PLOS ONE

RESEARCH ARTICLE

The Transcript Profile of a Traditional Chinese Medicine, *Atractylodes lancea*, Revealing Its Sesquiterpenoid Biosynthesis of the Major Active Components

Shakeel Ahmed[1,2,3,4], Chuansong Zhan[1,2,3,4], Yanyan Yang[1,3,4], Xuekui Wang[1,3,4], Tewu Yang[1,3,4], Zeying Zhao[1,3,4], Qiyun Zhang[1,2,3,4], Xiaohua Li[1,2,3,4], Xuebo Hu[1,2,3,4]*

1 Department of Medicinal Plant, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, 430070, P.R. China, **2** Center for Plant Functional Components, Huazhong Agricultural University, Wuhan, 430070, P.R. China, **3** National-Regional Joint Engineering Research Center in Hubei for Medicinal Plant Breeding and Cultivation, Huazhong Agricultural University, Wuhan, 430070, P.R. China, **4** Engineering Research Center for Medicinal Plants, Huazhong Agricultural University, Wuhan, 430070, P. R. China

* xuebohu@mail.hzau.edu.cn

Citation: Ahmed S, Zhan C, Yang Y, Wang X, Yang T, Zhao Z, et al. (2016) The Transcript Profile of a Traditional Chinese Medicine, *Atractylodes lancea*, Revealing Its Sesquiterpenoid Biosynthesis of the Major Active Components. PLoS ONE 11(3): e0151975. doi:10.1371/journal.pone.0151975

Editor: Ji-Hong Liu, Key Laboratory of Horticultural Plant Biology (MOE), CHINA

Received: January 3, 2016

Accepted: March 7, 2016

Published: March 18, 2016

Data Availability Statement: The raw Illumina sequencing data of A. lancea was submitted to NCBI Sequence Read Archive (SRA), and the accession number is SRP070094. The assembled contigs for samples from leaves, roots and stems are loaded into TSA database with accession numbers GEFW01000000, GEFZ01000000, GEGA00000000.

Funding: This research was supported by the Fundamental Research Funds for the Central Universities Program No. 2014PY057 (XH) and Natural Science Foundation of Hubei Province, China, No. 2015CFA091 (XH). The funders had no

Abstract

Atractylodes lancea (Thunb.) DC., named "Cangzhu" in China, which belongs to the Asteraceae family. In some countries of Southeast Asia (China, Thailand, Korea, Japan etc.) its rhizome, commonly called rhizoma atractylodis, is used to treat many diseases as it contains a variety of sesquiterpenoids and other components of medicinal importance. Despite its medicinal value, the information of the sesquiterpenoid biosynthesis is largely unknown. In this study, we investigated the transcriptome analysis of different tissues of non-model plant *A*. *lancea* by using short read sequencing technology (Illumina). We found 62,352 high quality unigenes with an average sequence length of 913 bp in the transcripts of *A*. *Lancea*. Among these, 43,049 (69.04%), 30,264 (48.53%), 26,233 (42.07%), 17,881 (28.67%) and 29,057(46.60%) unigenes showed significant similarity (E-value<$1e^{-5}$) to known proteins in Nr, KEGG, SWISS-PROT, GO, and COG databases, respectively. Of the total 62,352 unigenes, 43,049 (Nr Database) open reading frames were predicted. On the basis of different bioinformatics tools we identify all the enzymes that take part in the terpenoid biosynthesis as well as five different known sesquiterpenoids via cytosolic mevalonic acid (MVA) pathway and plastidal methylerythritol phosphate (MEP) pathways. In our study, 6, 864 Simple Sequence Repeats (SSRs) were also found as great potential markers in *A*. *lancea*. This transcriptomic resource of *A*. *lancea* provides a great contribution in advancement of research for this specific medicinal plant and more specifically for the gene mining of different classes of terpenoids and other chemical compounds that have medicinal as well as economic importance.

role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

The plant *Atractylodes lancea* (Thunb.) DC., known as "Cangzhu" in China, "Khod-Kha-Mao" in Thailand [1] and its name in Japan is "So-ju-tsu"[1, 2]. *A. lancea* belongs to the Asteraceae family. The rhizome of *A. lancea*, generally called rhizoma atractylodes is used for treatment of influenza, rheumatic diseases, night blindness and a few digestive problems [3–5]. The history of using rhizomes of *A. lancea* as a drug can be traced back to Han dynasty (206BC-220AD), when it was described in Shen-nong-ben-cao-jing, the first Chinese pharmacopoeia. Later it was found that this herb include two species, *A. lancea* and *A. Chinensis* (DC.) Koids, known "Mao CangZhu" and "Bei CangZhu" separately in China and people have used these together as rhizoma Atractylodes [6].

Previous reports imply terpenoids and their glycosidal derivatives are the major active components [7, 8]. Terpenoids are the natural products that are mostly present in plants with specific structures [9]. Plant-derived Terpenoids show a diversity of medicinal effects that comprise of multiple industrial and pharmaceutical applications including antiparasitic, anticancer, antifungal, antiviral and antibacterial activities. Terpenoids are grouped as monoterpenoids, sesquiterpenoids, diterpenoids, triterpenoids, and others [10]. In general, terpenoids are synthesized in plants by way of MVA pathway and MEP pathway. In the MVA pathway, terpenoid is synthesized starting from primary metabolic product acetyl-CoA to crucial precursors such as isopentyl diphosphate (IPP) and dimethylallyl diphosphate. The reaction is catalyzed by a large variety of enzymes with special product specificities (Fig 1). In the MEP pathway, glycolysis products glyceraldehydyde-3-phosphate and pyruvate are catalyzed into 1-deoxy-D-xylulose-5-phosphate. After a few enzymatic steps the pathway runs into the same chemicals as MVA pathway (Fig 1). The 1-deoxy-D-xylulose-5-phosphate synthase (DXS) and hydroxymethylglutaryl-CoA synthase (HMGR), the rate-limiting enzymes in MEP and MVA pathway, respectively, are usually encoded by a group of small multigene families [11]. Sesquiterpenes synthase are universally expressed family of different proteins which are able to convert the universal precursor farnesyl diphosphate (FPP) into more than three hundred various sesquiterpenes skeletons [12]. The current study mainly emphasises on the biosynthetic pathway of sesquiterpenoids in *A. lancea*.

Sesquiterpenoids have a wide variety of benefits such as pharmaceuticals, flavors, fragrances, industrial chemicals and nutraceuticals [13–15]. β-caryophyllene is an essential sesquiterpene that is present in different essential oils of many plants like cinnamon (*Cinnamomum cassia*), thyme (*Thymus mongolicus*), clove (*Syringa*spp.) and black pepper (*Piper nigrum*), of which mostly have been used for cure of different health problems as well as for fragrances [16, 17]. It is also prominently used as anti-carcinogenic & anti-microbial antioxidant, as well as skin penetration enhancer [18]. Germacrene D is another kind of sesquiterpene. It is a chiral compound, which is produced from FPP by enantionmers particular synthase [19]. Germacrene D has a sturdy effect on insect activities [20]. FPP can be converted into cyclic sesquiterpene, (E)-β-farnesene, which is catalyzed by (E)-β-farnesene synthase (β-FS)[21]. (E)-β-farnesene occurs in a variety of plants and animals & is widely used as a semio-chemical in insects and plants [22].

Sesiquiterpenes are the major components of the volatile essential oil from *A. Lancea*. In an effort to identify the chemical profiles of essential oil from *A. Lancea*, the wild grown plants produced mostly significant amount of sesiquiterpenes with the top three hinesol (68.5%), β-eudesmol (13.1%) and elemol (6.2%)[23]. However, the content of these chemicals is greatly influenced by the geographic location where the sample was taken from [23, 24], as *A. Lancea* is widely distributed in the vast area between Yellow River and Yangtze River of China. Among these diverse sesiquiterpenes, atractylenolides (I, II & III), atractylon, biatractylolide screened

