

RESEARCH ARTICLE

De Novo Transcriptome Assembly and Annotation of the Leaves and Callus of *Cyclocarya Paliurus* (Bata1) Iljinskaja

Xiaoxiang Xu^{1,2}, Zhongping Yin^{1,2*}, Jiguang Chen^{1,2}, Xiaoqiang Wang³, Dayong Peng¹, Xinchun Shangguan¹

1 Jiangxi Key Laboratory of Natural Products and Functional Food; Jiangxi Agricultural University, Nanchang 330045, China, **2** College of Food Science and engineering, Jiangxi Agricultural University, Nanchang 330045, China, **3** State Key Laboratory of Medicinal Chemical Biology and College of Pharmacy, Nankai University, Tianjin 300000, China

* yin_zhongping@163.com



OPEN ACCESS

Citation: Xu X, Yin Z, Chen J, Wang X, Peng D, Shangguan X (2016) De Novo Transcriptome Assembly and Annotation of the Leaves and Callus of *Cyclocarya Paliurus* (Bata1) Iljinskaja. PLoS ONE 11 (8): e0160279. doi:10.1371/journal.pone.0160279

Editor: Xiang Jia Min, Youngstown State University, UNITED STATES

Received: March 25, 2016

Accepted: July 15, 2016

Published: August 2, 2016

Copyright: © 2016 Xu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. The raw sequence data was deposited into NCBI SRA database (<http://www.ncbi.nlm.nih.gov/sra>) with the project accession number PRJNA326184 with SRA Study accession number SRR3709217 for the leaves and SRR3709206 for the callus. The assembled transcripts were deposited into NCBI Transcriptome Shotgun Assembly Sequence Database (<http://www.ncbi.nlm.nih.gov/genbank/tsa/>), with the accession numbers (GEUI00000000).

Abstract

Cyclocarya Paliurus (Bata1) Iljinskaja contains various bioactive secondary metabolites especially in leaves, such as triterpenes, flavonoids, polysaccharides and alkaloids, and its leaves are widely used as an hyperglycemic tea in China. In the present paper, we sequenced the transcriptome of the leaves and callus of *Cyclocarya Paliurus* using Illumina HiSeq 4000 platform. After sequencing and de novo assembly, a total of 65,654 unigenes were generated with an N50 length of 1,244bp. Among them, 35,041 (53.37%) unigenes were annotated in NCBI Non-Redundant database, 19,453 (29.63%) unigenes were classified into Gene Ontology (GO) database, and 7,259 (11.06%) unigenes were assigned to Clusters of Orthologous Group (COG) categories. Furthermore, 11,697 (17.81%) unigenes were mapped onto 335 pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG), among which 1,312 unigenes were identified to be involved in biosynthesis of secondary metabolites. In addition, a total of 11,247 putative simple sequence repeats (SSRs) were detected. This transcriptome dataset provides a comprehensive sequence resource for gene expression profiling, genetic diversity, evolution and further molecular genetics research on *Cyclocarya Paliurus*.

Introduction

Cyclocarya Paliurus (Bata1) Iljinskaja, a unique genus of Juglandaceae, is a well-known edible and medicinal plant growing in southern China. The leaves of *Cyclocarya Paliurus* have been often used to produce teas with health benefits, which is named “sweat tea” for its slight sweetness [1–4]. It has been demonstrated that *C. Paliurus* exhibits various pharmacological activities such as anti-hypertensive, hypoglycemic, antioxidant and enhancement of mental efficiency [5–7], which may be mostly attributed to its various bioactive components, e.g. flavonoids [4], triterpenoids [8], polysaccharides [9] and polyphenols [10]. Among these compounds, triterpenoids are an important group of health-promoting chemicals. In our previous

Funding: This work was funded by the Natural Science Foundation of PR China (No. 31260368; No. 31460436). ZP Yin received the funding.

Competing Interests: The authors have declared that no competing interests exist.

studies, eight triterpenoids named β -amyryn, ursolic acid, oleanolic acid, betulinic acid, corosolic acid, maslinic acid, β -boswellic acid, and arjunolic acid were isolated and identified from the *C. Paliurus* leaves [11], which have been proven to have valuable pharmacological and biological activities [12–14], and are now widely used in drugs, food and cosmetics.

However, because of the great difficulty in propagation and cultivation, *C. Paliurus* is endangered, and there are very few populations scattered in remote forested mountains, which seriously impaired the utilization of this resource [14]. We investigated the cutting of *C. Paliurus* [15], and achieved a rooting rate of 64.57% by the stimulation of SJCL (a rooting agent) [16]. We studied another rapid propagation method named *in vitro* culture of stem segment and obtained a high bud ratio, but found it was hard to root and grow into a tree [17]. Although significant efforts have been made, applicable large-scale breeding and cultivation techniques of *C. Paliurus* have not been established yet. In order to alleviate the resource shortage, we tried to produce the *C. Paliurus* secondary metabolites by the plant cell culture. Our group succeeded in inducing callus and establishing the cell culture technology, and found cell culture was a promising way to yield triterpenic acids [18, 19]. Five triterpenic acids have been isolated and identified in the cell cultures of *C. Paliurus* in our recent studies [20].

However, there are few studies on the synthetic mechanism of *C. Paliurus* secondary metabolites, and very limited information is available about the metabolism and related biosynthetic genes. So far, only one *C. Paliurus* gene named CpFPS was reported [21]. It's very hard to do more further researches at the molecular level on metabolic mechanism and regulation to achieve high metabolite production. RNA-seq (High-throughput RNA sequencing technology) provides us a feasible way to carry out some secondary metabolism investigations without genomic sequence, and therefore is particularly attractive for non-model organisms like *C. Paliurus*.

RNA-seq is a powerful tool for transcriptome analysis based on second-generation sequencing technology, which have already made substantial contributions to our understanding of genome expression and regulation [22]. This technology can be used to estimate the expression of genes or isoforms, detect differentially expressed genes, and determine novel splice junctions [22–24]. In recent years, RNA-seq has been widely applied in the genome-wide quantification of absolute transcript levels, and the mining of molecular markers and identification of genes involved in biosynthesis of various secondary metabolites in plants, such as *Salvia miltiorrhiza* [25], *Panax notoginseng* [26], *Asparagus racemosus* [27], *Cunninghamia lanceolata* [28], *Gentiana rigescens* [29], and *Astragalus membranaceus* [30].

In the present study, high-quality transcriptome data of the leaves and callus from *C. Paliurus* were obtained using Illumina Hiseq 4000 platform, and a total of 65,654 assembled unigenes were generated and annotated against public protein databases followed by GO, COG and KEGG classification. Moreover, 11,247 putative simple sequence repeats (SSRs) were detected. These transcriptome data provide a valuable public genomic resource for understanding the metabolic mechanisms and facilitating the discovery of genes involved in secondary metabolism pathway and its regulatory, as well as the future gene expression profiling, functional genomic studies of *C. Paliurus*.

Materials and Methods

Plant Material and RNA Extraction

The *C. Paliurus* leaves for RNA-seq were harvested from the arboretum of Jiangxi Agricultural University in July. Calluses were induced from the leaves collected between April and May, then inoculated on fresh agar-based MS medium (Murashige and Skoog medium) [31] supplemented with 2,4-dichlorophenoxy (2,4-D 0.5 mg/L), 1-Naphthaleneacetic acid (NAA 0.3 mg/L) and

6-Furfurylamino-purine (KT 1.0 mg/L) and cultured under a 12/12 h (light/dark) photoperiod. The subculture interval was initially 20 days, then gradually decreased to 10 days with the increase of subculture time [32]. The collected samples were immediately frozen in liquid nitrogen and stored at -80°C . Total RNA of each sample were extracted and purified using TRIzol[®] reagent (Plant RNA purification reagent, Invitrogen, Carlsbad, CA, USA) according the manufacturer's instructions. The RNA concentration and purity were detected by Nanodrop 2000 (Thermo Fisher, America), and the quality of RNA was further verified by gel electrophoresis and Agilent 2100. Only high-quality RNA samples ($\text{OD } 260/280 = 1.8\text{--}2.2$, $\text{OD } 260/230 \geq 2.0$, $\text{RIN} \geq 6.5$, $28\text{S}:18\text{S} \geq 1.0$, $> 10\mu\text{g}$) were used to construct sequencing library.

cDNA library construction and Illumina sequencing

RNA-seq transcriptome library was prepared from $5\mu\text{g}$ of total RNA using TruSeq[™] RNA sample preparation kit from Illumina (San Diego, CA). The poly(A) mRNA was isolated from total RNA using Oligo (dT) magnetic beads. Following purification, the mRNA was randomly cleaved into short fragments (100 to 400 bp) after adding fragmentation buffer. These short fragments were used as templates to synthesize the first-strand cDNA using reverse transcriptase and random primers. The second-strand was synthesized subsequently using a SuperScript double-stranded cDNA synthesis kit (Invitrogen, CA) with random hexamer primers (Illumina). The cDNA fragments were purified and resolved with EB buffer for end repair and A-tailing addition, and then connected with paired-end adapters. After PCR amplification using Phusion DNA polymerase (NEB) for 15 PCR cycles, the cDNA target fragments of 200–300 bp were size-selected to establish the cDNA library on 2% Low Range Ultra Agarose. After quantification by TBS380, the paired-end RNA-seq sequencing library was sequenced from the 5' to 3' ends using an Illumina HiSeq 4000 platform with the 2×151 bp paired-end read module.

Data filtering and De novo assembly

The raw paired-end reads, which were transformed by the Base Calling into sequence data, were cleaned to high-quality reads by removing the joint sequences and adaptor sequences, reads containing more than 10% N rate, and trimming the low-quality reads (quality value < 20) from the 3' end of the sequence and the raw reads with an average length less than 30bp. Then, de novo transcriptome assembly was conducted using software Trinity (<http://trinityrnaseq.sourceforge.net/>, Version number: trinityrnaseq, release-20140413) without reference genome [33], Trinity consists of three software modules: Inchworm, Chrysalis and Butterfly, and has been regarded as the authoritative software for the efficient and robust de novo reconstruction of transcriptome.

De novo assembly was carried out according to the established method [33], which was briefly described as follows. Firstly, linear transcript contigs were efficiently reconstructed by Inchworm in the following seven steps: (1) Constructing a k-mer dictionary ($k = 25$) from all sequence reads; (2) Removing likely error-containing k-mers from the dictionary; (3) Selecting the most frequent k-mer to seed a contig assembly (excluding both low-complexity and singleton k-mers); (4) Extending the seed in each direction with the highest occurring k-mer of a k-1 overlap; (5) Extending the sequence in either direction until it cannot be extended further; (6) reporting the linear contig; (7) Repeating steps three to six with the next most abundant k-mer until the entire k-mer dictionary has been exhausted. Secondly, Inchworm contigs were recursively grouped into connected components by Chrysalis. Contigs with a perfect overlap of k-1 bases or a minimal number of reads that span the junction across both contigs were deemed to be derived from the same gene and clustered into the same group. After grouping, complete de Bruijn graphs were constructed for each component. Thirdly, Butterfly processed the

individual graphs independently, and extracted full-length isoforms and teased apart transcripts derived from paralogous genes. Redundant sequences were eliminated, and the longest transcript that could not be extended on either end was defined as unigenes. The assembled unigenes (longer than 200 bp) had been deposited into the NCBI Transcriptome Shotgun Assembly Sequence Database (<http://www.ncbi.nlm.nih.gov/genbank/tsa/>) with the accession numbers (GEUI00000000).

