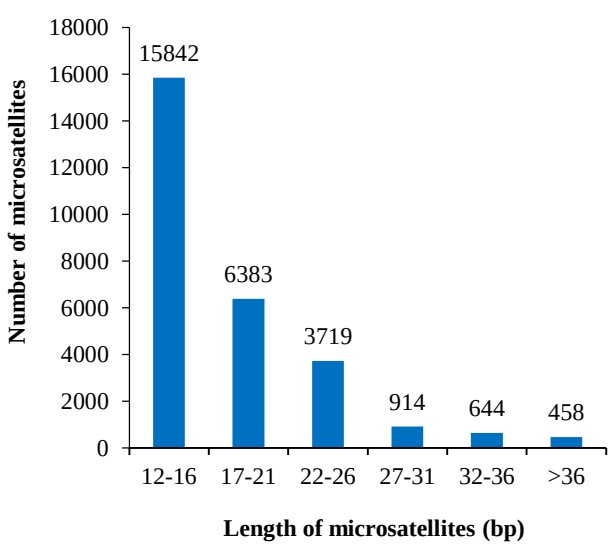

Fig. 9. The number of different length of microsatellites.

EST-SSRs polymorphism and availability test: In preliminary experiments, a total of 60 primers were validated for population amplification, among them 36 pairs of primers failed to amplify, and 11 pairs of primers amplified a single band, which was considered not helpful for genetic research, 13 pairs of primers showed high polymorphism (Table 2). The selective proportion was similar to *Populus euphratica* (Du *et al*., 2013). The PIC values of these primers were more significant than 0.5. We examined the availability of these 13 pairs of primers among three populations of *C. paliurus* (Table 3). No significant linkage disequilibrium was found in all individuals of the three populations. The range of alleles was 2-10 (mean: 6.05 SD 0.42). Expected heterozygosity (H_e) of POP-JX, POP-HN, and POP-AH populations was 0.492-0.872, 0.562-0.847, and 0.397-0.860, respectively. Observed heterozygosity (H_o) of POP-JX, POP-HN, and POP-AH populations were 0-1, 0.09-1, and 0.364-1, respectively (Table 3).

Significant deviations from HWE (Hardy–Weinberg equilibrium) were found among primers CP005, CP008, and CP009 for the POP-JX population and CP005 for the POP-HN population. Significant inbreeding and the presence of null alleles may contribute to the partial deviations from HWE. Here, these 10 EST-SSRs (CP001, CP002, CP003, CP004, CP006, CP007, CP010, CP011, CP012, CP013), due to their high information content and low frequency of null alleles, are recommended to be the optimal candidate primers for further population genetic diversity and inferring the evolutionary history of *C. paliurus*. Meanwhile, each locus had an average of 6.05 alleles (Table 3), which is more than that previous report (average number of alleles: 3.3) for *C. paliurus* (Fan *et al*., 2013). The average Shannon's Information Index (I) of the three populations was 1.51, indicating that these markers had high polymorphism and could better depict population genetic of *C. paliurus*.

Table 2. Characteristics for thirteen polymorphic microsatellite loci and primer pairs developed for *Cyclocarya paliurus.*