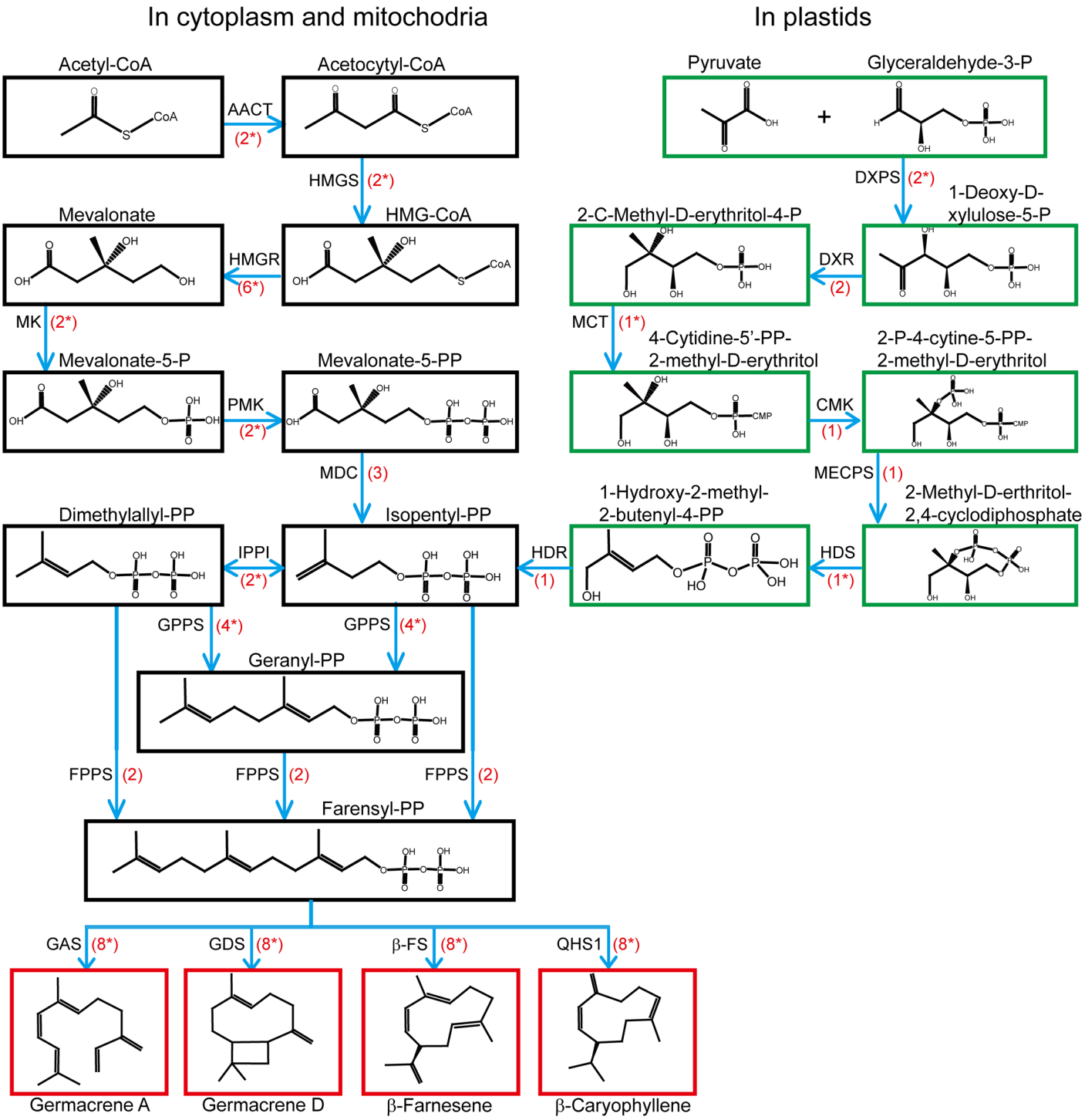

Fig 1. Putative sesquiterpenoid biosynthetic pathway in *Atractylodes lancea*. A flow diagram of biosynthetic pathway of terpenoid backbone and sesquiterpenoids biosynthesis in *Atractylodes lancea*. The structures of chemicals in the pathway are shown in boxes. The green boxes represent the plasticidal pathway while the black boxes show the pathway in cytoplasm & mitochondria. The words on the boxes are enzymes for the reaction while the numbers in red color represent the number of transcripts for that specific gene. Reactions in cytoplasm, mitochondria and plastids are shown in green. The boxes with red border show the structure of various sesquiterpenoids of *A. lancea*.

doi:10.1371/journal.pone.0151975.g001

from *A*. *Lancea* were demonstrated to having a good protection against ethanol-induced gastric ulcer [25]. Atractylenolides were also proved to be insect repellents [26]. Recently it was found that a new sesquiterpenoid, hinesol, was responsible for the apoptosis in human cancer cells. On the contrary, the activity of β-eudesmol, a more commonly found sesiquiterpene in other medicinal plants also available from *A*. *Lancea*, was less effective as compared to hinesol [27]. There are other kinds of sesquiterpenes as guaiane, eudesmane, tricyclic carbon skeleton types, but the physiological activity remained to be elucidated [28]. Furthermore, it was proved that some other chemicals from *A*. *Lancea*, like atractylochromene, methylphenol derivatives, cyclohexadiene derivatives, polyacetylenes, atractylodin and acidic polysaccharides, showed diverse activities against inflammation, bacterial, fungi or obesity, but their structures are different from sesquiterpenes [29–31]. The chemical and structural diversity in *A*. *Lancea* correlates with its multiplex medicinal functions. Owing to recent advances in molecular biology and decreasing cost of next generation sequencing technology, RNA sequencing (RNA-seq) become a popular choice for the transcriptome studies especially in non-model species [32]. Consequently, RNA-seq has been extensively deployed in various TCM species, for example Chinese sage (*Salvia miltiorrhiza*)[33], Chinese Ginseng (*Panax ginseng*)[34] and Sanchi (*Panax notoginseng*). [35]. Deep transcriptome analysis also helps to discover various genetic profile, including alternative splicing isoforms [36], strand-specific expression [37] and microRNA discovery [38]. With the help of transcriptome sequencing, comprehensive information can be obtained on gene expression, molecular mechanisms and biological pathways, even in the absence of reference genome [39–43]. However, to date the study of *A*. *lancea* transcriptome is not reported yet. Here we report on the Illumina transcriptome sequencing, functional annotation and differential expression profiles in different tissues i.e. stem root, and leaf of *A*. *lancea* which will be an important resource for gene mining, genetic improvement and development of different molecular markers. Additionally, to further explore the differences of candidate unigenes in terpenoid biosynthesis among these *A*. *lancea* tissues, the transcriptional levels of all the related unigenes were concretely discussed. The results from our work could contribute to the discovery of genes dedicated to the terpenoid pathway and its accumulative regulation of volatile constituents in specific tissues of *A*. *lancea*. According to our information, the current research work is first report of secondary metabolic analysis in *A*. *lancea* based on *de novo* transcriptome analysis.

Materials and Methods

Collection of the *A*. *lancea* tissues

Prior to the experiment, the Institute of Science & Technology Development of HZAU university assured us that no specific permission is needed for the field experiment with *A*. *Lancea* in Hubei Province as it is commonly planted in China as a medicine sources. With the permission from Hubei Jintuyuan Forest Medicine & Seed Co. Ltd. (52 Jinyuanbao Avenue, Yuanbao, Lichuan City, Enshi autonomous district, Hubei province of China), experimental materials of *A*. *Lancea* was taken from a herbal medicine planting field (E08°56′, N30°18′) belongs to the company. The roots, stems and leaves of *A*. *lancea* were immediately frozen in liquid nitrogen after collection until use. The *A*. *lancea* was authenticated by Prof. Xuebo Hu, Assoc. Prof. Tewu Yang and Xuekui Wang.

cDNA library preparation and sequence data analysis and assembly

To extract the total RNA present, equivalent weight of three tissue samples were mixed by using RNeasy Plant Mini Kits (Qiagen, Inc., Valencia, CA, USA) according to the manufacturer's protocol. All the samples of extracted RNA were qualified and quantified using a

Nanodrop ND-1000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA), they showed a 260/280 nm ratio from 1.9 to 2.1. No sign of degradation was found when RNA samples were analyzed by electrophoresis. Transcriptome analysis was done by taking equal amounts of all the three samples by using Illumina's kit following manufacturer's protocol. Briefly, the poly-(A) mRNA was purified from the total RNA by Oligotex mRNA Mini Kit (Qiagen, Inc., Valencia, CA, USA) following the manufacturer's protocol. The cDNA library construction and normalization were performed using protocols described previously [44].

Transcriptome *de novo* assembly

Trinity, a short read assembly package after sequencing was used for assembling Transcriptome mechanism, which consists of Inch-worm, a huge amount of RNA-seq reads were generated when processed sequentially by Chrysalis and Butterfly programs [45]. Consequent analysis of clean reads was carried out once they were filtered from the raw reads. Inchworms were the first to be used to assemble short reads with over-lapping sequences having longest contigs without gaps. Each cluster was used to construct a full de Bruijn graph after the clusters were grouped. Reads and pairs of reads were compared in equivalence to outline the pathways they had common. On the other hand full length transcripts were spliced isoforms, matching to paralogous genes, were generated by splicing apart transcripts. All such sequences from Trinity were defined as unigenes. In this study three samples of *A*. *lancea* were sequenced, sequence splicing was carried out for unigenes from each sample. Excess unigenes are separated from the required unigenes by using sequence clustering software. Unigenes are grouped into two classes after clustering genes into families: clusters (prefixed by CL) and single-tons (prefixed by unigene). Finally, we carried out alignment via BLASTx (E.value p 0.00001) between unigenes and protein databases of NR, Swiss-Prot, KEGG, and COG, and the course of unigene sequence was by using the best aligned results. If there is an incongruity among various databases, a priority order of NR, Swiss-Prot, KEGG, and COG was used to check the direction of the sequence. The unigenes whose sequences could not be determined by the above data base were aligned and their sequence directions determined using ESTScan [46].

Unigene differential expression analysis

Differential expression of gene function was performed using gene ontology (GO) functional analysis, and these differentially expressed genes were mapped in each term using GO database (http://www.geneontology.org) and then correspondent number of gene with each GO term was determined. Following the creation of gene list which includes the number of genes linked with every GO term, the significance of GO enriched in differentially expressed gene in comparison with genomic background hyper-geometric test was applied.

SSRs mining and primer design

SSRs consist of one to six nucleotide motifs, having minimum five tandem repeats. We used Microsatellite (MISA) detection tool for SSRs mining [47] and we design primer pairs using software primer3 (V.2.3.6) for each SSRs under default settings, with a range in the size of products of PCR from 100–250 bp [48, 49].