Functional annotation and classification

All assembled unigenes were aligned with BLASTX program for homology searches against publicly available protein databases Non-redundant (<http://www.ncbi.nlm.nih.gov/>), Swissprot (<http://www.ebi.ac.uk/uniprot/>), Pfam (<http://pfam.sanger.ac.uk/>), String(<http://string-db.org/>) with identity set at >30% and a cutoff E-value of 10^{-5} , and annotated and classified on Gene Ontology (<http://www.geneontology.org/>), Clusters of Orthologous Group (<http://www.ncbi.nlm.nih.gov/COG/>), and the KEGG pathway (<http://www.genome.jp/kegg/>) with a threshold E-value of 10^{-5} . The aligning results were used to identify the sequence direction and to predict the coding regions. If the aligning results from different databases were conflicted with each other, a priority order of alignments from Nr, SwissPort, KEGG, GO and COG was followed. Based on the Nr annotations, the Blast2GO program was used to obtain GO annotations according to biological process, molecular function and cellular component [34]. The unigenes were also aligned to the COG database to predict and classify functions, and the secondary metabolic pathways were annotated according to the KEGG pathway database. Transcription Factors of *C. Paliurus* were extracted from Plant Transcription Factor Database (PlantTFDB), and unigenes were mapped to them using Blastn program.

Detection of SSR markers

The unigenes were scanned for microsatellites using the MISA software (<http://pgrc.ipkgatersleben.de/misa/>) with the default parameters. The parameters were adjusted for identification of perfect di-nucleotide, tri-nucleotide, tetra-nucleotide, penta-nucleotide, and hexa-nucleotide motifs with a minimum of 6, 5, 5, 5, and 5 repeats, respectively. Primer pairs were designed using Primer 3.0.

Results and Discussion

RNA sequencing and de novo assembly

To generate a comprehensive overview of *C. Paliurus* transcriptome, total RNA were extracted from leaves and callus, then the mRNA was isolated, and cDNA libraries were established and sequenced separately using Illumina Hiseq 4000 platform, which generated 39.0 and 49.4 million raw reads, respectively (Table 1). After removing adaptor sequences, ambiguous reads and low-quality reads, the quality of reads was assessed successively. The clean reads were individually generated with the average GC percentage of 46.49% and 47.90% (Table 1). By using the Trinity program [1], all high-quality reads were assembled into 65,654 unigenes with an N50 of 1,244 bp and average length of 704 bp, and 84,223 transcripts were constructed with an N50 of 1,362 bp and average length of 792 bp (Table 2). The average GC content of the unigenes was 42.95%. Furthermore, the length of these unigenes ranged from 201 to 10,000 bp (Fig 1), and the majority were disturbed in 201-400bp. However, there are still 14,155 unigenes (21.56%) whose lengths were more than 1,000bp. These data indicated that the generated unigenes in our experiments were of fine quality and therefore suitable for further annotation.

Table 1. Summary of transcriptome sequencing of *C. Paliurus* leaves and callus.

	leaves	callus
Total raw reads	39,053,944	49,408,746
Total clean reads	38,324,176	48,392,438
Q20 percentage (%)	98.11	98.09
Q30 percentage (%)	94.18	94.12
Error percentage (%)	0.0116	0.0117
GC percentage (%)	46.49	47.90

(Note: Q20, the proportion of nucleotides with quality value larger than 20 in reads, Q30, the proportion of nucleotides with quality value larger than 30 in reads)

doi:10.1371/journal.pone.0160279.t001

Functional annotation and classification

In order to predict and analyze the function of assembled unigenes, the total annotated unigenes were aligned against the NCBI non-redundant (Nr) database, the String and SwissPort protein database, the Gene Ontology (GO) database, the Clusters of Orthologous Group (COG) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database using the BLASTX program with an E-value cut-off of 10^{-5} . In total, 35,041 (53.37%) unigenes were annotated in Nr database (S1 Table). Beyond that, 17,709 (26.97%), 15,529 (23.65%), 11,697 (17.81%), 20,629 (31.42%) unigenes were annotated in Pfam, String, KEGG, SwissPort databases, respectively (Table 3). Further blast statistics indicated that 80.60% of the annotated unigenes in NR exhibited high homology with the E-value < $1e-20$, and 52.69% with a strong E-value (E-value = 0) (Fig 2A). The similarity distribution showed that 63.97% of the annotated sequences had similarities higher than 80%, while 36.02% had a similarity between 40% and 80% (Fig 2B). Additionally, the annotated unigenes were compared to known nucleotide sequences of other plant species, which were best matched to the known nucleotide sequences from *Vitis vinifera* (11.40%), followed by *Theobroma cacao* (11.28%), *Prunus persica* (8.97%), *Prunus mume* (8.19%), and *Morus notabilis* (5.69%) (Fig 2C).

Gene Ontology (GO) is an international standardized gene functional classification system. The GO terms for *C. Paliurus* unigenes were retrieved using Blast2GO [35]. A total of 19,453 (29.63%) assembled unigenes were annotated and classified into three main categories: Biological Processes, Cellular Component, and Molecular Function, and then distributed into 58 sub-categories (Fig 3). Within the Biological Processes classification, metabolic process (12,696, 65.32%), cellular process (11,001, 56.59%), and single-organism process (8,854, 45.60%) were the most significantly represented (Fig 3), which indicated that these unigenes played an important metabolic activity in *Cyclocarya Paliurus*. Under the Cellular Component classification, the unigenes were mainly related to “cell” (7,090, 36.45%) and “cell part” (7,090, 36.45%),

Table 2. Summary of the sequence assembly after Illumina sequencing.

	Unigenes	Transcripts
Total sequence num	65,654	84,223
Total sequence base	46,199,033	66,728,179
GC percentage (%)	42.94	42.96
Average length (bp)	704	792
Smallest length (bp)	201	201
Largest length (bp)	9184	9184
N50 (bp)	1,244	1,362

doi:10.1371/journal.pone.0160279.t002

Sequence length distribution

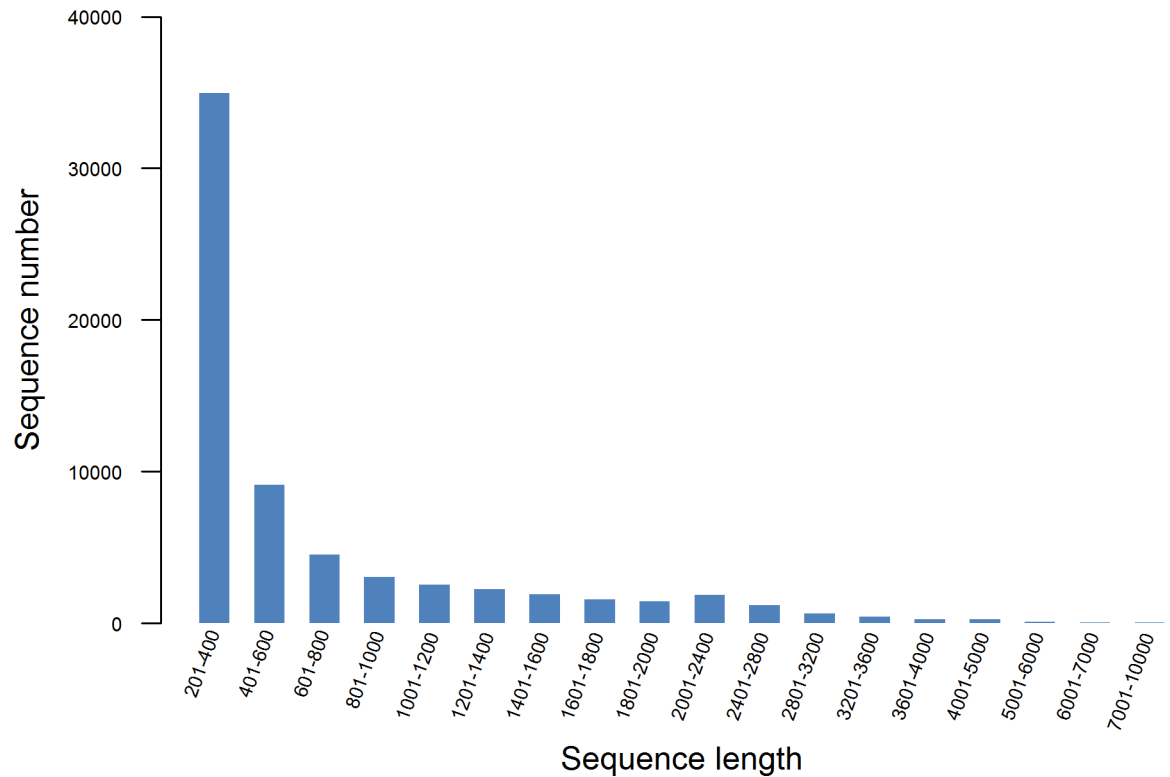


Fig 1. Length distribution of assembled unigenes. Among 65,654 unigenes, 44,010 (67.03%) were shorter than 600bp and 14,155(21.56%) were longer than 1,000bp.

doi:10.1371/journal.pone.0160279.g001

followed by “organelle” (5,031, 25.86%) and “membrane” (4,324, 22.23%). Only a few unigenes were assigned to extracellular matrix, extracellular matrix part and collagen trimer. In the Molecular Function category, the majority were assigned to “binding” (10,265, 52.77%) and “catalytic activity” (10,428, 53.61%) prominently. The above-mentioned findings were similar to the recent report of Chinese Chive transcriptome functional annotation [35].

COG is a database which is widely used to predict and classify functional genes. Every protein in the COG database is assumed to be evolved from an ancestor, and the whole database is built on coding proteins with complete genomes as well as system evolution relationships of bacteria, algae, and eukaryotes [36, 37]. Out of the 65,654 unigenes, 7,259 (11.06%) were annotated and classified into 24 functional categories (Fig 4). Among the aligned COG classifications, “general function prediction only” category (1,006, 13.86%) was the largest group,

Table 3. Summary of unigenes annotation.

Database	Total unigenes	Annotated unigenes	Percentage
Pfam	65,654	17,709	26.97%
String	65,654	15,529	23.65%
KEGG	65,654	11,697	17.81%
Swissprot	65,654	20,629	31.42%
NR	65,654	35,041	53.37%