Locus	Repeat		Primer sequences (5'-3')	*Ta*(°C)	Range (bp)	PIC	GenBank accession no.
CP001	$(AG)_{12}$	F:	GCTTGTTGATAATTGGAGAGCTG	55.8	166-215	0.75	MT601844
		R:	TTGTCATGTCCGTCTAGTCTTCA				
CP002	$(AG)_{10}$	F:	AGTGAACAAATCAAGCTCGTGAC	55.8	140-206	0.86	MT601843
		R:	ATTCTTGTATCTGCGGGAAAAAT				
CP003	$(AG)_{11}$	F:	GAAAAGGAGGGAAACCGAGT	56	168-278	0.53	MT601845
		R:	CTCTTATCTCTCACGTCTCTCCG				
CP004	$(AT)_8$	F:	GTCACAATAACTCAAGTGTGCGA	56	154-210	0.80	MT601846
		R:	TGACAGAACAAGTAGCCTTTGGT				
CP005	$(AG)_{10}$	F:	AAAAGTGAACAAATCAAGCTCGT	56	160-220	0.75	MT601847
		R:	ATTCTTGTATCTGCGGGAAAAAT				
CP006	$(CT)_{12}$	F:	TAACAATAAGACGATGAGGGCAT	58.3	160-230	0.81	MT679697
		R:	AAGCTCGTGCTAGGTTGAAAGTT				
CP007	$(CT)_{12}$	F:	AGAGTGCCTCTGATAAGAAAGCC	58.5	158-188	0.71	MT601848
		R:	AATTCAGAGATAAAGCCCCATTC				
CP008	$(TTC)_9$	F:	ACCGTGACAATGAAGATGAAGAT	59.3	168-208	0.85	MT679698
		R:	GCTAAATTGAAAAGCAGAGCAGA				
CP009	$(TC)_{14}$	F:	GCTATCATGCTACTAACCCAACG	61	129-192	0.82	MT679699
		R:	TCGTACGATCAAACAGAATGTCA				
CP010	$(TTC)_8$	F:	GGAATTCTGACTGGCATCGT	56.7	161-176	0.64	MT679700
		R:	GGTGGTGATCAGGTAGATGAAGA				
CP011	$(CT)_{12}$	F:	TATGGACGTGTTTTGTCTAGGGT	56	132-184	0.88	MT679701
		R:	TGATGATTGATGCAAGTTCGTTA				
CP012	$(CT)_{14}$	F:	CTCCAAAGCACCACTCCAAT	54.8	150-290	0.93	MT679702
		R:	ATCAAAGACCATACCGAAACTGA				
CP013	$(CT)_8$	F:	GCATAACATGGAGGCATTAAAGT	59	158-180	0.83	MT679703
		R:	AGTTCACTTGATGCTTTTTCCTC				

Table 3. Genetic diversity parameters in three populations of *Cyclocarya paliurus*.

Locus	POP-JX(n=11)						POP-HN(n=11)						POP-AH(n=11)					
	A	H_O	H_e	F_{is}	I	Null	A	H_O	H_e	F_{is}	I	Null	A	H_O	H_e	F_{is}	I	Null
CP001	6	0.900	0.795	-0.132	1.670	-0.059	7	0.636	0.773	0.176	1.669	0.077	5	0.727	0.632	-0.150	1.244	-0.058
CP002	9	0.700	0.870	0.195	2.112	0.124	9	0.818	0.810	-0.010	1.893	0.028	4	0.727	0.517	-0.408	0.923	-0.062
CP003	3	0.364	0.492	0.261	0.792	0.197	4	0.667	0.562	-0.187	1.040	-0.061	2	0.364	0.397	0.083	0.586	0.075
CP004	6	0.333	0.660	0.495	1.345	0.091	4	0.727	0.628	-0.158	1.123	-0.005	7	0.636	0.769	0.172	1.654	-0.139
CP005	4	0.000	0.667*	1.000	1.215	0.400	4	0.091	0.607*	0.850	1.073	0.321	4	0.800	0.695	-0.151	1.240	-0.062
CP006	5	0.545	0.764	0.286	1.518	-0.093	7	0.727	0.777	0.064	1.693	-0.057	7	0.900	0.790	-0.139	1.709	0.006
CP007	4	0.900	0.690	-0.304	1.275	0.086	4	0.727	0.645	-0.128	1.162	-0.067	7	0.545	0.661	0.175	1.460	0.024
CP008	8	1.000	0.831*	-0.204	1.895	0.014	8	0.909	0.806	-0.128	1.830	-0.072	7	0.800	0.810	0.012	1.782	-0.076
CP009	6	0.818	0.752*	-0.088	1.538	-0.038	6	0.364	0.798	0.544	1.669	0.241	7	0.909	0.731	-0.243	1.603	-0.103
CP010	4	0.700	0.575	-0.217	1.063	0.094	4	0.273	0.657	0.585	1.211	-0.095	4	0.545	0.607	0.102	1.097	-0.093
CP011	6	0.600	0.765	0.216	1.583	-0.068	7	1.000	0.827	-0.209	1.831	-0.083	7	1.000	0.831	-0.204	1.855	-0.041
CP012	10	1.000	0.872	-0.147	2.161	-0.124	8	1.000	0.847	-0.180	1.961	-0.050	8	0.900	0.825	-0.091	1.895	0.070
CP013	8	0.727	0.752	0.033	1.677	-0.079	7	0.909	0.781	-0.164	1.677	0.232	9	1.000	0.860	-0.163	2.071	0.039
Mean	6.08	0.661	0.73	0.107	1.526	0.042	6.08	0.681	0.732	0.081	1.526	0.032	6	0.758	0.702	-0.077	1.471	-0.032
SD	0.21	0.081	0.03	0.1	0.11	0.15	0.51	0.08	0.03	0.1	0.1	0.14	0.55	0.05	0.04	0.05	0.12	0.07