Results and Discussion

A. *lancea* transcriptome sequencing and unigene assembly

To clarify a comprehensive overview of gene expression profiles in *A*. *lancea* tissues, the construction of cDNA libraries were made from different samples of leaf, root and stem of

A. lancea, respectively and sequenced by the Illumina transcriptome platform in our experiments. After removal of adaptor sequences and low quality reads, a total of 43,921,277, 37,866,604 and 40,135,278 clean reads were acquired from leaf, root and stem tissues, respectively (Table 1). These data sizes are bigger than those from peanut (*Arachis hypogaea*) [44], yellow horn (*Xanthoceras sorbifolium*) [50], siberian apricot (*Prunus sibirica*) [51] and Centella (*Centella asiatica*) [52], suggesting that the relatively complete read databases were successfully constructed from different tissues of *A. lancea* by Illumina sequencing. Subsequently, Trinity software was used for assembly of these clean reads (Trinityrnaseq_r2013_08_14) and low density and quality reads were filtered out, resulting in 64,106, 55,409 and 56,565 unigenes in the leaves, root and stem respectively. After *de novo* assembly of three *A. lancea* tissues, 62,352 unigenes were finally obtained with an average length of 913 bp (Table 1). Among these, 42127 unigenes having a length range between 300 nt to 1000 nt and 15263 unigenes having a length longer than 1 kb (>1000 nt) as shown in Fig 2. Furthermore, we found that the sum of unigenes (62,352) in *A. lancea* is more than the identified number of unigenes 59,236 in peanut (*A. hypogaea*) [44], 51,867 unigenes in yellow horn (*X. sorbifolium*) [50] and 46,940 unigenes in Siberian apricot (*P. sibirica*) [51].

Functional annotation of *A. lancea* unigenes

The species distribution of the non-redundant (Nr) annotation is shown in Fig 3. There was 23.79% of unigenes shown the highest homology to genes from grape (*Vitis vinifera*), 9.5% of unigenes matched to potato (*Solanum tuberosum*), 8.1% of unigenes matched to cacao (*Theobroma cacao*), 6.6% of unigenes matched to tomato (*Solanum lycopersicum*) and 5.8% & 5.1% of unigenes matched to populus (*Populus trichocarpa*) and peach (*Prunus persica*), respectively. All the *A. lancea* unigenes from different tissues were predicted via BLAST (basic local alignment search tool) with a cut-off E-value of 10^{-5} in public databases such as non-redundant (NR), SWISS-PROT, kyoto encyclopedia of genes and genomes (KEGG), classification of Orthologous Group (COG), and gene ontology (GO), which retrieved higher sequence similarity proteins among specific unigenes beside their functional annotations. From the BLAST results, a total of 43,049 (69.04%), 30,264 (48.53%), 26,233 (42.07%), 17,881(28.67%) and 29,057(46.60%) unigenes showed diverse similarity to well-known proteins in above mentioned databases, respectively (Table 2). However, 44,482 unigene (71.34%) sequence orientations are still unknown, which is higher than the peanut (*A. hypogaea*) (27.8%) [44] but lower than that of Chinese tulip tree (*Liriodendron chinense*) (73.60%) [53]. This is because of the lack of *A. lancea* genomic information, and few or no effective characterized protein domains of the shorter sequences for getting BLAST hits. Also, it is possible that some un-matched unigenes are the novel genes specific for *A. lancea*.

Table 1. Statistic of sequencing and de novo assembling of transcriptome in *Atractylodes lancea*.

	Sample	Total number	Total length (nt)	Mean Length (nt)	N50	Total consensus sequences	Distinct Clusters	Distinct Singletons
Contigs	Leaf	112883	42287508	375	806	0	0	0
	Root	94663	37505566	396	837	0	0	0
	Stem	101679	39492194	388	359	0	0	0
Unigenes	Leaf	64106	43921277	685	1258	64106	19718	44388
	Root	55409	37866604	684	1221	55409	16802	38607
	Stem	56565	40135278	710	1328	56565	16947	39618
Total		62352	56923290	913	1494	62352	23974	38378

doi:10.1371/journal.pone.0151975.t001

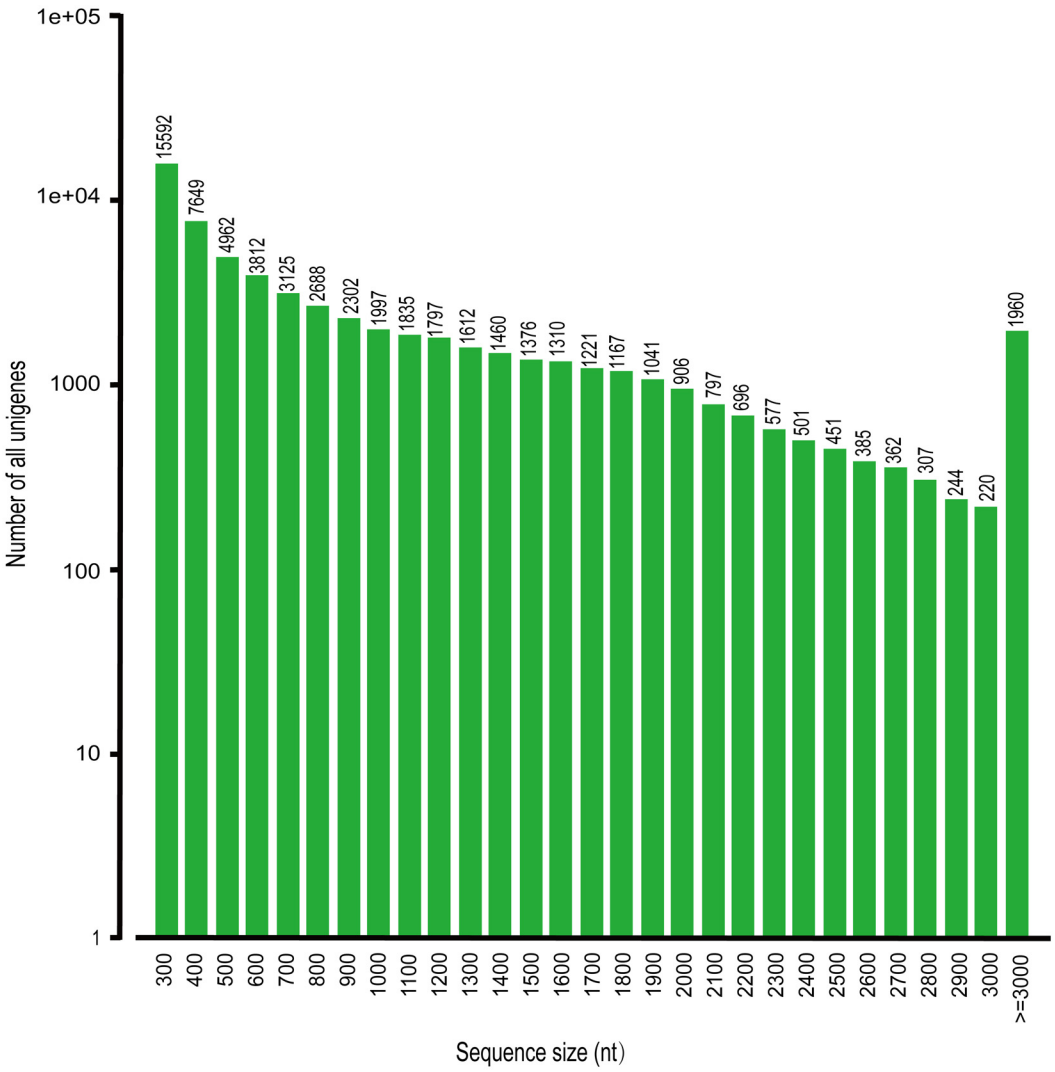

Fig 2. Length distribution of unigenes in *Atractylodes lancea*. The x-axis represent the size of the all assembled sequences and the y-axis indicates the corresponding number of unigenes.

doi:10.1371/journal.pone.0151975.g002

Functional classification of *A. lancea* unigenes by GO, COG and KEGG

For categorizing the function of predicted *A. lancea* unigenes gene, ontology (GO) annotation was used [54]. In total, 29,057 unigenes were selected for three main GO categories and 56 sub-categories (Fig 4). It shows that "metabolic process", "cellular process", "binding" and "catalytic activity" are the most dominant category involving more than 180,000 unigenes, while a small portion of genes were linked with terms such as "pigmentation", "receptor regulator activity" and "protein tag". It is interesting to observe that 20,169 unigenes from GO analysis had not been annotated in the Swiss-Prot database, which could be explained by the fact that the prediction quality could be significantly improved by GO annotation as the clustering of proteins determine their sub cellular locations reflection in a better way [55].

To further expose the value of annotation process and predict possible functions of unigenes, we looked for the annotated sequences for genes involved in the classification of orthologous group (COG) to classify the orthologous products of genes [56]. COG database was used

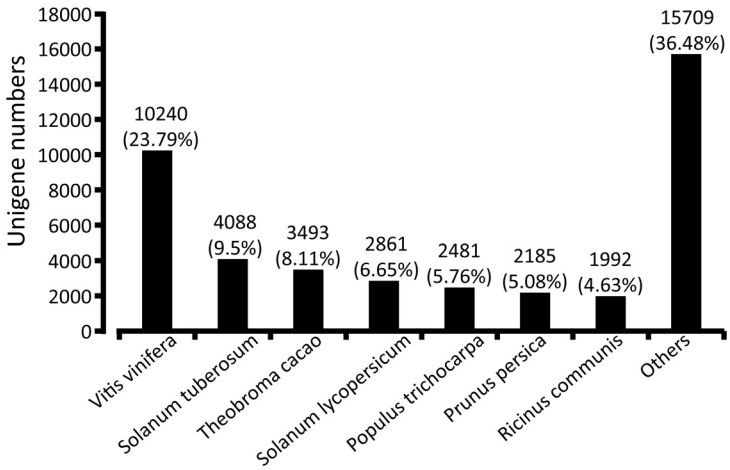

Fig 3. The species distribution of the non-redundant unigene annotation. The column shows the homology of *Atractylodes lancea* unigene number with that from other species. The numbers inside parentheses indicate the percentage of the homology to different species.

doi:10.1371/journal.pone.0151975.g003

for the alignment and for prediction and classification of possible function of all *A*. *lancea* unigenes. Results revealed that 17,881 unigenes were recognized as 25 COG classifications (Fig 5). In 25 COG categories, the largest group represents "general function prediction (5837 unigenes)", second cluster was 'transcription' (3105 unigenes) and then 'replication, recombination & repair' (2732 unigenes). It was also observed that just a few genes found related to the terms as "extracellular structures" and "Nuclear structures".