doi:10.1371/journal.pone.0160279.t003

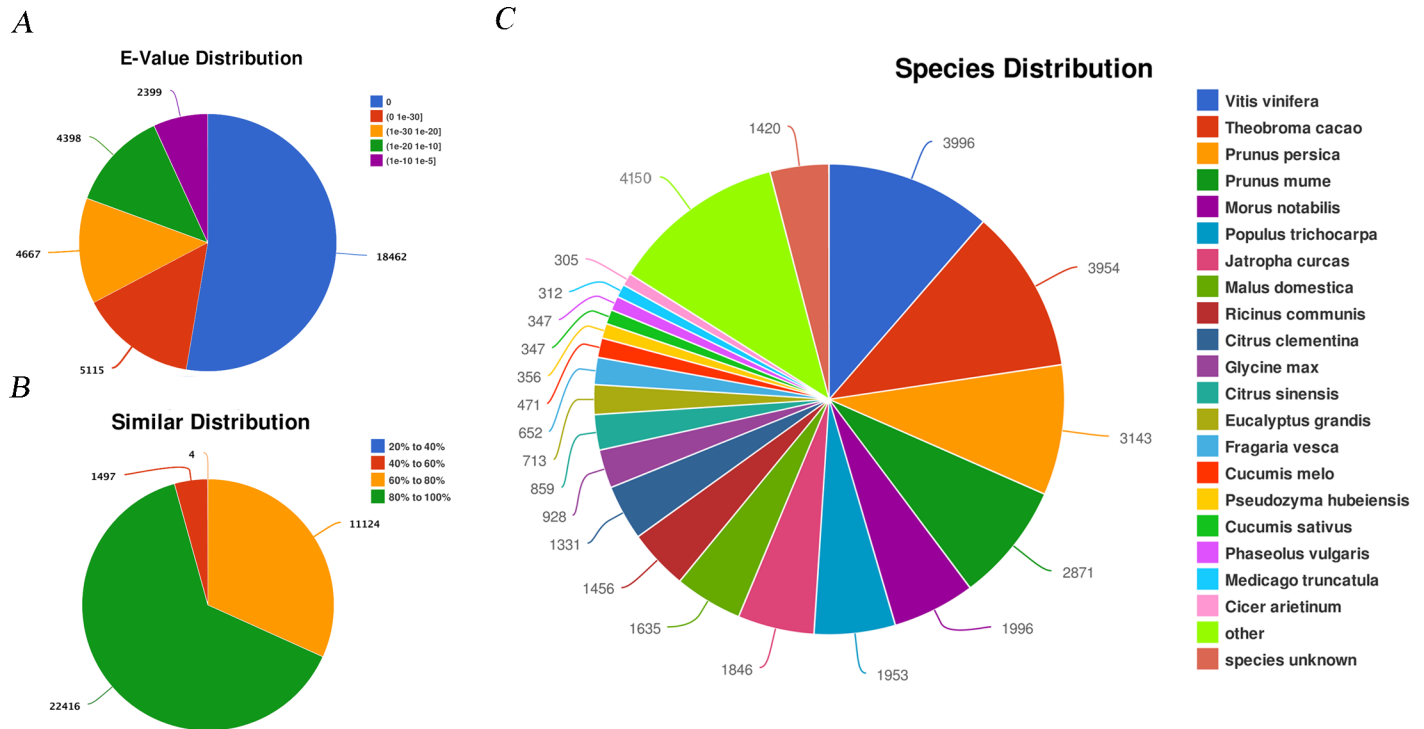


Fig 2. Unigenes homology searches against the NR database. (A) The E-value distribution of BLAST hits. (B) The similar distribution of BLAST hits. (C) Species distribution of the top BLASTX hits.

doi:10.1371/journal.pone.0160279.g002

followed by “signal transduction mechanisms” (972, 13.39%), “posttranslational modification, protein turnover, chaperones” (757, 10.43%), “translation, ribosomal structure and biogenesis” (664, 9.15%), and “carbohydrate transport and metabolism” (510, 7.03%), whereas only few unigenes were assigned to “cell motility” (7, 0.096%), “extracellular structures” (1, 0.014%), and “nuclear structure” (1, 0.014%), respectively. In addition, 334 unigenes were classified into the unknown function.

Metabolic pathway analysis by KEGG

The KEGG pathway database provides a wealth of information on molecular interaction and reaction networks to further understand the biological functions of unigenes. The mapped results indicated that 11,697 (17.81%) unigenes were predominantly annotated with Enzyme Commission (EC) numbers and divided into five branches according to the metabolic pathway (Fig 5) (Metabolism; Genetic Information Processing; Environmental Information Processing; Cellular Processes; Organismal Systems), and further grouped into 335 KEGG pathways (S2 Table). It was noteworthy that 7,746 (66.34%) mapped unigenes participated in the metabolism and 1,312 (11.24%) were involved in the biosynthesis of secondary metabolites such as phenylpropanoid biosynthesis (197), flavonoid biosynthesis (42), terpenoid backbone biosynthesis (58), which were the important information we are especially interested in. Furthermore, the top 20 largest annotated pathway groups of *C. Paliurus* were presented in Fig 6. The most representative KEGG pathway was “Ribosome”, followed by “Plant hormone signal transduction” (328), “Protein processing in endoplasmic reticulum” (301), “RNA transport” (299), and “Plant pathogen interaction” (280). As shown in Fig 6, there were 328 (2.81%) unigenes mapped into the “Plant hormone signal transduction” pathway, some of which might be stimulated to express by the supplemented plant hormones in the callus culture medium. In our

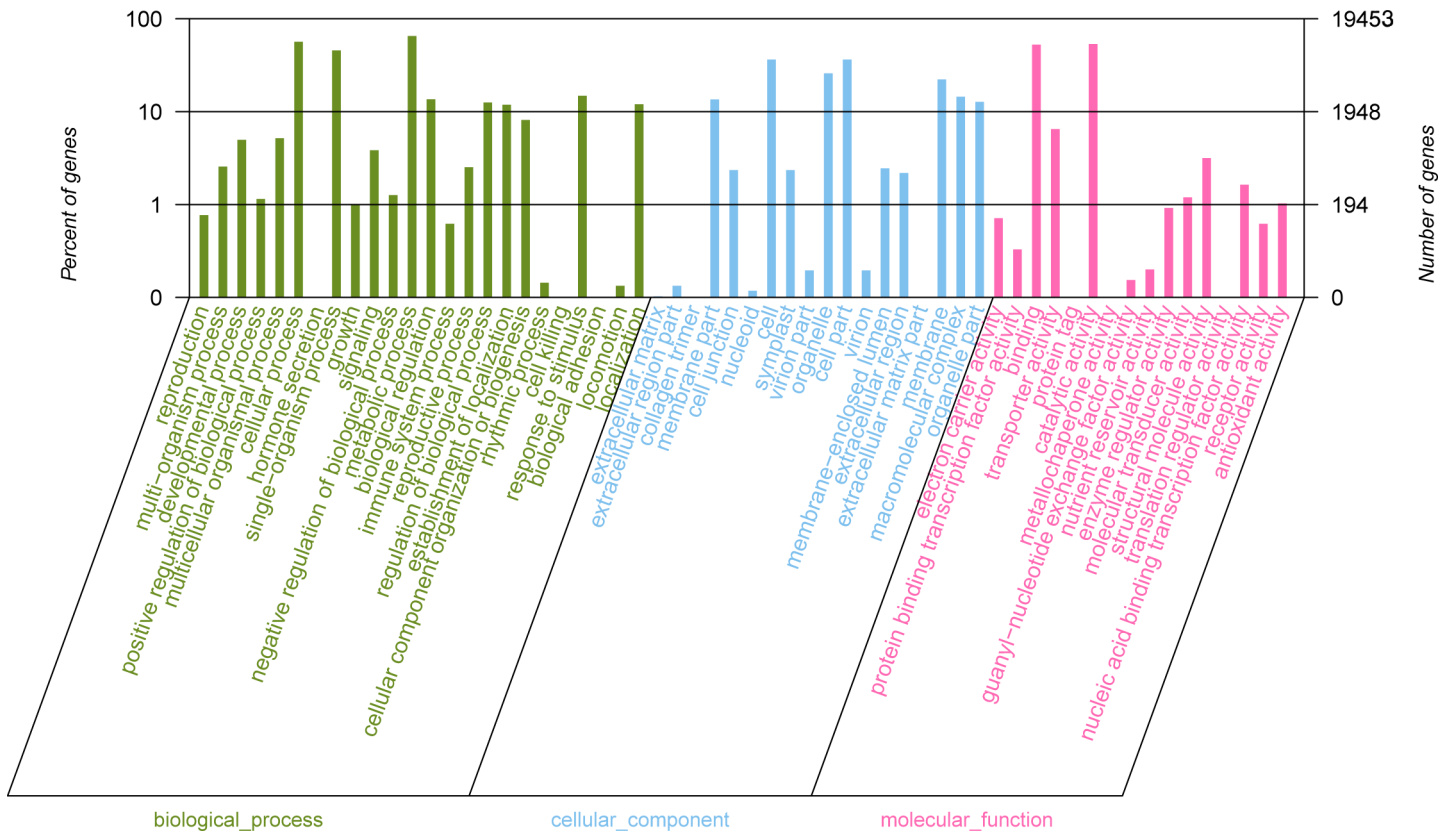


Fig 3. GO classification of assembled sequences. A total of 19,453 unigenes were classified into three main GO categories: Biological Processes, Cellular Component, and Molecular Function.

doi:10.1371/journal.pone.0160279.g003

experiments, plant hormones such as rootone, indole acetic acid, and cytokinins were added into the callus culture medium, which have been presumed to participate in the regulation of “Plant growth”, “Cell culture”, “Cell differentiation”, and “Recession”. These predicted results indicated that numerous unigenes were involved in the secondary metabolite biosynthesis of *C. Paliurus* leaves and callus. The above mapped information with KEGG pathway database would be very useful for future researches on gene function and its regulatory mechanism of *C. Paliurus*.

Transcription factor analysis

Transcription factors (TFs) play critical roles in plant growth, bioactive component synthesis and gene expression regulation, especially in the secondary metabolism regulation. Plants show various TFs expression patterns when growing in different environments or facing stress, which further significantly affect the synthesis of secondary metabolites [38–40]. Therefore, the identification of putative TF genes is useful for understanding the regulatory mechanism of secondary metabolites. TFs are often classified into different families according to the features of DNA-binding domains. In this study, a total of 21,843 unigenes were annotated and further classified into 60 transcription factor families (Plant Transcription Factor Database, PlantTFDB, <http://planttfdb.cbi.pku.edu.cn>) in this paper (Fig 7). Among these TF families, the most abundant transcription factors of *C. Paliurus* were found in the bHLH family which includes 2,120 unigenes. The second was NAC, followed by the bZIP, MYB-related, WRKY, C3H, B3 and C2H2 TF families, which contains 1,734, 1,422, 1,311, 1,071, 1,036, 1,007 and 990

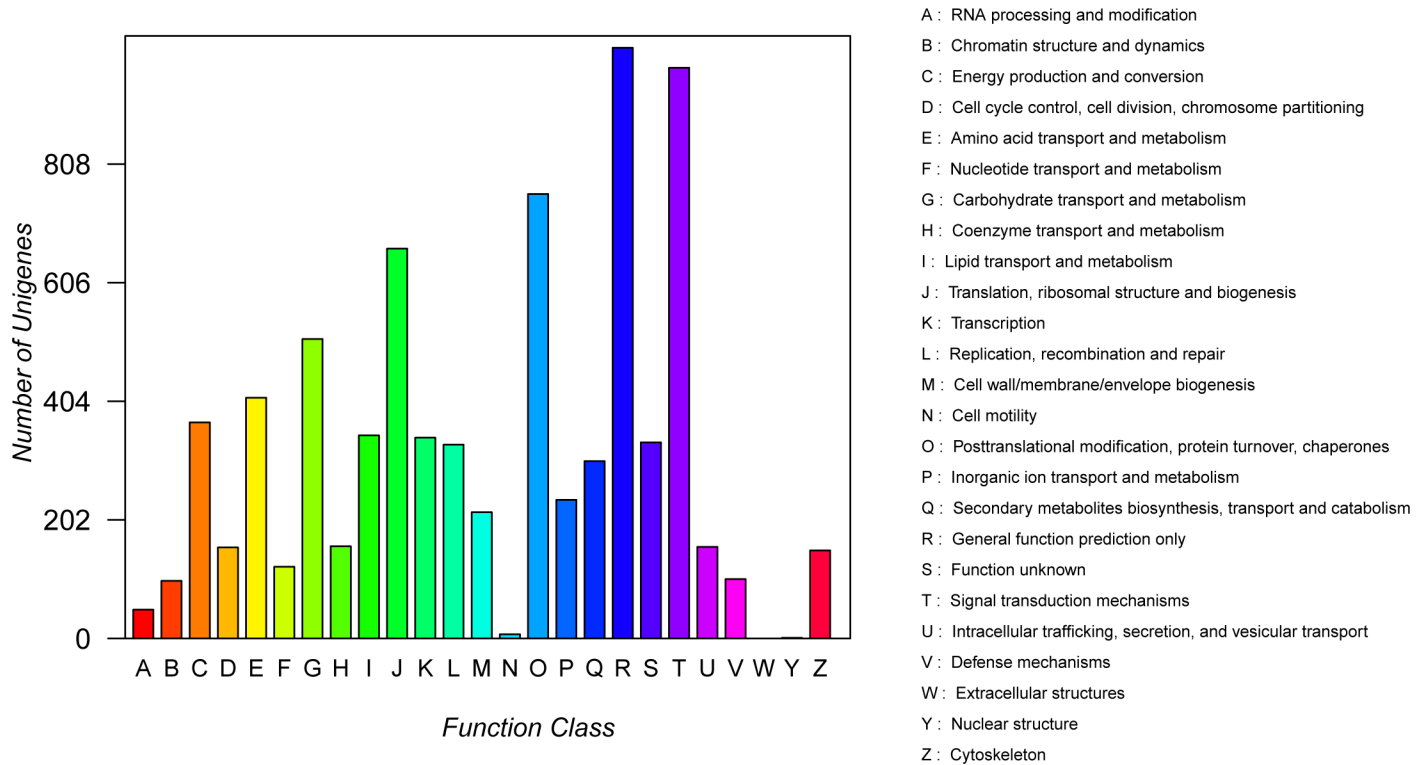


Fig 4. COG functional classification. A total of 7,259 unigenes were assigned to 25 classification categories.