Conclusion

In the present study, we obtained 174.75 million raw reads using the BGISEQ-500 sequencing platform, and a total of 108,003 unigenes were generated by *de novo* assembly. 77,441 (Nr), 31,223(GO), 60,160 (KOG), and 60,580 (KEGG) unigenes were successfully annotated to NCBI databases. According to Nr annotations, *Juglans regia* (75.8%) showed high gene similarity to *C. paliurus*, which may imply a closer relationship between *J. regia* and *C. paliurus*. According to KEGG pathway annotation, 1,795 unigenes were enriched in metabolic pathways, which will be useful for metabolic engineering. In addition, 27,960 SSRs were detected among 21,517 unigenes, and 60 pairs of primers were experimentally validated. Finally, 13 primer pairs showed polymorphism and stability and can be selected for different population genetic analyses of *C. paliurus*.

Acknowledgment

We thank ManPing Ding for her assistance in the field sampling of *C. paliurus* at Xionglu Nursery Garden. We thank Dr. Ying Wang for her helpful guidance in data analysis and figure diagramming. In addition, we thank Dr. Maroof Ali for his language polishing in our manuscript. This work was supported National Science Foundation of China (No. 41401062) and the Anhui Provincial Natural Science Foundation of China (No.1508085MD66). Xiaohong Li was the moderator of the above funds.

References

Ali, M., Y.J. Liu, Q.P. Xia, S. Bahadur, A. Hussain, J.W. Shao and M. Shuaib. 2021. Pollen micromorphology of eastern Chinese *Polygonatum* and its role in taxonomy by using scanning electron microscopy. *Microsc. Res. Tech.*, 81: 469-473.

Ali, M., S. Bahadur, A. Hussain, S. Saeed, I. Khuram, M. Ullah and N. Akhtar. 2020. Foliar epidermal micromorphology and its taxonomic significance in *Polygonatum* (Asparagaceae) using scanning electron microscopy. *Microsc. Res. Tech.*, 83(11): 1381-1390.

Bae, D.Y., S.M. Eum, S.W. Lee, J.H. Paik and J.K. Na. 2018. Enrichment of genomic resources and identification of simple sequence repeats from medicinally important *Clausena* excavata. *3 Biotech.*, 8(3): 133.

Blanca, J., J. Cañizares C. Roig, P. Ziarsolo, F. Nuez and B. Picó. 2011. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). *B.M.C.*, *Genome*, 12(1): 104.