For further recognition of the interaction and biological functions of genes in the *A*. *lancea*, KEGG was used to make canonical pathways as reference mapping of all annotated sequences [57]. KEEG was employed as a reference database of pathway networks for integration and interpretation of large scale datasets generated by high-throughput sequencing technology [58, 59]. On the fact that some unigenes were recruited in several KEGG pathways during the analysis, 26,233 unigenes were assigned to 128 KEGG pathways (S1 Table), of which most represented by Metabolic Pathway (5971 unigenes, 22.76% of annotated to KEGG database), followed by "biosynthesis of secondary metabolites" (2957 unigenes, 11.27% of annotated to KEGG database), "plant-pathogen interactions" (1608 unigenes, 6.13% of annotated to KEGG database), "Plant hormone signal transduction" (1396 unigenes, 5.32%of annotated to KEGG database) and "Ribosome" (1174 unigenes, 4.48% of annotated to KEGG database).

Table 2. Statistics of annotations for assembled unigenes of *Atractylodes lancea* in different public databases.

Database	Unigenes	Percentage(%)
NR	43049	69.04
SWISS-PROT	30264	48.53
KEGG	26233	42.07
COG	17881	28.67
GO	29057	46.6
ALL	44482	71.34

doi:10.1371/journal.pone.0151975.t002

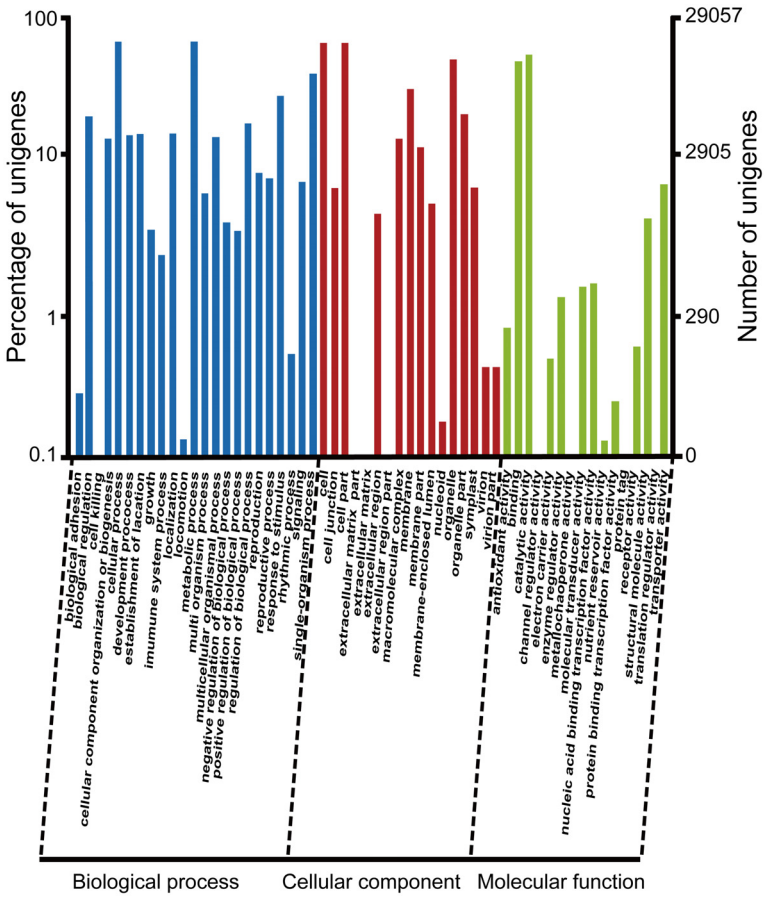

Fig 4. Distributions of GO annotation of all unigenes. The results were classified into three main categories: biological process, cellular component, and molecular function. The left y-axis indicates the percentage of a specific category of genes in that category. The right y-axis indicates the number of genes in a category.

doi:10.1371/journal.pone.0151975.g004

Differentially expressed genes (DEGs) in the leaf vs. stem, leaf vs. root and root vs. stem in *A. lancea*

A major function of the transcriptome sequencing is for comparison of levels of gene expression among different samples. To check the differences in expression of gene among three libraries from the leaf, stem and root, the tag frequencies of leaf vs. stem, leaf vs. root and root vs. stem were used. Through FPKM method (fragments per kb per million reads) all-unigene expressions were calculated. Firstly fragments density measures was normalized and for judgment of significance of gene expression false discovery rate(FDR) < 0.001 were used and the total value of $|\log2\text{Ratio}| \geq 1$ was used as a threshold. In Fig 6 the result shows a two-fold transcript difference among three libraries. We identified 22543, 18263 and 16370 unigenes in leaf vs. stem, leaf vs. root and root vs. stem libraries respectively that were differently expressed in all three libraries (S2 Table). Of these 11642, 8668 and 9038 unigenes were up-regulated and 10901, 9605 and 7232 unigenes in three libraries were down-regulated regulated by the log2 ratio bigger than 2 or less than 0.5 of leaf vs. stem, leaf vs. root and root vs. stem, respectively. It also showed among these differential expression genes, most were found expressed in the root, and then the stem and leaf. One assumption is that the diverse chemical synthesis of the plant is largely processed in the root.

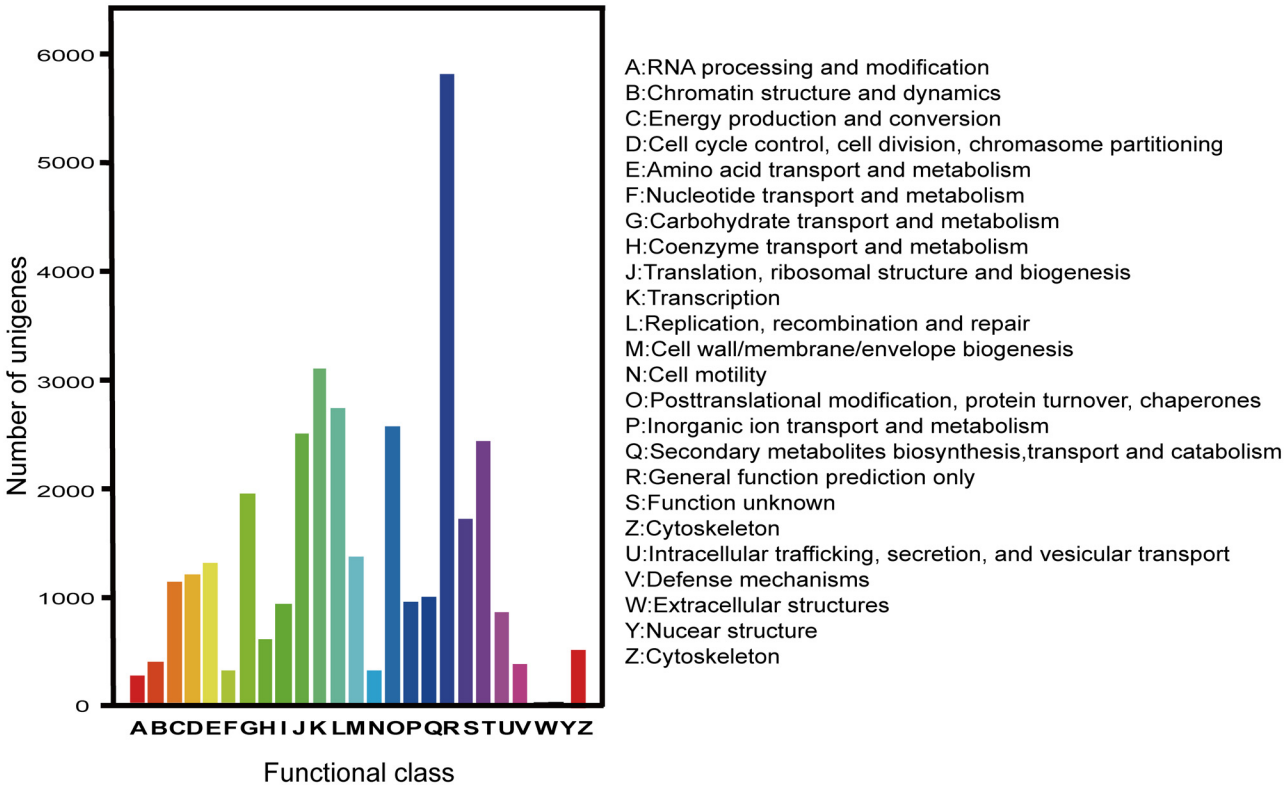

Fig 5. COG function classification of all unigenes. The annotated unigenes are divided into a variety of functional orthologous groups, which are indicated by letters A-Z and annotated besides the figure.