doi:10.1371/journal.pone.0160279.g004

unigenes, respectively. Researches have validated that the bHLH, WRKY, MYB, bZIP, and C2H2 TF families play a major role in the regulation of many genes which participate in the plant secondary metabolism [41, 42], especially for the regulation of the bioactive component synthesis, such as flavonoids [43], alkaloids [44, 45], and terpenoids [46]. The expression level of these TFs in *C. Paliurus* may be associated with the biosynthesis of secondary metabolites, and the discovery of these putative TFs may provide valuable information for the future researches on gene expression regulation, particularly those TFs related to the flavonoid pathway which will be described below.

Analysis of flavonoid biosynthesis pathway

Flavonoids are polyphenolic secondary metabolites derived from the phenylalanine via the phenylpropanoid pathway, which become a research focus in recent years for their various biological and pharmacological activities [47]. It was reported that the total flavonoid content of *C. Paliurus* leaves ranged from 0.73 to 4.73%, depending on the growing region, harvest time, determination method and so on [48–50]. The most abundant flavonoid was isoquercitrin in the *C. Paliurus* leaves [12]. Up to now, sixteen flavonoids have been isolated and identified from *C. Paliurus* (Table 4). In the present study, 197 unigenes were found to be involved in phenylpropanoid pathway by mapping with KEGG pathway database. Among them, 42, 3, 1 and 3 unigenes were involved in the biosynthesis of flavonoids, anthocyanin, isoflavonid and flavone and flavonol, respectively, which represented different enzymes in the different pathways. The unigenes associated with flavonoid biosynthetic pathway were shown in the Fig 8. A total of 10 candidate genes with annotations matching enzymes in the flavonoid biosynthesis, i.e., phenylalanine ammonialyase (PAL), cinnamate-4-hydroxylase (C4H), 4- coumaroyl:

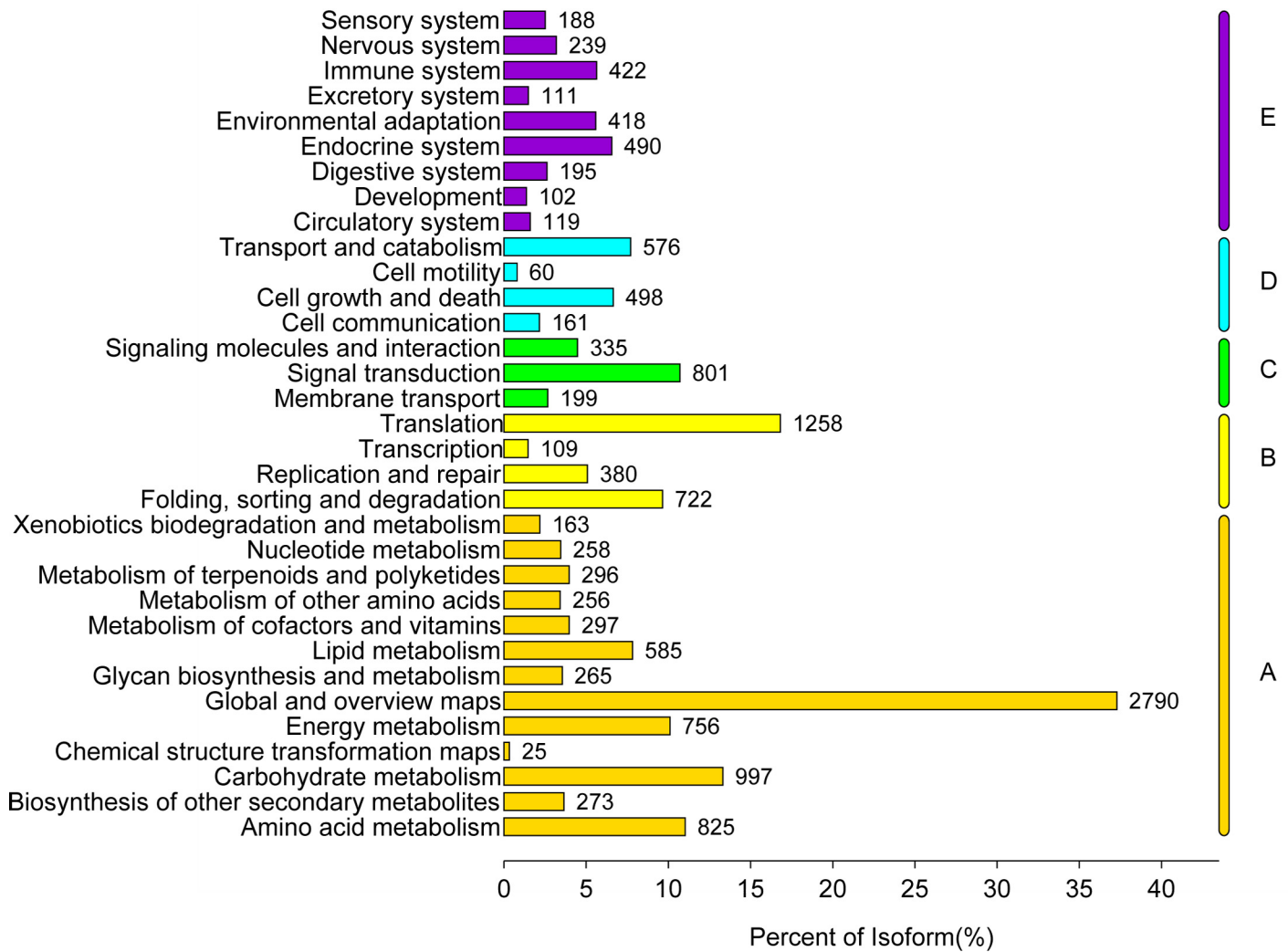


Fig 5. KEGG pathway annotation of the unigenes. These unigenes were divided into five branches (A, Metabolism; B, Genetic Information Processing; C, Environmental Information Processing; D, Cellular Processes; E, Organismal System.)

doi:10.1371/journal.pone.0160279.g005

coenzyme A ligase(4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavanone -3-hydroxylase (F3H), flavanone-3'-hydroxylase (F3'H), flavonoid 3',5'-hydroxylase (F3'5'H), dihydroflavonol 4-reductase (DFR) and flavonol synthase (FLS), were annotated in this paper. Moreover, the unique putative unigenes encoding anthocyanidin synthase (ANS), leucoanthocyanidin reductase (LAR), anthocyanidin reductase (ANR) and anthocyanidin 3-O-glucosyltransferase (3GT) were also prominently found. It's important to clone these unique genes and further analyze their functions in the future studies. Isoflavonoids, a subclass of flavonoids with special structure and function, have been normally found in leguminous plants. It's interesting that one isoflavonoid 7-hydroxyl-4'-methoxy isoflavone was identified in *C. Paliurus* leaves [51]. However, 2-hydroxyisoflavanone synthase (IFS), which catalyzes the conversion of flavanones to isoflavones, hasn't been found in our mapping experiments. A further study is needed in the future.

Transcription factors play an important role in flavonoid biosynthesis. These factors regulate some genes co-expression to stimulate or inhibit the accumulation of flavonoids when combining with the special structure gene function. So far, researchers have isolated and

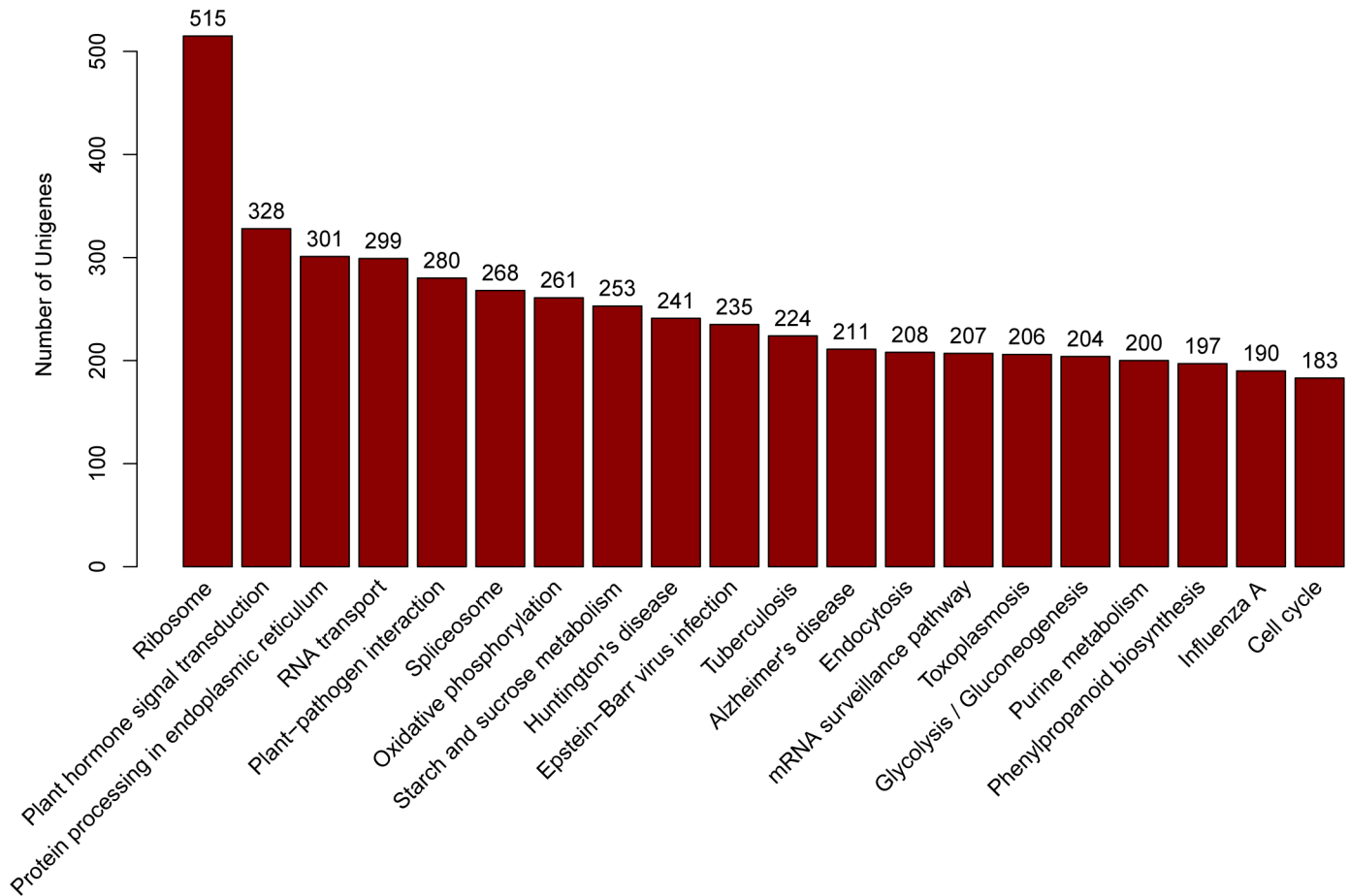


Fig 6. The top 20 largest KEGG pathways. Annotated unigenes were classified into 335 KEGG pathways. The top 20 pathways containing unigenes are displayed.

doi:10.1371/journal.pone.0160279.g006

identified many transcription factor genes involved in regulation of flavonoids in *Maize* [52], *Arabidopsis* [53], *Petunia* [54], *Rice* [55] and other crops, such as MYB, MYC, bZIP, WD40, zinc finger TFs. Among these transcription factors, MYB and MYC (bHLH) families became the research focus.