Du, F.K., F. Xu., H. Qu, S. Feng, J. Tang and R. Wu. 2013. Exploiting the transcriptome of Euphrates poplar, *Populus euphratica* (Salicaceae) to develop and characterize new EST-SSR markers and construct an EST-SSR database. *PLoS One*, 8(4): e61337.

Excoffier, L. and H.E. Lischer. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.*, 10(3): 564-567.

Fang, S., J. Wang, Z. Wei and Z. Zhu. 2006. Methods to break seed dormancy in *Cyclocarya paliurus* (Batal) Iljinskaja. *Sci. Hort.*, 110(3): 305-309.

Fan, D.M., L.J. Ye, Y. Luo, W. Hu, S. Tian and Z.Y. Zhang. 2013. Development of microsatellite loci for *Cyclocarya paliurus* (Juglandaceae), a monotypic species in subtropical China. *Appl. Plant Sci.*, 1(6): 1200524.

Gupta, P.K., S. Rustgi, S. Sharma, R. Singh, N. Kumar and H. Balyan. 2003. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol. Genet. Genom.*, 270(4): 315-323.

Han, X.J., Wang Y.D., Y.C. Chen, L.Y. Lin and Q.K. Wu. 2013. Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in *Litsea cubeba*. *PLoS One*, 8(10): e76890.

Jan, S.A., N. Habib, Z.K. Shinwari, M. Ali and N. Ali. 2021. The anti-diabetic activities of natural sweetener plant Stevia: an updated review. *SN Appl. Sci*. 3, 517: https://doi.org/10.1007/s42452-021-04519-2

Jiang, Z.Y., X.M. Zhang, J. Zhou, S.X. Qiu and J.J. Chen. 2006. Two new triterpenoid glycosides from *Cyclocarya paliurus*. *J. Asian Nat. Prod. Res.*, 8(1-2): 93-98.

Jordon-Thaden, I.E., A.S. Chanderbali, M.A. Gitzendanner and D.E. Soltis. 2015. Modified CTAB and TRIzol protocols improve RNA extraction from chemic *SN Appl. Sci.*, 3(4): 517. ally complex embryophyta. *Appl. Plant. Sci.*, 3(5): 1400105.

Khan, I., Z.K. Shinwari, N.B. Zahra, S.A. Jan, S. Shinwari and S. Najeebullah. 2019. DNA barcoding and molecular systematics of selected species of family Acanthaceae. *Pak. J. Bot.,* 52(1): 205-212.

Kou, Y., S. Cheng, S. Tian, B. Li, D. Fan, Y. Chen, D.E. Soltis, P.S. Soltis and Z. Zhang. 2016. The antiquity of *Cyclocarya paliurus* (Juglandaceae) provides new insights into the evolution of relict plants in subtropical China since the late Early Miocene. *J. Biogeogr.*, 43(2): 351-360.

Kumpatla, S.P. and S. Mukhopadhyay. 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome Res.*, 48(6): 985-998.

Liu, Y., Y. Cao, S. Fang, T. Wang, Z. Yin, X. Shang, W. Yang and X. Fu. 2018. Antidiabetic effect of *Cyclocarya paliurus* leaves depends on the contents of antihyperglycemic flavonoids and antihyperlipidemic triterpenoids. *Molecules*, 23(5): 1042.

Mao, X., J. Liu, X. Li, J. Qin and X. Fu. 2016. Flowering biological characteristics and mating system in immature plantations of heterodichogamous *Cyclocarya paliurus. J. Nanjing U. Techno. Nat. Sci. Ed.*, 40: 47-55.

Metzker, M.L. 2010. Sequencing technologies the next generation. *Nat. Rev. Genet*, 11(1): 31-46.

Najeebullah, S., Z.K. Shinwari, S.A. Jan, I. Khan and M. Ali. 2020. Ethno-medicinal and phytochemical properties of genus allium: a review of recent advances. *Pak. J. Bot.*, 53(1): 135-144.