doi:10.1371/journal.pone.0151975.g005

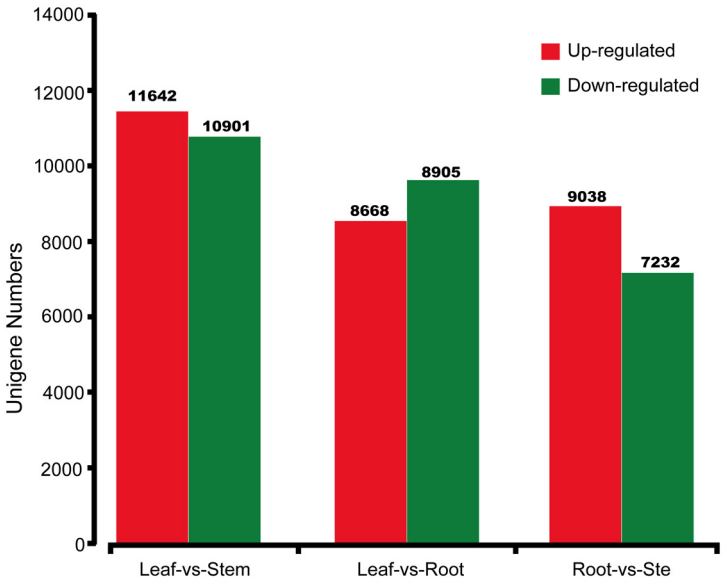

Fig 6. Differentially expressed genes profiling of three libraries of leaf, root and stem of *Atractylodes lancea*. The red and green columns indicate up- and down-regulated genes in comparisons of leaves, stem and root libraries in *A. lancea*. FDR$\leq$0.05 and the absolute value of Log2FC Ratio $\geq$1 were used as the threshold to judge the significance of gene expression difference from transcriptome data.

doi:10.1371/journal.pone.0151975.g006

Analysis of *A. lancea* unigenes related to terpenoid backbonebiosynthesis

Based on the Nr annotation, a total of 77 Contigs/unigenes were identified as the genes of MVA pathway, that include acetyl CoA C-acetyltransferase (AACT), 3-hydroxy-3-methylglutaryl CoA synthase (HMGS), 3-hydroxy-3-methylglutaryl CoA reductase (HMGR), mevalonate kinase (MK), phosphomevalonate kinase (PMK), mevalonate-5-pyrophosphate decarboxylase (MDC), isopentenyl diphosphate isomerase (IPPI), geranyl diphosphate synthase (GPPS), farnesyl diphosphate synthase (FPPS), beta-caryophyllene synthase (QHS1), germacrene D synthase (GDS), germacrene A synthase (GAS) and E-β-farnesene synthase (β-FS). These genes produce β-caryophyllene, germacrene D, germacrene A and E-β-farnesene four different types of sesquiterpenoids [21, 60–62]. It also has to be pointed out that due to the limitation of short reads of RNA-seq, some unigenes assembled by the software are too short to represent real transcripts. Other unigenes are long enough to cover one or two domains of usual protein size, but they are almost identical to a longer transcript except a small part of the fragments. These unigenes are likely from one gene, possibly generated with selective transcripts or assembly error. In the end, we predicted 33 unigenes that are responsible for the enzymatic synthesis of MVA pathway (Fig 1). Nevertheless, these unigenes need to be approved by future cloning. Based on our analysis, up to 10 non-redundant unigenes were present in the plastidal MEP pathway, liable for the synthesis of the isopentenyl diphosphate that is the building block of terpenoids. These included 2 unigenes for 1-deoxy-D-xylulose-5-phosphate synthase (DXPS), two unigenes for 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR), 1 unigene for 2-C-methyl-D-erythritol 4-phosphate cytidylyl transferase (MCT), 2 unigenes for 4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase (CMK), 1 unigene for 2-C-methyl-D-erythritol-2, 4-cyclodiphosphate synthase (MECPS), 1 unigenes for 4-hydroxy-3-methyl but-2-(E)-enyl diphosphate (HDS), and 1 unigene for 4-hydroxy-3-methyl but-2-(E)-enyl diphosphate reductase (HDR). It is shown that most of candidate genes from the MEP pathway were up-regulated in leaves except DXPS and DXR. One DXPS (CL8765.Contig3_All) is up-regulated in the root and one (Unigene15742_All) is down-regulated in the stem. While in case of HDR, out of two unigene, (CL5530.Contig1_All) is up-regulated in the root (S1 Table). We also found that the genes from MEP pathway showed higher expression in leaves than in root and stem at the transcriptional level. Only one unigene was found codifying the isopentenyl diphosphate delta-isomerase which catalyzes the alteration of isopentenyl diphosphate Dimethylallyl diphosphate. Moreover, we found that prenyl-transferases, which generates higher-order building blocks: farnesyl diphosphate synthase (2 contigs/unigene) and geranyl diphosphate synthase (8 unigenes), are the originator of different categories of terpenoids. The protein sequences of all the transcripts are provided in S1 Table.

Analysis of *A. lancea* unigenes related to sesquiterpenoids biosynthesis

Sesquiterpenoids are derived from FPP which can be cyclized to produce various structures by different types of enzymes [63]. In this study, 19 contigs/unigenes (≥200 bp) were annotated to be involved in four different types of sesquiterpenoid biosynthesis, which includes β-caryophyllene synthase, germacrene D synthase, E-β-farnesene synthase and germacrene A synthase (Fig 1). These contigs/unigenes are CL4471.contig1_ALL, CL8689.contig1_ALL, Unigene14966_All, Unigene20711_All, Unigene20756_All, Unigene20757_All Unigene21327_All, Unigene21328_All, Unigene21329_All, Unigene21417_All, Unigene23621_All, Unigene25222_All, Unigene32673_All, Unigene33174_All, Unigene33794_All, Unigene34375_All, CL1332.Contig1_All, CL1748.Contig6_All and CL5528.Contig1_All. We noticed that, the above four sesquiterpenoid synthases share a high homology and it is difficult to separate them from each other without experimental confirmation. Furthermore, these enzymes are also

homologous to sesquiterpene cyclase, β-pinene synthase, α-isocomene synthase, etc. All these enzymes are commonly grouped as sesquiterpene synthases. However, only one potential unigene (CL5528.Contig1_All) among the 19 sesquiterpene synthases is predicted to be germacrene A synthase with certainty. In another case, β-eudesmol synthase was reported for the specific sesquiterpene β-eudesmol biosynthesis [64]. But we were unable to find any candidate could match the enzyme with the current searching criteria. Previous study showed that βeudesmol was not always present in *A*. *lancea* samples [23, 24], In our study β-eudesmol synthase was not detected either because the gene expression was too low to be captured or because of the poor fragmentation or enrichment during the process of RNA-Seq.

It is very intriguing to pinpoint all enzymes especially for *A*. *lancea* sesquiterpenoid biosynthesis. Previous study indicates that cytochrome P450 oxidase (CYP) plays an important role in generation of all kinds of terpenoid derivatives [65]. In a search of all possible CYPs encoded by the *A*. *lancea* transcriptome, a total of 3,241 CYP contigs were found in the Nr annotation. Further filtration of redundant contigs with possibly the same predicted functions, the CYP quantity was narrowed down to 369 (S1 Table), however, it is still 1.5 times more than that of Arabidopsis thaliana [65], indicating a more sophisticated chemical process and diversity. Besides the sesquiterpene, there are other kinds of terpenes with less content in the *A*. *Lancea*. It correlates with the discovery of a large number of CYP genes, of which some are predicted to be terpene modifiers. All these components lay the foundation of chemical diversity for the fact that it treats various diseases.

The study on the biochemical properties of enzymes for sesiquiterpenes biosynthesis has made substantial progress in the past years, such as discovery of committed enzymatic steps in the biosynthesis of sesiquiterpenes [66, 67]. However, the identification and cloning of these enzymes are more challenging. Other than β-caryophyllene synthase, germacrene D synthase, E-β-farnesene synthase and germacrene A synthase, there are a few similar genes have been elucidated like tomato sesquiterpene synthase (Sst1) and Sst2 [68], aoeghum terpene synthase (SbTPS1-SbTPS7) [69], and a Cstps1, a sesquiterpene synthase-encoding genes for citurs aroma formation [70]. In our annotation database, we could sort out a bunch of sesquiterpene synthases. But due to the structural similarity between the sesquiterpenes, the sesquiterpene synthases also come with a close homology. Future study may explain whether those sesquiterpene synthase candidates can be grouped into further subgroups of each with the specificity to one kind of sesquiterpene.

SSR markers development in *Atractylodes lancea*

SSRs are used as chief molecular markers. These repetitive DNA sequences symbolize a vital section of an advanced eukaryote genome. These typically co ascendant and highly polymorphic are widely utilized for marker systems of genetic mapping, molecular breeding in a wide variety of species [71–76]. In order to develop SSR markers in *A*. *lancea* and find potential microsatellites, all the 62,352 unigenes produced in current study were utilized, for all motifs they were defined as bi-hexa nucleotide SSR with at least four repeating units (except for di-nucleotide with a minimum of six repeating units, and tri-nucleotide with a minimum of five repeats). By using different primers (S1 Table), total of 6,864 microsatellites were identified in 5,970 unigenes, 757 unigenes contained more than 1 SSR. Di-nucleotide motifs were found to be the most abundant types (3,122, 45.48%), followed by tri-nucleotide (2,307, 33.61%), hexa-nucleotide (432, 6.29%), penta-nucleotide (303, 4.41%) and tetra-nucleotide (130, 1.89%), (Table 3). In our current study, *AG/CT* repeat was found the most abundant motif among all the searched SSRs, (2252, 32.80%), followed by *AC/GT* (523, 7.61%), *ACC/GGT* (505, 7.35%), and *AAG/CTT* (484, 7.05%) (Fig 7). Conventional methods for SSR marker development are

Table 3. A summary of SSRs identified in *Atractylodes lancea*.