MYC (bHLH) is one of the biggest transcription factor families in plant. This family regulates not only the plant growth and developmental processes including formation of trichome and light signal transduction, but also the stress responses and secondary metabolism [56, 57]. Plant bHLH proteins have been classified into 32 subfamilies according to genome-wide classification and the evolutionary analysis [58], whereas members of the same plant bHLH subfamilies had similar functional characteristic [59]. There were some bHLH subfamilies involved in the regulation of phenylpropanoid and terpenoids biosynthesis pathways, such as AtTT8 in *Arabidopsis* [60], OsRa-c in rice [61] and TcJAMYC in yew [46]. Furthermore, the above mentioned subfamilies have also been found to be related to the regulation of anthocyanin biosynthesis in various plants, for example, *Zea mays* L, *Oryza sativa* L, *Dahlia variabilis* Hort. and *Brassica oleracea* L [62–65].

It was reported that there were approximately 197 and 155 MYB genes in *Arabidopsis* and rice, respectively [66]. The first plant MYB gene, C1, encodes a MYB-like TF, was isolated from *Zea mays*, which regulated the synthesis of anthocyanin with the synergy of R/B protein family

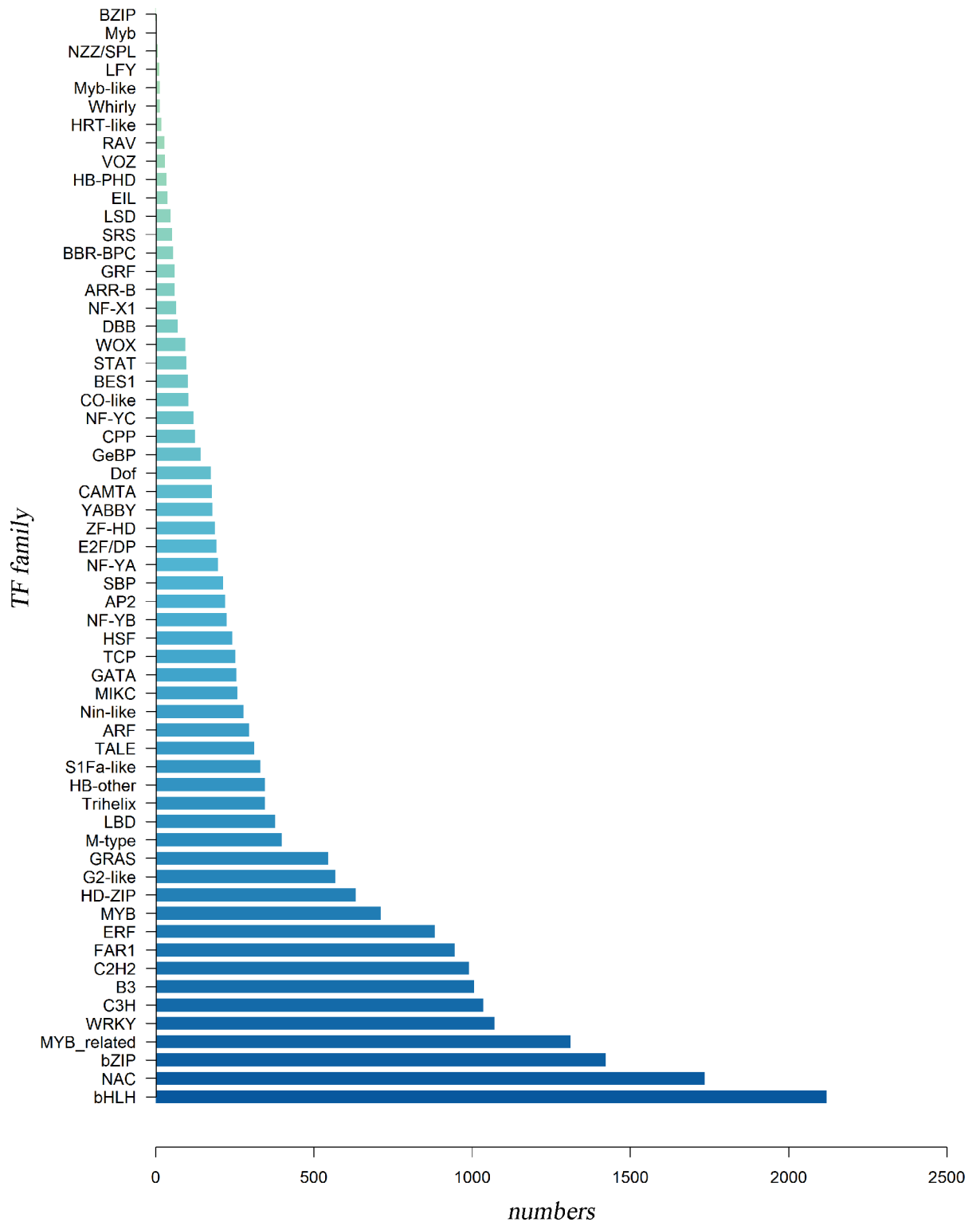


Fig 7. Summary of transcription factor unigenes of *C. Paliurus*. The number of unigenes related to TFs in each TF family. Among the TF families bHLH, NAC, bZIP and MYB_related proteins were the most abundant.

doi:10.1371/journal.pone.0160279.g007

Table 4. Summary of flavonoids identified in *C. Paliurus*.

Number	Name	Chemical Formula	Reference
1	Kaempferol	C ₁₅ H ₁₀ O ₆	[71]
2	Quercetin	C ₁₅ H ₁₀ O ₇	[71]
3	Isoquercitrin	C ₂₁ H ₂₀ O ₁₂	[71]
4	3,6,3',5'-tetramethoxy-5,7,4' trihydroxyflavone	C ₁₉ H ₁₈ O ₉	[72]
5	Kaempferol-7-O-α-L-rhamnoside	C ₂₁ H ₂₀ O ₁₀	[73]
6	Kaempferol-4'-Methylether-7-O-β-D-Mannose	C ₂₂ H ₂₂ O ₁₁	[73]
7	Quercetin-3-O-α-L-rhamnopyranoside	C ₂₁ H ₂₀ O ₁₁	[74]
8	Kaempferol-3-β-D-glucuronide	C ₂₁ H ₁₈ O ₁₂	[75]
9	Kaempferol-3-O-α-L-Rhamnopyranoside	C ₂₁ H ₂₀ O ₁₀	[76]
10	Kaempferol-3-O-β-D-Galactopyranoside	C ₂₁ H ₁₈ O ₁₁	[76]
11	Kaempferol 3-O-β-D-glucopyranoside	C ₂₁ H ₂₀ O ₁₁	[76]
12	Kaempferol-3-O-α-L-(4''-E-p-coumaroyl) rhamnopyranoside	C ₃₃ H ₃₂ O ₁₅	[76]
13	Quercetin-3-O-β-D-glucuronate sodium	C ₂₁ H ₁₇ O ₁₄ Na	[77]
14	myricetin-3-O-β-D-glucuronate sodium	C ₁₅ H ₁₀ O ₁₅ Na	[77]
15	7-hydroxyl-4'-methoxy isoflavone	C ₂₁ H ₁₇ O ₁₅	[51]
16	Myricetin	C ₁₅ H ₁₀ O ₈	[51]

doi:10.1371/journal.pone.0160279.t004

[67]. Grotewold et al characterized the maize P gene encode proteins with homology to the DNA-binding domain of MYB-like transcription factors [68]. According to Grotewold et al [69], maize P1 (R2R3-MYB transcription factor) improved the expression of CHS (chalcone synthase), CHI (chalcone isomerase), DFR(dihydroflavonol 4-reductase), but didn't stimulate the expression of branching enzyme such as F3H (flavanone 3-hydroxylase) in anthocyanin biosynthesis. However, it has been reported that many R2R3-MYB transcription factors, e.g., AN2 in *Petunia* hybrid [70] and PAP1/PAP2 in *Arabidopsis thaliana* [53], were involved in the regulation of anthocyanin biosynthesis. In our analysis, a total of 711 MYB genes were found in *C. Paliurus*. Therefore, more study remains to be done to identify the MYB genes from *C. Paliurus* for studying the regulation of flavonoid biosynthesis.

In the present paper, we found many transcription factors involved in flavonoid metabolism of *C. Paliurus*, providing critical information for further regulation study on the gene expression and secondary metabolism. We have established the cell suspension culture of *C. Paliurus* to produce the bioactive components such as flavonoids, triterpenoids and polysaccharides, and the above mentioned transcription factor information will be helpful to further develop and improve our current approach to achieve a higher yield of these compounds by gene expression regulation. Therefore, we can use genetic engineering for improving plant flavonoid secondary metabolic pathways to effectively increase the content of secondary metabolites by studying the transcription factors involved in flavonoid secondary metabolism, and understanding the mechanisms of plant secondary metabolic regulation.

Detection of simple sequence repeats (SSRs)

SSRs, or microsatellites are important molecular markers for genetics and biology researches, including gene mapping, genetic diversity assessment, comparative genomics, and molecular breeding. In this paper, 65,654 assembled unigene sequences from *C. Paliurus* were scanned to explore the SSR profiles using MISA software and the results were shown in Table 5. A total of 9,688 sequences containing 11,247 SSRs were identified. Of all 9,688 SSR containing sequences, 1,353 had more than one SSR. In addition, 494 SSRs were present in compound forms. Among these SSRs, dinucleotide repeat motifs (5,198, 46.21%) were the most abundant, followed by