Ovais, M., A.T. Khalil, S.A. Jan, M. Ayaz, I. Ullah, W. Shinwari and Z.K. Shinwari. 2019. Traditional Chinese medicine going global: opportunities for belt and road countries. *Proc. Pak. Acad. Sci.*, 56(3 SI): 17-26.

Peakall, R. and P.E. Smouse. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes*, 6(1): 288-295.

Porebski, S., L.G. Bailey and B.R. Baum. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.*, 15(1): 8-15.

Rice, W.R. 1989. Analyzing tables of statistical tests. *Evolution*, 43(1): 223-225.

Schuelke, M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.*, 18(2): 233-234.

Seca, A.M.L. and D. Pinto. 2019. Biological potential and medical use of secondary metabolites. *Medicines (Basel)*, 6(2): 66.

Shinwari, Z.K., S.A. Jan, A.T. Khalil, A. Khan, M. Ali, M. Qaiser and N.B. Zahra. 2018. Identification and phylogenetic analysis of selected medicinal plant species from Pakistan: DNA barcoding approach. *Pak. J. Bot.*, 50(2): 553-560.

Tian, X., J. Zheng, Z. Jiao, J. Zhou, K. He, B. Wang and X. He. 2019. Transcriptome sequencing and EST-SSR marker development in *Salix babylonica* and *S. suchowensis*. *Tree Genet. Genomes*, 15(1): 9.

Varshney, R.K., T. Thiel, N. Stein, P. Langridge and A. Graner. 2002. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell. Mol Bio. Lett*., 7(2A): 537-546.

Varshney, R.K., A. Graner and M.E. Sorrells. 2005. Genic microsatellite markers in plants: Features and applications. *Trends Biotechnol.*, 23(1): 48-55.

Xie, J.H., Z.J. Wang, M.Y. Shen, S.P. Nie, B. Gong, H.S. Li, Q. Zhao, W.J. Li and M.Y. Xie. 2016. Sulfated modification, characterization and antioxidant activities of polysaccharide from *Cyclocarya paliurus*. *Food Hydrocoll.*, 53: 7-15. 202-209.

Xing, W., J. Liao, M. Cai, Q. Xia, Y. Liu, W. Zeng and X. Jin. 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(3): 53.

Xu, X., Z. Yin, J. Chen, X. Wang, D. Peng and M. Shangguan Jia. 2016. De novo transcriptome assembly and annotation of the leaves and callus of *Cyclocarya paliurus* (Bata1) Iljinskaja. *PLoS One*, 11(8): e0160279.

Yang, Z.W., K.H. Ouyang, J. Zhao, H. Chen, L. Xiong and W.J. Wang. 2016. Structural characterization and hypolipidemic effect of *Cyclocarya paliurus* polysaccharide in rat. *Int. J. Biol. Macromol.*, 91: 1073-1080.

Yeh, F.C., R. Yang, T. Boyle, Z.H. Ye, J.X. Mao, C. Yeh, B. Timothy and X. Mao. 1999. POPGENE, reversion 1.32: the user-friendly shareware for population genetic analysis.

Zalapa, J.E., H. Cuevas, H. Zhu, S. Steffan, D. Senalik, E. Zeldin, B. McCown, R. Harbut and P. Simon. 2012. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Amer. J. Bot.*, 99(2): 193-208.

Zhu, B., X. Luo, Z. Gao, X. Hu and Q. Wang. 2022. De novo transcriptome assembly and development of EST-SSR markers of the endangered *Dendrebium nobile* (Orchidaceae). *Pak. J. B.*, 54(2): 483-489.

Zhu, F.Y., M.X. Chen, N.H. Ye, W.M. Qiao, B. Gao, W.K. Law, T. Yuan, Z. Dong and T.Y. Liu. 2018. Comparative performance of the BGISEQ-500 and Illumina HiSeq4000 sequencing platforms for transcriptome analysis in plants. *Plant Methods*, 14(1): 69.

(Received for publication 10 March 2022)