Searching Items	Numbers
Total number of sequence examined	62352
Total size of examined sequence	56,9232,90
Total number of identified cSSRs	6864
Number of cSSRs containing sequences	5970
Number of sequences containing more than one cSSRs	757
Number of cSSRs present in compound formation	303
Mono-nucleotides	570
Di-nucleotides	3122
Tri-nucleotides	2307
Tetra-nucleotides	130
Penta-nucleotides	303
Hexa-nucleotides	432

doi:10.1371/journal.pone.0151975.t003

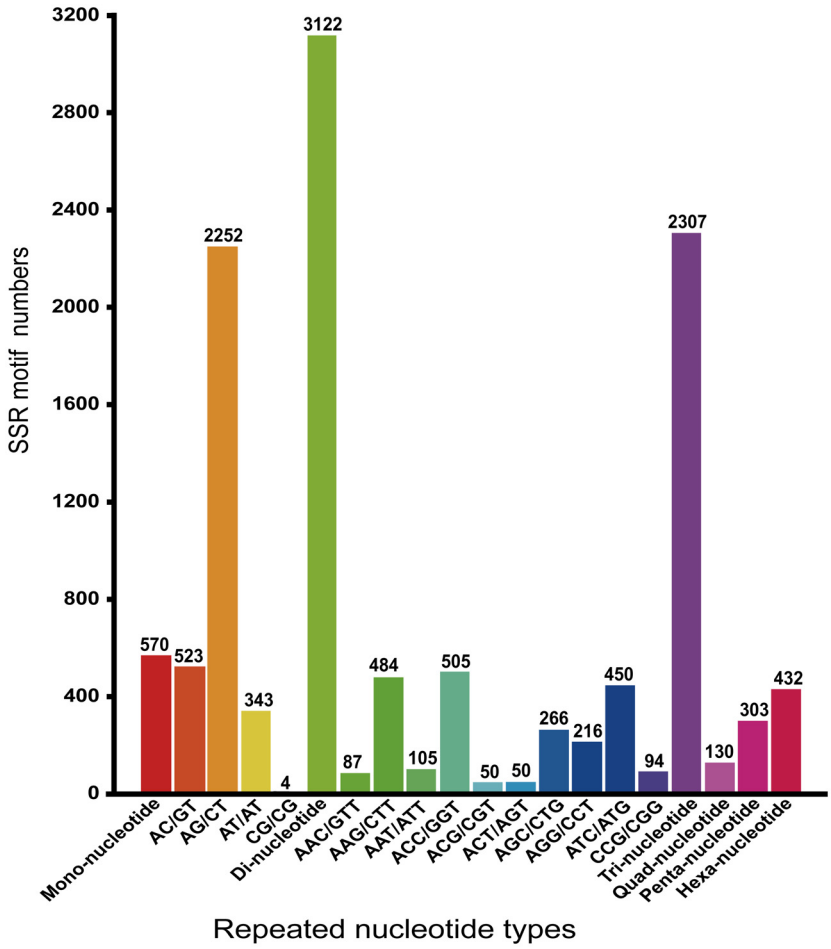

Fig 7. Quantity statistics of SSR classification: The X-axis is the repeat times of repeat units; the Y-axis is the number of SSRs from *Atractylodes lancea*. The di-nucleotide category was found in large number and among the di-nucleotide (AG/CT) was most abundant one in our SSRs.

doi:10.1371/journal.pone.0151975.g007

Table 4. A summary of SNP results in *Atractylodes lancea*.

SNP Type	Leave	Root	Stem	Total
Transition	57839	51702	55579	1,65,120
AG	29135	26042	28015	83192
CT	28704	25660	27564	81928
Transversions	33701	29976	31750	95427
AC	8428	7402	7902	23732
AT	9601	8564	8989	27154
GC	7281	6629	6952	20862
GT	8391	7381	7907	23679
Total	91540	81678	87329	2,60,547

doi:10.1371/journal.pone.0151975.t004

expensive, arduous and time-consuming. The newly discovered and developed high-throughput sequencing technique is a powerful and cheap tool for transcriptome sequencing [77]. For microsatellite mining, SSR markers are being developed by the transcriptome data, and had been utilized in many species [55, 78, 79].

Discovery of simple nucleotide polymorphisms (SNPs)

Innovating SNPs from cDNA libraries mapping revealed 91,540; 81,678 and 87,329 SNPs across 112,883; 94,663 and 101,679 contigs in leaves, root and stem of *A. lancea*, respectively (Table 4). A total of 260,547 heterozygous SNPs were detected from all three samples and out of these 165,120 were transitions and 95,427 were transversions (Fig 8). We also found several prospective SNP markers, which can be beneficial for the phylogenetic and population genetic studies of *A. lancea*. The identified SNP markers can be constructive to assist in genetic mark of selection for genetic association analysis in further research and also for identification of

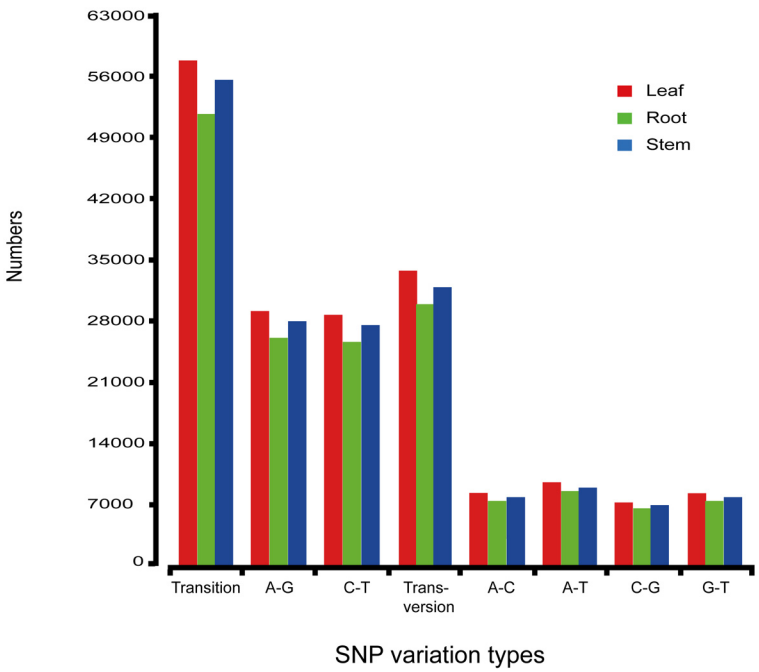

Fig 8. Statistics of SNP number. The X-axis is SNP types; the Y-axis is the number of SNP.

doi:10.1371/journal.pone.0151975.g008

functional variations [80, 81]. The identification of huge SNPs provides affluence of potential markers to be helpful in various applications, such as linkage mapping, population genetics, and gene-based association studies and comparative genomics.

Conclusion

People in South-East Asian countries like China, Japan and Thailand make use of *A*. *lancea* as a traditional medicine for different diseases for a long time. Here we report the Illumina transcriptome sequencing, functional annotation and differential expression profiles in the different tissues i.e. stem root, and leaf of *A*. *lancea* which will be an important resource for gene mining, genetic improvement and development of different molecular markers. In current study 62,352 high quality unigenes were obtained from these tissues. Additionally we found the unigenes that are responsible for encoding the different enzymes that are involved it the biosynthesis of terpenoid backbone pathway as well as sesquiterpenoids which will help future functional & comparative genomic research on this important plant.

Supporting Information

S1 Table. Unigenes from the transcriptome of *Atractylodes lancea* by KEGG. 26,233 unigenes were assigned to 128 KEGG pathways.
(XLSX)

S2 Table. Differentially expressed unigenes from three libraries of *Atractylodes lancea* showing up and down regulated unigenes (leaf vs stem), (leaf vs root), (root vs stem).
(XLSX)

S3 Table. Putative unigenes along their enzymes of the terpenoid backbone biosynthesis as well as four different types of sesquiterpenoids with their FPKM value showing their expression in three different tissues of *Atractylodes lancea*.
(XLSX)

S4 Table. Protein sequences of all the transcripts involved in the sesquiterpenoid biosynthesis.
(TXT)

S5 Table. All contigs and unigenes for CYP450 in *Atractylodes lancea*.
(XLSX)

S6 Table. Primers used for SSR analysis in *Atractylodes lancea*.
(XLS)

Author Contributions

Conceived and designed the experiments: XH SA CZ XW. Performed the experiments: SA CZ YY XW TY ZZ QZ. Analyzed the data: SA CZ XL XH. Contributed reagents/materials/analysis tools: SA CZ YY XW TY ZZ QZ. Wrote the paper: SA CZ XL XH.

References

1. Na-Bangchang K, Karbwang J. Traditional Herbal Medicine for the Control of Tropical Diseases. Tropical medicine and health. 2014; 42(2 Suppl):3. doi: 10.2149/tmh.2014-S01 PMID: 25425945
2. Yu Y, Jia T-Z, Cai Q, Jiang N, Ma M-y, Min D-y, et al. Comparison of the anti-ulcer activity between the crude and bran-processed Atractylodes lancea in the rat model of gastric ulcer induced by acetic acid. Journal of ethnopharmacology. 2015; 160:211–8. doi: 10.1016/j.jep.2014.10.066 PMID: 25481080