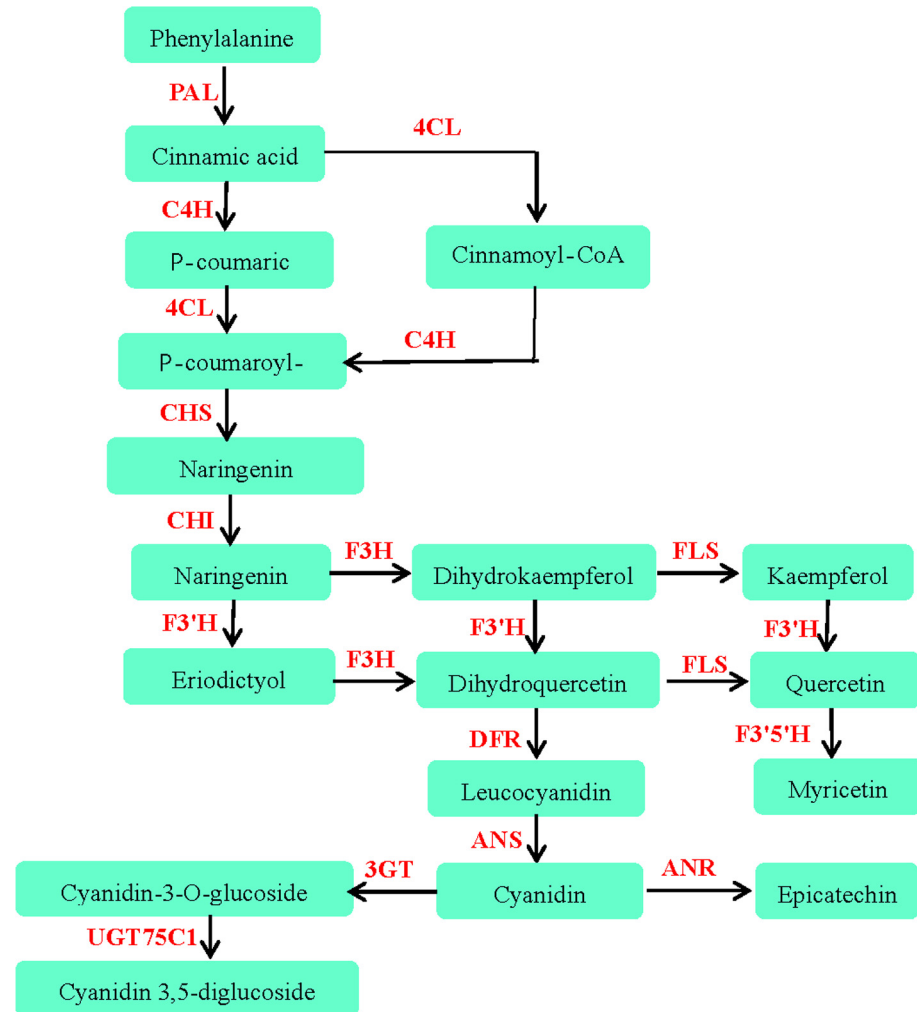


Fig 8. Putative biosynthesis pathway of flavonoids and phenolic compounds in *C. Paliurus*. PAL, phenylalanine ammonialyase; C4H, cinnamate-4-hydroxylase; 4CL, 4- coumaroyl:coenzyme A ligase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone -3- hydroxylase; DFR, dihydroflavonol 4-reductase; FLS, flavonol synthase; ANS, anthocyanidin synthase; LAR, leucoanthocyanidin reductase; ANR, anthocyanidin reductase; 3GT, anthocyanidin 3-O-glucosyltransferase; F3'5'H, flavonoid 3',5'-hydroxylase and UGT75C1,UDP-glucose: anthocyanidin 3-O-glucoside 5-O-glucosyltransferase.

doi:10.1371/journal.pone.0160279.g008

4,217 (37.49%) mononucleotide repeat motifs, 1,681 (14.95%) trinucleotide repeat motifs, tetra-nucleotide, hexa-nucleotide and penta-nucleotide repeat motifs. The main repeat motifs were AG/CT which accounted for 36.90% (4,150 SSRs), followed by A/T (4,004, 35.60%), AT/AT,

Table 5. Summary of the SSRs identified in the transcriptome sequences.

Item	Number
Total number of sequences examined	65,654
Total size of examined sequences (bp)	46,199,033
Total number of identified SSRs	11,247
Number of SSR containing sequences	9,688
Number of sequences containing more than 1 SSR:	1,353
Number of SSRs present in compound formation	494

doi:10.1371/journal.pone.0160279.t005

Table 6. Frequency of SSRs repeat motifs in *C. Paliurus*.

Repeat motif length	Repeat number							Total	Frequency%
	5	6	7	8	9	10	>10		
AC/GT		134	73	45	22	12	21	307	2.72
AG/CT		1,109		737	746	879	679	4,150	36.90
AT/AT		216	132	121	113	110	29	721	6.41
CG/CG		10	8	2				20	0.18
AAC/GTT	44	18	5	1	1			69	0.61
AAG/CTT	283	170	100	5				558	4.96
AAT/ATT	77	25	21	1				124	1.10
ACC/GGT	96	35	23	3	1			158	1.40
ACG/CGT	26	10	7	4				47	0.42
ACT/AGT	17	9	5	1				32	0.28
AGC/CTG	126	40	25	1				192	1.71
AGG/CCT	146	65	46	4				261	2.32
ATC/ATG	107	48	30	4				189	1.68
CCG/CGG	35	6	10					51	0.45
Tetranucleotide		108	16		1			125	1.11
Pentanucleotide		11	1					12	0.11
Hexanucleotide			10	4				14	0.12

doi:10.1371/journal.pone.0160279.t006

AAG/CTT, AC/GT and ACC/GGT repeat (Table 6). A total of 5341 primer pairs, which contain three sets of primers respectively, were designed from 9,688 sequences using Primer 3 (S3 Table). These results provided plenty of reliable markers for genetic linkage mapping, analysis of genetic polymorphism and functional gene mining of *C. Paliurus* and its closely related species.

Conclusions

C. Paliurus is a well-known edible and medicinal plant, but no genomic information is available yet. In this paper, the transcriptome of *C. Paliurus* leaves and callus without a reference genome was analyzed using Illumina Hiseq 4000 platform. A total of 65,654 assembled unigenes were generated. The annotated unigenes were functionally classified in the GO, COG, and KEGG databases. Moreover, the putative simple sequence repeats (SSRs) were detected. To our knowledge, this is the first attempt to de novo assemble the whole transcriptome of *C. Paliurus*. This study provided not only a comprehensive enough coverage for gene cloning, expression, and functional analysis, but also a valuable public platform to understand the biosynthesis and regulation of secondary metabolites in *C. Paliurus*, especially those important bioactive components.

Supporting Information

S1 Table. Unigene annotation by the NCBI NR, Swiss-Port, Pfam, String, COG, GO and KEGG databases.
(XLSX)

S2 Table. List of KEGG pathway in *C. Paliurus*.
(XLSX)

S3 Table. Designed SSR primers for *C. Paliurus*.
(XLSX)

Acknowledgments

The authors are grateful for the financial supports from the Natural Science Foundation of PR China (No. 31260368; No. 31460436). The authors also acknowledge Majorbio for technical assistance in Illumina sequencing.

Author Contributions

Conceived and designed the experiments: ZPY XQW. Performed the experiments: XXX JGC. Analyzed the data: XXX DYP. Contributed reagents/materials/analysis tools: XCS ZPY. Wrote the paper: XXX XQW ZPY.

References

1. Fang S, Wang J, Wei Z, Zhu Z. Methods to break seed dormancy in *Cyclocarya Paliurus* (Batal.) Iljinskaja. *Scientia Horticulturae*. 2006; 110(3):305–9. doi: [10.1016/j.scienta.2006.06.031](https://doi.org/10.1016/j.scienta.2006.06.031)
2. Birari RB, Bhutani KK. Pancreatic lipase inhibitors from natural sources: unexplored potential. *Drug Discov Today*. 2007; 12(19–20):879–89. doi: [10.1016/j.drudis.2007.07.024](https://doi.org/10.1016/j.drudis.2007.07.024) PMID: [17933690](https://pubmed.ncbi.nlm.nih.gov/17933690/)
3. Xie JH, Xie MY, Nie SP, Shen MY, Wang YX, Li C. Isolation, chemical composition and antioxidant activities of a water-soluble polysaccharide from *Cyclocarya Paliurus* (Batal.) Iljinskaja. *Food Chemistry*. 2010; 119(4):1626–32. doi: [10.1016/j.foodchem.2009.09.055](https://doi.org/10.1016/j.foodchem.2009.09.055)
4. Fang S, Yang W, Chu X, Shang X, She C, Fu X. Provenance and temporal variations in selected flavonoids in leaves of *Cyclocarya Paliurus*. *Food Chemistry*. 2011; 124(4):1382–6. doi: [10.1016/j.foodchem.2010.07.095](https://doi.org/10.1016/j.foodchem.2010.07.095)
5. Kurihara H, Fukami H, Kusumoto A, Toyoda Y, Shibata H, Matsui Y, et al. Hypoglycemic action of *Cyclocarya Paliurus* (Batal.) Iljinskaja in normal and diabetic mice. *Biosci Biotechnol Biochem*. 2003; 67(4):877–80. doi: [10.1271/bbb.67.877](https://doi.org/10.1271/bbb.67.877) PMID: [12784631](https://pubmed.ncbi.nlm.nih.gov/12784631/)
6. Xie MY, Li L, Nie SP, Wang XR, Lee FSC. Determination of speciation of elements related to blood sugar in bioactive extracts from *Cyclocarya Paliurus* leaves by FIA-ICP-MS. *European Food Research & Technology*. 2006; 223(2):202–9. doi: [10.1007/s00217-005-0173-0](https://doi.org/10.1007/s00217-005-0173-0)
7. Xie JH, Xie MY, Shen MY, Nie SP, Li C, Wang YX. Optimisation of microwave-assisted extraction of polysaccharides from *Cyclocarya Paliurus* (Batal.) Iljinskaja using response surface methodology. *J Sci Food Agric*. 2010; 90(8):1353–60. doi: [10.1002/jsfa.3935](https://doi.org/10.1002/jsfa.3935) PMID: [20474055](https://pubmed.ncbi.nlm.nih.gov/20474055/)
8. Yin ZP, Shangguan XC, Dong Ming LI, Shao Fu WU, Chen JG, Zhang YH. A Study on Ultrasonic-assisted Extraction of Total Triterpenoids from *Cyclocarya Paliurus* Leaves. *Acta Agriculturae Universitatis Jiangxiensis*. 2010; 32(2):373–7. doi: [10.13836/j.jjau.2010072](https://doi.org/10.13836/j.jjau.2010072)
9. Xie JH, Liu X, Shen MY, Nie SP, Zhang H, Li C, et al. Purification, physicochemical characterisation and anticancer activity of a polysaccharide from *Cyclocarya Paliurus* leaves. *Food Chem*. 2013; 136(3–4):1453–60. doi: [10.1016/j.foodchem.2012.09.078](https://doi.org/10.1016/j.foodchem.2012.09.078) PMID: [23194548](https://pubmed.ncbi.nlm.nih.gov/23194548/)
10. Zhang J, Shen Q, Lu JC, Li JY, Liu WY, Yang JJ, et al. Phenolic compounds from the leaves of *Cyclocarya Paliurus* (Batal.) Iljinskaja and their inhibitory activity against PTP1B. *Food Chemistry*. 2010; 119(4):1491–6. doi: [10.1016/j.foodchem.2009.09.031](https://doi.org/10.1016/j.foodchem.2009.09.031)
11. Li Tian JAU. The separation and identification of triterpenoids from leaves of *Cyclocarya Paliurus*. *Acta Agriculturae Universitatis Jiangxiensis*. 2013. doi: [10.13836/j.jjau.2013184](https://doi.org/10.13836/j.jjau.2013184)
12. Xie MY, Wang YX, Yi X, Wang XR. A Study on the Structure and Contents of Flavonoids in the Leaves of *Cyclocarya Paliurus* (Batal.) Iljinskaja. *Chinese Journal of Analytical Chemistry*. 2004; 32(8):1053–6.
13. Xie JH, Dong CJ, Nie SP, Li F, Wang ZJ, Shen MY, et al. Extraction, chemical composition and antioxidant activity of flavonoids from *Cyclocarya Paliurus* (Batal.) Iljinskaja leaves. *Food Chem*. 2015; 186:97–105. doi: [10.1016/j.foodchem.2014.06.106](https://doi.org/10.1016/j.foodchem.2014.06.106) PMID: [25976797](https://pubmed.ncbi.nlm.nih.gov/25976797/)
14. Xie MY, Xie JH. Review about the Research on *Cyclocarya Paliurus* (Batal.) Iljinskaja. *Journal of Food Science & Biotechnology*. 2008; 27(1):113–21.
15. Guo CL, Yang WY, Dong nan HU, Shangguan XC. A Study on Wand Cutting Techniques of *Cyclocarya Paliurus*. *Acta Agriculturae Universitatis Jiangxiensis*. 2006. doi: [10.13836/j.jjau.2006055](https://doi.org/10.13836/j.jjau.2006055)
16. Guo CL, Shangguan XC, Jiang Y, Yang WY, Zhang JW. A Study on the Effects of Growth Regulators on Cutting Techniques of *Cyclocarya Paliurus*. *Acta Agriculturae Universitatis Jiangxiensis*. 2008; 30(2):275–8. doi: [10.13836/j.jjau.2008056](https://doi.org/10.13836/j.jjau.2008056)
17. Hu DN, Shangguan XC, Liu LY, Xie F, Guo CL. In Vitro Culture of Stem Segments of *Cyclocarya Paliurus*. *Hubei Agricultural Sciences*. 2009. doi: [10.14088/j.cnki.issn0439-8114.2009.06.055](https://doi.org/10.14088/j.cnki.issn0439-8114.2009.06.055)

18. Yin ZP, Shangguan XC, Mi LX, Jiang Y, Wu SF, Zhang YH. Suspension culture of *Cyclocarya Paliurus* cells and accumulation of triterpenoids. *Journal of Shenzhen University Science & Engineering*. 2011; 28(5):430–5.
19. Chen JG, Shangguan XC, Yin ZP, Ren MH, Fu X. Establishment of the cell suspension culture system of *Cyclocarya Paliurus* and matrix consumption laws. *Modern Food Science & Technology*. 2014; 30(1):44–50. doi: [10.13982/j.mfst.1673-9078.2014.01.004](https://doi.org/10.13982/j.mfst.1673-9078.2014.01.004)
20. Yin ZP, Shangguan XC, Chen JG, Wu SF, Li DM. Separation and Identification of Triterpenic Acids from Suspended Cultured Cells of *Cyclocarya Paliurus*. *Scientia Silvae Sinicae*. 2013; 49(9):23–7.
21. Qian M, XiaoYing C, ChangGen L, Ting Y, XiaoChu L, JiHong J. Molecular Cloning and Functional Analysis of a Gene Encoding Farnesyl Diphosphate Synthase from *Cyclocarya Paliurus*. *Bulletin of Botanical Research*. 2011; 31((3)):323–9.
22. Marguerat S, Bahler J. RNA-seq: from technology to biology. *Cell Mol Life Sci*. 2010; 67(4):569–79. doi: [10.1007/s00018-009-0180-6](https://doi.org/10.1007/s00018-009-0180-6) PMID: [19859660](https://pubmed.ncbi.nlm.nih.gov/19859660/)
23. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009; 10(1):57–63. doi: [10.1038/nrg2484](https://doi.org/10.1038/nrg2484) PMID: [19015660](https://pubmed.ncbi.nlm.nih.gov/19015660/)
24. Spies D, Ciaudo C. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Comput Struct Biotechnol J*. 2015; 13:469–77. doi: [S2001-0370\(15\)00039-2](https://doi.org/S2001-0370(15)00039-2) PMID: [26430493](https://pubmed.ncbi.nlm.nih.gov/26430493/) doi: [10.1016/j.csbj.2015.08.004](https://doi.org/10.1016/j.csbj.2015.08.004)
25. Yang L, Ding G, Lin H, Cheng H, Kong Y, Wei Y, et al. Transcriptome analysis of medicinal plant *Salvia miltiorrhiza* and identification of genes related to tanshinone biosynthesis. *PLoS One*. 2013; 8(11): e80464. doi: [10.1371/journal.pone.0080464](https://doi.org/10.1371/journal.pone.0080464) PMID: [24260395](https://pubmed.ncbi.nlm.nih.gov/24260395/)
26. Liu MH, Yang BR, Cheung WF, Yang KY, Zhou HF, Kwok JS, et al. Transcriptome analysis of leaves, roots and flowers of *Panax notoginseng* identifies genes involved in ginsenoside and alkaloid biosynthesis. *BMC Genomics*. 2015; 16:265. doi: [10.1186/s12864-015-1477-5](https://doi.org/10.1186/s12864-015-1477-5) PMID: [25886736](https://pubmed.ncbi.nlm.nih.gov/25886736/)
27. Upadhyay S, Phukan UJ, Mishra S, Shukla RK. De novo leaf and root transcriptome analysis identified novel genes involved in steroidal sapogenin biosynthesis in *Asparagus racemosus*. *BMC Genomics*. 2014; 15:746. doi: [10.1186/1471-2164-15-746](https://doi.org/10.1186/1471-2164-15-746) PMID: [25174837](https://pubmed.ncbi.nlm.nih.gov/25174837/)
28. Huang HH, Xu LL, Tong ZK, Lin EP, Liu QP, Cheng LJ, et al. De novo characterization of the Chinese fir (*Cunninghamia lanceolata*) transcriptome and analysis of candidate genes involved in cellulose and lignin biosynthesis. *BMC Genomics*. 2012; 13:648. doi: [10.1186/1471-2164-13-648](https://doi.org/10.1186/1471-2164-13-648) PMID: [23171398](https://pubmed.ncbi.nlm.nih.gov/23171398/)
29. Zhang X, Allan AC, Li C, Wang Y, Yao Q. De Novo Assembly and Characterization of the Transcriptome of the Chinese Medicinal Herb, *Gentiana rigescens*. *Int J Mol Sci*. 2015; 16(5):11550–73. doi: [10.3390/ijms160511550](https://doi.org/10.3390/ijms160511550) PMID: [26006235](https://pubmed.ncbi.nlm.nih.gov/26006235/)
30. Chen J, Wu XT, Xu YQ, Zhong Y, Li YX, Chen JK, et al. Global transcriptome analysis profiles metabolic pathways in traditional herb *Astragalus membranaceus* Bge. var. *mongolicus* (Bge.) Hsiao. *BMC Genomics*. 2015; 16 Suppl 7:S15. doi: [10.1186/1471-2164-16-S7-S15](https://doi.org/10.1186/1471-2164-16-S7-S15) PMID: [26099797](https://pubmed.ncbi.nlm.nih.gov/26099797/)
31. Toshio M, Skoog F. A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. *Physiologia Plantarum*. 1962; 15(3):473–97. doi: [10.1111/j.1399-3054.1962.tb08052.x](https://doi.org/10.1111/j.1399-3054.1962.tb08052.x)
32. Yin ZP, Shangguan XC, Chen JG, Zhao Q, Li DM. Growth and triterpenic acid accumulation of *Cyclocarya Paliurus* cell suspension cultures. *Biotechnology & Bioprocess Engineering*. 2013; 18(3):606–14. doi: [10.1007/s12257-012-0751-5](https://doi.org/10.1007/s12257-012-0751-5)
33. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011; 29(7):644–52. doi: [10.1038/nbt.1883](https://doi.org/10.1038/nbt.1883) PMID: [21572440](https://pubmed.ncbi.nlm.nih.gov/21572440/)
34. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005; 21(18):3674–6. doi: [10.1093/bioinformatics/bti610](https://doi.org/10.1093/bioinformatics/bti610) PMID: [16081474](https://pubmed.ncbi.nlm.nih.gov/16081474/)
35. Zhou SM, Chen LM, Liu SQ, Wang XF, Sun XD. De Novo Assembly and Annotation of the Chinese Chive (*Allium tuberosum* Rottler ex Spr.) Transcriptome Using the Illumina Platform. *PLoS One*. 2015; 10(7):e0133312. doi: [10.1371/journal.pone.0133312](https://doi.org/10.1371/journal.pone.0133312) PMID: [26204518](https://pubmed.ncbi.nlm.nih.gov/26204518/)
36. Wang Y, Pan Y, Liu Z, Zhu X, Zhai L, Xu L, et al. De novo transcriptome sequencing of radish (*Raphanus sativus* L.) and analysis of major genes involved in glucosinolate metabolism. *BMC Genomics*. 2013; 14:836. doi: [10.1186/1471-2164-14-836](https://doi.org/10.1186/1471-2164-14-836) PMID: [24279309](https://pubmed.ncbi.nlm.nih.gov/24279309/)
37. Dutkowski J, Tiuryn J. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics*. 2007; 23(13):i149–58. doi: [10.1093/bioinformatics/btm194](https://doi.org/10.1093/bioinformatics/btm194) PMID: [17646291](https://pubmed.ncbi.nlm.nih.gov/17646291/)
38. Sun G, Yang Y, Xie F, Wen JF, Wu J, Wilson IW, et al. Deep sequencing reveals transcriptome re-programming of *Taxus x media* cells to the elicitation with methyl jasmonate. *PLoS One*. 2013; 8(4): e62865. doi: [10.1371/journal.pone.0062865](https://doi.org/10.1371/journal.pone.0062865) PMID: [23646152](https://pubmed.ncbi.nlm.nih.gov/23646152/)