3. Duan JA, Wang L, Qian S, Su S, Tang Y. A new cytotoxic prenylated dihydrobenzofuran derivative and other chemical constituents from the rhizomes of Atractylodes lancea DC. Arch Pharm Res. 2008; 31 (8):965–9. Epub 2008/09/13. doi: 10.1007/s12272-001-1252-z PMID: 18787781.
4. Nakai Y, Kido T, Hashimoto K, Kase Y, Sakakibara I, Higuchi M, et al. Effect of the rhizomes of Atractylodes lancea and its constituents on the delay of gastric emptying. Journal of ethnopharmacology. 2003; 84(1):51–5. PMID: 12499077
5. Resch M, Heilmann J, Steigel A, Bauer R. Further phenols and polyacetylenes from the rhizomes of Atractylodes lancea and their anti-inflammatory activity. Planta medica. 2001; 67(5):437–42. PMID: 11488458
6. Herb ECoNCMMBC. Editorial Committee of National Chinese Medical Manage Bureau Chinese Herb. Shanghai: Shanghai Science and Technology Publisher; 1999.
7. Wang H-X, Liu C-M, Liu Q, Gao K. Three types of sesquiterpenes from rhizomes of Atractylodes lancea. Phytochemistry. 2008; 69(10):2088–94. doi: 10.1016/j.phytochem.2008.04.008 PMID: 18511090
8. Kitajima J, Kamoshita A, Ishikawa T, Takano A, Fukuda T, Isoda S, et al. Glycosides of Atractylodes lancea. Chem Pharm Bull (Tokyo). 2003; 51(6):673–8. Epub 2003/06/17. PMID: 12808245.
9. Lange BM, Ahkami A. Metabolic engineering of plant monoterpenes, sesquiterpenes and diterpenes—current status and future opportunities. Plant biotechnology journal. 2013; 11(2):169–96. doi: 10.1111/pbi.12022 PMID: 23171352
10. Parshikov IA, Sutherland JB. The use of Aspergillus niger cultures for biotransformation of terpenoids. Process Biochemistry. 2014; 49(12):2086–100. doi: 10.1016/j.procbio.2014.09.005 WOS:000347761100010.
11. Zhao CL, Cui XM, Chen YP, Liang QA. Key Enzymes of Triterpenoid Saponin Biosynthesis and the Induction of Their Activities and Gene Expressions in Plants. Natural Product Communications. 2010; 5 (7):1147–58. WOS:000280117300036. PMID: 20734961
12. Cane DE. Sesquiterpene biosynthesis: cyclization mechanisms. Comprehensive natural products chemistry. 1999; 2:155–200.
13. Berger RG. Flavours and fragrances: chemistry, bioprocessing and sustainability: Springer Science & Business Media; 2007.
14. Simonsen HT, Weitzel C, Christensen SB. Guaianolide sesquiterpenoids: Pharmacology and biosynthesis. Natural Products: Springer; 2013. p. 3069–98.
15. Zwenger S, Basu C. Plant terpenoids: applications and future potentials. Biotechnol Mol Biol Rev. 2008; 3(1):1–7.
16. Di Sotto A, Evandri MG, Mazzanti G. Antimutagenic and mutagenic activities of some terpenes in the bacterial reverse mutation assay. Mutation Research/Genetic Toxicology and Environmental Mutagenesis. 2008; 653(1):130–3.
17. Legault J, Pichette A. Potentiating effect of β-caryophyllene on anticancer activity of α-humulene, isocaryophyllene and paclitaxel. Journal of Pharmacy and Pharmacology. 2007; 59(12):1643–7. PMID: 18053325
18. Cornwell P, Barry B. Sesquiterpene components of volatile oils as skin penetration enhancers for the hydrophilic permeant 5-fluorouracil. Journal of pharmacy and pharmacology. 1994; 46(4):261–9. PMID: 8051608
19. Prosser I, Altug IG, Phillips AL, König WA, Bouwmeester HJ, Beale MH. Enantiospecific (+)-and (−)-germacrene D synthases, cloned from goldenrod, reveal a functionally active variant of the universal isoprenoid-biosynthesis aspartate-rich motif. Archives of Biochemistry and Biophysics. 2004; 432 (2):136–44. PMID: 15542052
20. Picaud S, Olsson ME, Brodelius M, Brodelius PE. Cloning, expression, purification and characterization of recombinant (+)-germacrene D synthase from Zingiber officinale. Arch Biochem Biophys. 2006; 452 (1):17–28. doi: 10.1016/j.abb.2006.06.007 PMID: 16839518.
21. Picaud S, Brodelius M, Brodelius PE. Expression, purification and characterization of recombinant (E)-beta-farnesene synthase from Artemisia annua. Phytochemistry. 2005; 66(9):961–7. doi: 10.1016/j.phytochem.2005.03.027 PMID: 15896363.
22. Crock J, Wildung M, Croteau R. Isolation and bacterial expression of a sesquiterpene synthase cDNA clone from peppermint (Mentha x piperita, L.) that produces the aphid alarm pheromone (E)-β -farnesene. Proceedings of the National Academy of Sciences. 1997; 94(24):12833–8.
23. Jia C, Mao D, Zhang W, Sun X. [Studies on chemical constituents in essential oil from wild Atractylodes lancea in dabie mountains]. Zhong Yao Cai. 2004; 27(8):571–4. Epub 2005/01/22. PMID: 15658816.
24. Liu Y, Chen W, Zeng M, Xu K. [Pharmacodynamics of water extracts from Atractylodes lancea before and after processing]. Zhongguo Zhong Yao Za Zhi. 2012; 37(15):2276–9. Epub 2012/11/30. PMID: 23189733.

25. Wang KT, Chen LG, Wu CH, Chang CC, Wang CC. Gastroprotective activity of atractylenolide III from Atractylodes ovata on ethanol-induced gastric ulcer in vitro and in vivo. J Pharm Pharmacol. 2010; 62 (3):381–8. Epub 2010/05/22. doi: 10.1211/jpp.62.03.0014 JPHP381 [pii]. PMID: 20487223.

26. Chen HP, Zheng LS, Yang K, Lei N, Geng ZF, Ma P, et al. Insecticidal and repellant activities of polyacetylenes and lactones derived from Atractylodes lancea rhizomes. Chem Biodivers. 2015; 12(4):593–8. Epub 2015/04/17. doi: 10.1002/cbdv.201400161 PMID: 25879503.

27. Masuda Y, Kadokura T, Ishii M, Takada K, Kitajima J. Hinesol, a compound isolated from the essential oils of Atractylodes lancea rhizome, inhibits cell growth and induces apoptosis in human leukemia HL-60 cells. J Nat Med. 2015; 69(3):332–9. Epub 2015/04/03. doi: 10.1007/s11418-015-0897-5 PMID: 25833731.

28. Wang HX, Liu CM, Liu Q, Gao K. Three types of sesquiterpenes from rhizomes of Atractylodes lancea. Phytochemistry. 2008; 69(10):2088–94. Epub 2008/05/31. doi: 10.1016/j.phytochem.2008.04.008 S0031-9422(08)00191-X [pii]. PMID: 18511090.

29. Resch M, Heilmann J, Steigel A, Bauer R. Further phenols and polyacetylenes from the rhizomes of Atractylodes lancea and their anti-inflammatory activity. Planta Med. 2001; 67(5):437–42. Epub 2001/08/08. doi: 10.1055/s-2001-15817 PMID: 11488458.

30. Resch M, Steigel A, Chen ZL, Bauer R. 5-Lipoxygenase and cyclooxygenase-1 inhibitory active compounds from Atractylodes lancea. J Nat Prod. 1998; 61(3):347–50. Epub 1998/04/17. doi: 10.1021/np970430b np970430b [pii]. PMID: 9544564.

31. Inagaki N, Komatsu Y, Sasaki H, Kiyohara H, Yamada H, Ishibashi H, et al. Acidic polysaccharides from rhizomes of Atractylodes lancea as protective principle in Candida-Infected mice. Planta Med. 2001; 67(5):428–31. Epub 2001/08/08. doi: 10.1055/s-2001-15822 PMID: 11488456.

32. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics. 2013; 14(6):671–83. doi: 10.1093/bib/bbs046 PMID: 22988256.

33. 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 PMID: 24260395; PubMed Central PMCID: PMCPMC3834075.

34. Li C, Zhu Y, Guo X, Sun C, Luo H, Song J, et al. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in Panax ginseng CA Meyer. BMC genomics. 2013; 14(1):245.

35. 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(1):265. doi: 10.1186/s12864-015-1477-5 PMID: 25886736; PubMed Central PMCID: PMCPMC4399409.

36. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456(7221):470–6. doi: 10.1038/nature07509 PMID: 18978772; PubMed Central PMCID: PMCPMC2593745.

37. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature methods. 2010; 7(9):709–15. doi: 10.1038/nmeth.1491 PMID: 20711195

38. Kuchenbauer F, Morin RD, Argiropoulos B, Petriv OI, Griffith M, Heuser M, et al. In-depth characterization of the microRNA transcriptome in a leukemia progression model. Genome Res. 2008; 18 (11):1787–97. doi: 10.1101/gr.077578.108 PMID: 18849523; PubMed Central PMCID: PMCPMC2577858.

39. Metzker ML. APPLICATIONS OF NEXT-GENERATION SEQUENCING Sequencing technologies—the next generation. Nature Reviews Genetics. 2010; 11(1):31–46.

40. Sun X, Zhou S, Meng F, Liu S. De novo assembly and characterization of the garlic (Allium sativum) bud transcriptome by Illumina sequencing. Plant Cell Rep. 2012; 31(10):1823–8. doi: 10.1007/s00299-012-1295-z PMID: 22684307.

41. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437(7057):376–80. doi: 10.1038/nature03959 PMID: 16056220; PubMed Central PMCID: PMCPMC1464427.

42. Wheat CW. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica. 2010; 138(4):433–51. doi: 10.1007/s10709-008-9326-y PMID: 18931921.

43. 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 PMID: 19015660; PubMed Central PMCID: PMCPMC2949280.

44. Yin D, Wang Y, Zhang X, Li H, Lu X, Zhang J, et al. De novo assembly of the peanut (Arachis hypogaea L.) seed transcriptome revealed candidate unigenes for oil accumulation pathways. PLoS One. 2013; 8 (9):e73767. doi: 10.1371/journal.pone.0073767 PMID: 24040062; PubMed Central PMCID: PMCPMC3769373.

45. 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 PMID: 21572440; PubMed Central PMCID: PMCPMC3571712.

46. Iseli C, Jongeneel CV, Bucher P, editors. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. ISMB; 1999.

47. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theoretical and Applied Genetics. 2003; 106(3):411–22. PMID: 12589540

48. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007; 23(10):1289–91. doi: 10.1093/bioinformatics/btm091 PMID: 17379693.

49. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. Nucleic Acids Research. 2012; 40(15):e115–e. ARTN e115 doi: 10.1093/nar/gks596 WOS:000308958600003. PMID: 22730293

50. Liu Y, Huang Z, Ao Y, Li W, Zhang Z. Transcriptome analysis of yellow horn (Xanthoceras sorbifolia Bunge): a potential oil-rich seed tree for biodiesel in China. PLoS One. 2013; 8(9):e74441. doi: 10.1371/journal.pone.0074441 PMID: 24040247; PubMed Central PMCID: PMCPMC3770547.

51. Dong S, Liu Y, Niu J, Ning Y, Lin S, Zhang Z. De novo transcriptome analysis of the Siberian apricot (Prunus sibirica L.) and search for potential SSR markers by 454 pyrosequencing. Gene. 2014; 544 (2):220–7. doi: 10.1016/j.gene.2014.04.031 PMID: 24746601.

52. Sangwan RS, Tripathi S, Singh J, Narnoliya LK, Sangwan NS. De novo sequencing and assembly of Centella asiatica leaf transcriptome for mapping of structural, functional and regulatory genes with special reference to secondary metabolism. Gene. 2013; 525(1):58–76. doi: 10.1016/j.gene.2013.04.057 PMID: 23644021.

53. Yang Y, Xu M, Luo Q, Wang J, Li H. De novo transcriptome analysis of Liriodendron chinense petals and leaves by Illumina sequencing. Gene. 2014; 534(2):155–62. doi: 10.1016/j.gene.2013.10.073 PMID: 24239772.

54. Consortium GO. The Gene Ontology (GO) database and informatics resource. Nucleic acids research. 2004; 32(suppl 1):D258–D61.

55. Chou KC. Some remarks on predicting multi-label attributes in molecular biosystems. Molecular Biosystems. 2013; 9(6):1092–100. doi: 10.1039/c3mb25555g WOS:000318557100005. PMID: 23536215

56. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003; 4(1):41. doi: 10.1186/1471-2105-4-41 PMID: 12969510; PubMed Central PMCID: PMCPMC222959.

57. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28 (1):27–30. PMID: 10592173; PubMed Central PMCID: PMCPMC102409.

58. Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC genomics. 2010; 11(1):400.

59. Xie F, Burklew CE, Yang Y, Liu M, Xiao P, Zhang B, et al. De novo sequencing and a comprehensive analysis of purple sweet potato (Impomoea batatas L.) transcriptome. Planta. 2012; 236(1):101–13. doi: 10.1007/s00425-012-1591-4 PMID: 22270559

60. Prosser I, Phillips AL, Gittings S, Lewis MJ, Hooper AM, Pickett JA, et al. (+)-(10R)-Germacrene A synthase from goldenrod, Solidago canadensis; cDNA isolation, bacterial expression and functional analysis. Phytochemistry. 2002; 60(7):691–702. PMID: 12127586.

61. Bennett MH, Mansfield JW, Lewis MJ, Beale MH. Cloning and expression of sesquiterpene synthase genes from lettuce (Lactuca sativa L.). Phytochemistry. 2002; 60(3):255–61. PMID: 12031443.

62. Irmisch S, Krause ST, Kunert G, Gershenzon J, Degenhardt J, Kollner TG. The organ-specific expression of terpene synthase genes contributes to the terpene hydrocarbon composition of chamomile essential oils. BMC Plant Biol. 2012; 12(1):84. doi: 10.1186/1471-2229-12-84 PMID: 22682202; PubMed Central PMCID: PMCPMC3423072.

63. Xia JH, Zhang SD, Li YL, Wu L, Zhu ZJ, Yang XW, et al. Sesquiterpenoids and triterpenoids from Abies holophylla and their bioactivities. Phytochemistry. 2012; 74:178–84. doi: 10.1016/j.phytochem.2011.11.011 WOS:000300815700020. PMID: 22169016

64. Yu F, Harada H, Yamasaki K, Okamoto S, Hirase S, Tanaka Y, et al. Isolation and functional characterization of a beta-eudesmol synthase, a new sesquiterpene synthase from Zingiber zerumbet Smith. FEBS Lett. 2008; 582(5):565–72. Epub 2008/02/05. S0014-5793(08)00045-8 [pii]. PMID: 18242187.

65. Mizutani M, Ohta D. Diversification of P450 genes during land plant evolution. Annu Rev Plant Biol. 2010; 61:291–315. Epub 2010/03/03. doi: 10.1146/annurev-arplant-042809-112305 PMID: 20192745.

66. de Kraker JW, Franssen MC, Dalm MC, de Groot A, Bouwmeester HJ. Biosynthesis of germacrene A carboxylic acid in chicory roots. Demonstration of a cytochrome P450 (+)-germacrene a hydroxylase and NADP+-dependent sesquiterpenoid dehydrogenase(s) involved in sesquiterpene lactone biosynthesis. Plant Physiol. 2001; 125(4):1930–40. Epub 2001/04/12. PMID: 11299372; PubMed Central PMCID: PMC88848.

67. de Kraker JW, Franssen MC, de Groot A, Konig WA, Bouwmeester HJ. (+)-Germacrene A biosynthesis. The committed step in the biosynthesis of bitter sesquiterpene lactones in chicory. Plant Physiol. 1998; 117(4):1381–92. Epub 1998/08/14. PMID: 9701594; PubMed Central PMCID: PMC34902.

68. van Der Hoeven RS, Monforte AJ, Breeden D, Tanksley SD, Steffens JC. Genetic control and evolution of sesquiterpene biosynthesis in Lycopersicon esculentum and L. hirsutum. Plant Cell. 2000; 12 (11):2283–94. Epub 2000/11/23. PMID: 11090225; PubMed Central PMCID: PMC150174.

69. Zhuang X, Kollner TG, Zhao N, Li G, Jiang Y, Zhu L, et al. Dynamic evolution of herbivore-induced sesquiterpene biosynthesis in sorghum and related grass crops. Plant J. 2012; 69(1):70–80. Epub 2011/09/02. doi: 10.1111/j.1365-313X.2011.04771.x PMID: 21880075.

70. Sharon-Asa L, Shalit M, Frydman A, Bar E, Holland D, Or E, et al. Citrus fruit flavor and aroma biosynthesis: isolation, functional characterization, and developmental regulation of Cstps1, a key gene in the production of the sesquiterpene aroma compound valencene. Plant J. 2003; 36(5):664–74. Epub 2003/11/18. 1910 [pii]. PMID: 14617067.

71. Chapman MA, Hvala J, Strever J, Matvienko M, Kozik A, Michelmore RW, et al. Development, polymorphism, and cross-taxon utility of EST-SSR markers from safflower (Carthamus tinctorius L.). Theor Appl Genet. 2009; 120(1):85–91. doi: 10.1007/s00122-009-1161-8 PMID: 19820913.

72. Smith J, Chin E, Shu H, Smith O, Wall S, Senior M, et al. An evaluation of the utility of SSR loci as molecular markers in maize (Zea mays L.): comparisons with data from RFLPs and pedigree. Theoretical and Applied Genetics. 1997; 95(1–2):163–73.

73. Wang Y, Georgi LL, Zhebentyayeva TN, Reighard GL, Scorza R, Abbott AG. High-throughput targeted SSR marker development in peach (Prunus persica). Genome. 2002; 45(2):319–28. PMID: 11962629.

74. Neeraja CN, Maghirang-Rodriguez R, Pamplona A, Heuer S, Collard BC, Septiningsih EM, et al. A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. Theor Appl Genet. 2007; 115(6):767–76. doi: 10.1007/s00122-007-0607-0 PMID: 17657470.

75. Bushman BS, Larson SR, Tuna M, West MS, Hernandez AG, Vullaganti D, et al. Orchardgrass (Dactylis glomerata L.) EST and SSR marker development, annotation, and transferability. Theor Appl Genet. 2011; 123(1):119–29. doi: 10.1007/s00122-011-1571-2 PMID: 21465186.

76. Senior M, Murphy J, Goodman M, Stuber C. Utility of SSRs for determining genetic similarities an relationships in maize using an agarose gel system. Crop science. 1998; 38(4):1088–98.

77. Verma VK, Behera T, Munshi A, Parida SK, Mohapatra T. Genetic diversity of ash gourd [Benincasa hispida (Thunb.) Cogn.] inbred lines based on RAPD and ISSR markers and their hybrid performance. Scientia horticulturae. 2007; 113(3):231–7.

78. Guo S, Liu J, Zheng Y, Huang M, Zhang H, Gong G, et al. Characterization of transcriptome dynamics during watermelon fruit development: sequencing, assembly, annotation and gene expression profiles. BMC Genomics. 2011; 12(1):454. doi: 10.1186/1471-2164-12-454 PMID: 21936920; PubMed Central PMCID: PMCPMC3197533.

79. Kaur S, Cogan NO, Pembleton LW, Shinozuka M, Savin KW, Materne M, et al. Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genomics. 2011; 12(1):265. doi: 10.1186/1471-2164-12-265 PMID: 21609489; PubMed Central PMCID: PMCPMC3113791.

80. Zhao W, Park EJ, Chung JW, Park YJ, Chung IM, Ahn JK, et al. Association analysis of the amino acid contents in rice. J Integr Plant Biol. 2009; 51(12):1126–37. doi: 10.1111/j.1744-7909.2009.00883.x PMID: 20021560.

81. Chung J, Cho Y, Lee J, Lee S, Ma K, Lee K, et al. Characters of RVA from grain mutants of a rice variety Shindongjin. The Journal of the Korean Society of International Agriculture. 2009.