39. Fukushima A, Nakamura M, Suzuki H, Saito K, Yamazaki M. High-Throughput Sequencing and De Novo Assembly of Red and Green Forms of the *Perilla frutescens* var. *crispa* Transcriptome. *PLoS One*. 2015; 10(6):e0129154. doi: [10.1371/journal.pone.0129154](https://doi.org/10.1371/journal.pone.0129154) PMID: [26070213](https://pubmed.ncbi.nlm.nih.gov/26070213/)
40. Li SW, Shi RF, Leng Y. De Novo Characterization of the Mung Bean Transcriptome and Transcriptomic Analysis of Adventitious Rooting in Seedlings Using RNA-Seq. *PLoS One*. 2015; 10(7):e0132969. doi: [10.1371/journal.pone.0132969](https://doi.org/10.1371/journal.pone.0132969) PMID: [26177103](https://pubmed.ncbi.nlm.nih.gov/26177103/)
41. He L, Xu X, Li Y, Li C, Zhu Y, Yan H, et al. Transcriptome analysis of buds and leaves using 454 pyrosequencing to discover genes associated with the biosynthesis of active ingredients in *Lonicera japonica* Thunb. *PLoS One*. 2013; 8(4):e62922. doi: [10.1371/journal.pone.0062922](https://doi.org/10.1371/journal.pone.0062922) PMID: [23638167](https://pubmed.ncbi.nlm.nih.gov/23638167/)
42. Hu L, Li H, Chen L, Lou Y, Amombo E, Fu J. RNA-seq for gene identification and transcript profiling in relation to root growth of bermudagrass (*Cynodon dactylon*) under salinity stress. *BMC Genomics*. 2015; 16:575. doi: [10.1186/s12864-015-1799-3](https://doi.org/10.1186/s12864-015-1799-3) PMID: [26238595](https://pubmed.ncbi.nlm.nih.gov/26238595/)
43. Zhang F, Gonzalez A, Zhao M, Payne CT, Lloyd A. A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. *Development*. 2003; 130(20):4859–69. doi: [10.1242/dev.00681](https://doi.org/10.1242/dev.00681) PMID: [12917293](https://pubmed.ncbi.nlm.nih.gov/12917293/)
44. Yamada Y, Kokabu Y, Chaki K, Yoshimoto T, Ohgaki M, Yoshida S, et al. Isoquinoline alkaloid biosynthesis is regulated by a unique bHLH-type transcription factor in *Coptis japonica*. *Plant Cell Physiol*. 2011; 52(7):1131–41. doi: [10.1093/pcp/pcr062](https://doi.org/10.1093/pcp/pcr062) PMID: [21576193](https://pubmed.ncbi.nlm.nih.gov/21576193/)
45. Todd AT, Liu E, Polvi SL, Pammett RT, Page JE. A functional genomics screen identifies diverse transcription factors that regulate alkaloid biosynthesis in *Nicotiana benthamiana*. *Plant J*. 2010; 62(4):589–600. doi: [10.1111/j.1365-3113.2010.04186.x](https://doi.org/10.1111/j.1365-3113.2010.04186.x) PMID: [20202168](https://pubmed.ncbi.nlm.nih.gov/20202168/)
46. Nims E, Vongpaseuth K, Roberts S, Walker E. TcJAMYC: a bHLH transcription factor that activates paclitaxel biosynthetic pathway genes in yew. *J Biol Chem*. 2015; 290(33):20104. doi: [10.1074/jbc.A109.026195](https://doi.org/10.1074/jbc.A109.026195) PMID: [26276858](https://pubmed.ncbi.nlm.nih.gov/26276858/)
47. Ikegami A, Eguchi S, Kitajima A, Inoue K, Yonemori K. Identification of genes involved in proanthocyanidin biosynthesis of persimmon (*Diospyros kaki*) fruit. *Plant Science*. 2007; 172(5):1037–47. doi: [10.1016/j.plantsci.2007.02.010](https://doi.org/10.1016/j.plantsci.2007.02.010)
48. Li TT, Shi YR, Wu CE, Fang SZ, Pan HX, Zhu ZJ. Analysis of chemical and volatile components of *Cyclocarya Paliurus* tea. *Food & Machinery*. 2015; 31(4):10–3. doi: [10.13652/j.issn.1003-5788.2015.04.003](https://doi.org/10.13652/j.issn.1003-5788.2015.04.003)
49. Chen JH, Huang JA, Liu ZH. Analysis on chemical constituents in the leaves of *Cyclocarya Paliurus*. *Science & Technology of Food Industry*. 2009; 30(7):159–61. doi: [10.13386/j.issn1002-0306.2009.07.099](https://doi.org/10.13386/j.issn1002-0306.2009.07.099)
50. Li X, I. M, Shangguan XC, Shi LX, Yin ZP, Zhou BY. Studies on the Determination and Distribution of Total Flavonoids of *Cyclocarya Paliurus* Vegetative Organs. *Acta Agriculturae Universitatis Jiangxiensis*. 2009; 31(5):896–900. doi: [10.13836/j.jjau.2009176](https://doi.org/10.13836/j.jjau.2009176)
51. Li S, Li J, Guan XL, Li J, Deng SP, Li LQ, et al. Hypoglycemic effects and constituents of the barks of *Cyclocarya Paliurus* and their inhibiting activities to glucosidase and glycogen phosphorylase. *Fitoterapia*. 2011; 82(7):1081–5. doi: [10.1016/j.fitote.2011.07.002](https://doi.org/10.1016/j.fitote.2011.07.002) PMID: [21784137](https://pubmed.ncbi.nlm.nih.gov/21784137/)
52. Cone KC, Cocciolone SM, Burr FA, Burr B. Maize anthocyanin regulatory gene *pl* is a duplicate of *c1* that functions in the plant. *Plant Cell*. 1993; 5(12):1795–805. doi: [10.1105/tpc.5.12.1795](https://doi.org/10.1105/tpc.5.12.1795) PMID: [8305872](https://pubmed.ncbi.nlm.nih.gov/8305872/)
53. Borevitz JO, Xia Y, Blount J, Dixon RA, Lamb C. Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell*. 2000; 12(12):2383–94. PMID: [11148285](https://pubmed.ncbi.nlm.nih.gov/11148285/)
54. Spelt C, Quattrocchio F, Mol JN, Koes R. anthocyanin1 of petunia encodes a basic helix-loop-helix protein that directly activates transcription of structural anthocyanin genes. *Plant Cell*. 2000; 12(9):1619–32. PMID: [11006336](https://pubmed.ncbi.nlm.nih.gov/11006336/)
55. Sweeney MT, Thomson MJ, Pfeil BE, McCouch S. Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell*. 2006; 18(2):283–94. doi: [10.1105/tpc.105.038430](https://doi.org/10.1105/tpc.105.038430) PMID: [16399804](https://pubmed.ncbi.nlm.nih.gov/16399804/)
56. Khanna R, Huq E, Kikis EA, Al-Sady B, Lanzatella C, Quail PH. A novel molecular recognition motif necessary for targeting photoactivated phytochrome signaling to specific basic helix-loop-helix transcription factors. *Plant Cell*. 2004; 16(11):3033–44. doi: [10.1105/tpc.104.025643](https://doi.org/10.1105/tpc.104.025643) PMID: [15486100](https://pubmed.ncbi.nlm.nih.gov/15486100/)
57. Chinnusamy V, Ohta M, Kanrar S, Lee BH, Hong X, Agarwal M, et al. ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis. *Genes Dev*. 2003; 17(8):1043–54. doi: [10.1101/gad.1077503](https://doi.org/10.1101/gad.1077503) PMID: [12672693](https://pubmed.ncbi.nlm.nih.gov/12672693/)
58. Carretero-Paulet L, Galstyan A, Roig-Villanova I, Martinez-Garcia JF, Bilbao-Castro JR, Robertson DL. Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in

- Arabidopsis, poplar, rice, moss, and algae. *Plant Physiol.* 2010; 153(3):1398–412. doi: [10.1104/pp.110.153593](https://doi.org/10.1104/pp.110.153593) PMID: [20472752](https://pubmed.ncbi.nlm.nih.gov/20472752/)
59. Pires N, Dolan L. Origin and diversification of basic-helix-loop-helix proteins in plants. *Mol Biol Evol.* 2010; 27(4):862–74. doi: [10.1093/molbev/msp288](https://doi.org/10.1093/molbev/msp288) PMID: [19942615](https://pubmed.ncbi.nlm.nih.gov/19942615/)
 60. Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L. The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in Arabidopsis siliques. *Plant Cell.* 2000; 12(10):1863–78. PMID: [11041882](https://pubmed.ncbi.nlm.nih.gov/11041882/)
 61. Hu J, Anderson B, Wessler SR. Isolation and characterization of rice R genes: evidence for distinct evolutionary paths in rice and maize. *Genetics.* 1996; 142(3):1021–31. PMID: [8849907](https://pubmed.ncbi.nlm.nih.gov/8849907/)
 62. Ludwig SR, Habera LF, Dellaporta SL, Wessler SR. Lc, a member of the maize R gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the myc-homology region. *Proc Natl Acad Sci U S A.* 1989; 86(18):7092–6. PMID: [2674946](https://pubmed.ncbi.nlm.nih.gov/2674946/)
 63. Furukawa T, Maekawa M, Oki T, Suda I, Iida S, Shimada H, et al. The Rc and Rd genes are involved in proanthocyanidin synthesis in rice pericarp. *Plant J.* 2007; 49(1):91–102. doi: [10.1111/j.1365-313X.2006.02958.x](https://doi.org/10.1111/j.1365-313X.2006.02958.x) PMID: [17163879](https://pubmed.ncbi.nlm.nih.gov/17163879/)
 64. Ohno S, Hosokawa M, Hoshino A, Kitamura Y, Morita Y, Park KI, et al. A bHLH transcription factor, Dv1VS, is involved in regulation of anthocyanin synthesis in dahlia (*Dahlia variabilis*). *J Exp Bot.* 2011; 62(14):5105–16. doi: [10.1093/jxb/err216](https://doi.org/10.1093/jxb/err216) PMID: [21765172](https://pubmed.ncbi.nlm.nih.gov/21765172/)
 65. Yuan Y, Chiu LW, Li L. Transcriptional regulation of anthocyanin biosynthesis in red cabbage. *Planta.* 2009; 230(6):1141–53. doi: [10.1007/s00425-009-1013-4](https://doi.org/10.1007/s00425-009-1013-4) PMID: [19756724](https://pubmed.ncbi.nlm.nih.gov/19756724/)
 66. Katiyar A, Smita S, Lenka SK, Rajwanshi R, Chinnusamy V, Bansal KC. Genome-wide classification and expression analysis of MYB transcription factor families in rice and Arabidopsis. *BMC Genomics.* 2012; 13:544. doi: [10.1186/1471-2164-13-544](https://doi.org/10.1186/1471-2164-13-544) PMID: [23050870](https://pubmed.ncbi.nlm.nih.gov/23050870/)
 67. Paz-Ares J, Ghosal D, Wienand U, Peterson PA, Saedler H. The regulatory c1 locus of *Zea mays* encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *EMBO J.* 1987; 6(12):3553–8. PMID: [3428265](https://pubmed.ncbi.nlm.nih.gov/3428265/)
 68. Grotewold E, Athma P, Peterson T. Alternatively spliced products of the maize P gene encode proteins with homology to the DNA-binding domain of myb-like transcription factors. *Proc Natl Acad Sci U S A.* 1991; 88(11):4587–91. PMID: [2052542](https://pubmed.ncbi.nlm.nih.gov/2052542/)
 69. Grotewold E, Chamberlin M, Snook M, Siame B, Butler L, Swenson J, et al. Engineering secondary metabolism in maize cells by ectopic expression of transcription factors. *Plant Cell.* 1998; 10(5):721–40. PMID: [9596632](https://pubmed.ncbi.nlm.nih.gov/9596632/)
 70. Quattrocchio F, Wing J, van der Woude K, Souer E, de Vetten N, Mol J, et al. Molecular analysis of the anthocyanin2 gene of petunia and its role in the evolution of flower color. *Plant Cell.* 1999; 11(8):1433–44. PMID: [10449578](https://pubmed.ncbi.nlm.nih.gov/10449578/)
 71. Yi X, Shi JG, Zhou GX, Xie MY. Studies on the Chemical Constituents in the Leaves of *Cyclocarya Paliurus*. *Chin J Chin Mater Med.* 2002; 27(1):43–5.
 72. Li J, Lu YY, Li F, Zhang YJ, Huang XS, Su Xj. Study on chemical constituents of *Cyclocarya Paliurus*. *Journal of Chinese Medicinal Materials.* 2006; 29(5):441–2. PMID: [16981452](https://pubmed.ncbi.nlm.nih.gov/16981452/)
 73. Li J, Lu YY, Xu ZJ, Li F, Wen GQ, Su XJ. Studies on flavones in the leaves of *Cyclocarya paliurus* (batal.) Iljinsk. *Journal of Chinese medicinal materials.* 2005; 28(12):1058–9. doi: [10.13863/j.issn1001-4454.2005.12.006](https://doi.org/10.13863/j.issn1001-4454.2005.12.006) PMID: [16568659](https://pubmed.ncbi.nlm.nih.gov/16568659/)
 74. Shu RG, Song ZR, Shu JC. Study on the Chemical Constituents of the Butanol Extraction of *Cyclocarya Paliurus* (Batal.) Iljinsk. *Journal of Chinese Medicinal Materials.* 2006; 29(12):1304–7. doi: [10.13863/j.issn1001-4454.2006.12.017](https://doi.org/10.13863/j.issn1001-4454.2006.12.017)
 75. Zhang XQ, Ye WC, Yin ZQ, Zhang ZH, Zhao SX. Studies on the Chemical Constituents in the Leaves of *Cyclocarya Paliurus*. *China Journal of Chinese Materia Medica.* 2005; 30(10):791–2.
 76. He Y, Yin ZQ, Zhang J, Ye WC, Fang SZ. Chemical Constituents from the Aerial Parts of *Cyclocarya Paliurus*. *Pharmaceutical and Clinical Research.* 2012; 20(3):187–9. doi: [10.3969/j.issn.1673-7806.2012.03.004](https://doi.org/10.3969/j.issn.1673-7806.2012.03.004)
 77. Zhang J. Studies on water-soluble constituents in the leaves of *Cyclocarya Paliurus*. *Journal of Pharmaceutical Practice.* 2007; 25(2):82–4.