Transcriptome Analyses to Reveal Genes Involved in Terpene Biosynthesis in Resin Producing Pine Tree *Pinus kesiya* var. *langbianensis*

Liukun Wen,[a] Rui Shi,[a,*] Juan Wang,[b] Youjie Zhao,[a] Huaibi Zhang,[c] Xi Ling,[a] and Zhi Xiong [a,*]

Pinus kesiya var. *langbianensis* is an important resin resource tree that belongs to the Pinaceae family. It produces a higher yield of resin per year compared to the rest of the pine trees from the same habitat. To identify genes that may be involved in this high resin yield production, the bark transcriptomes of *P. kesiya* var. *langbianensis* and a *P. kesiya* that produce a normal volume of resin were sequenced using RNA-Seq, and their gene expression profiles were compared in regards to specific interest in the resin synthetic metabolism pathways. The results showed that a total of 68,881 transcripts were assembled, 180 of which were involved in terpene metabolism. Surprisingly, in both the transcriptome analysis and the quantitative fluorescent polymerase chain reaction (QF-PCR), nine genes involved in resin biosynthesis were shown to be significantly down-regulated in *P. kesiya*. In addition, this study provided numerous gene candidates for the further study of resin production in pine trees.

Keywords: Pinus kesiya var. langbianensis; Resin biosynthetic pathway; Transcriptome sequencing; Gene expression

*Contact information: a: Light Industry and Food Engineering College of Southwest Forestry University, Kunming Yunnan, China, 650224; b: Yunnan Academy of Forestry, Kunming Yunnan, China, 650021; c: New Zealand Institute for Plant & Food Research Ltd., Private Bag 11600, Palmerston North, New Zealand, 4442; *Corresponding author: renee27@foxmail.com; Zhix@swfu.edu.cn*

INTRODUCTION

Pinus kesiya var. *langbianensis* is an economic evergreen gymnosperm tree belonging to the Pinaceae family (Jiang *et al*. 2006). The tree is widely distributed in the southwest region of the Yunnan province of China and is recognized as one of the varieties of *P. kesiya* (Liu *et al*. 2003). It is a quickly growing species whose commercial value is in the wood and resin industries. Pine resin is a semisolid inflammable organic substance obtained from traumatic resin canals, where it is deployed as a self-defensive measure. The resin is the basic component and has many properties, such as wettability, stickiness, and thermal and chemical resistance, for which reasons it is widely used in the adhesive, painting, printing, and rubber industries (Gen *et al*. 2015). Higher resin yields are desirable due to their commercial value in the chemical industries. It has been reported that from 2013 to 2016, the global value of the natural resin market was about $11,300 million, and is estimated to be $14,400 million by 2020 (Nie 2017). The significant demands of the global natural resin market have created work in pine tree selective breeding (Cai *et al*. 2017), where the resin yield is considered one of the most important influences on pine tree selection.

The metabolic pathway of terpenes may be divided into two routes. Pine resin biosynthesis is thought to be closely related to the methyl-erythritol 4-phosphate (MEP) pathway (Wang *et al.* 2015a). Chemical studies on *P. kesiya* var. *langbianensis* have been conducted, and terpenes have been identified as the major component in resin (Wang *et al.* 2012). When the volatile compounds of resin from *P. kesiya* were examined, the main constituents of resin discovered were α-pinene, β-pinene, β-phellandrene, 3-camphene, myrcene, longifolene, and caryophyllene (Zhao *et al*. 2016). Because they possessed similar pentanol isoprene units C5H8, these chemical compounds were determined to be monoterpenes, which were synthesized by terpene synthases (TPS) (Hall *et al*. 2013). Previous studies have shown that TPS genes encode monoterpene synthases *via* the MEP pathway.

With the development of molecular biology, the genetic identity and functional genomics have been able to be exploited for molecular assistance in breeding work (Burdon and Wilcox 2007). Next-generation sequencing (NGS) has been widely used in the study of gene expression, gene regulation, and gene networks in both model and non-model organisms. A large volume of sequences acquired in this way has already been deposited in the GenBank. Recently, some NGS transcriptome data of the *Pinus* species was published (Niu *et al*. 2013). However, when the genome sequence was tested on the loblolly pine (*Pinus taeda*) in 2014, the pine tree yielded the longest genome to match, 23 billion base pairs (Zimin *et al*. 2014), making it difficult to estimate the molecular markers associated with the target quantitative trait locus for resin production.

The present study aims to reveal the key genes involved in the terpene biosynthetic pathway, which could indicate a new molecular-assisted selection process for the breeding of high-resin yielding pine trees. Firstly, the RNA-seq experiments of high- and low-resin yield samples of *P. kesiya* var. *langbianensis* were conducted using the next-generation sequencing technology. The combination of transcriptome and metabolism analysis provided a precise genetic tool to trace the functional genes, and the differentially expressed genes were identified using enrichment analyses. To better understand and validate the transcriptome data, quantitative fluorescent PCR (QF-PCR) experiments were conducted to examine the key gene expression patterns in the terpene pathway. The results represent a means to accelerate the understanding of the molecular mechanism of pine resin production and allow for better application of molecular breeding in the future.

EXPERIMENTAL

Materials

Plant material and RNA extraction

Bark samples from high-resin yield pine trees (resin-producing capacity > 2 g/10 cm) and low-resin yield pine trees (resin-producing capacity < 1 g/10 cm) were collected from a *P. kesiya* var. *langbianensis* forest farm, which was located in Jinggu County in Pu'er City in the Yunnan province of China. The samples were immediately frozen in liquid nitrogen and stored at -80 °C until RNA extraction. The total RNA of each sample was isolated using RNAiso Plus (TaKaRa). The RNA quality was characterized initially on an agarose gel and NanoDrop ND1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and then further assessed by their RNA integrity number (RIN) value (8.0) using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

cDNA library construction and sequencing

The cDNA was first synthesized with a SuperScript cDNA Synthesis kit (Invitrogen, Camarillo, CA, USA). Then, cDNA library construction was built using an Illumina HiSeqTM 2000 device that was based on a GAII platform of Beijing Genomics Institute (Shenzhen, China).

Data filtering and de novo assembly

Image data output from the sequencing device was transformed into raw reads and stored in the FASTQ format. These data were filtered to remove raw reads that included adapter sequences or that were of low quality. The assembly of the transcriptome was achieved using the short-read assembly program Trinity (release-20130225) (Grabherr *et al*. 2011). The unigenes were divided into either clusters or singletons.

Methods

Gene annotation and analysis

Functional annotation provided the information on biological function to reveal the metabolic pathway in the organism. The basic local alignment search tool (BLASTX) (Cameron *et al.* 2005) alignment (applying an E-value of less than 10^{-5}) between each unigene sequence and those lodged in non-redundant protein database (NR; NCBI), non-redundant nucleotide database (NT; NCBI), Swiss-Prot, gene ontology (GO), clusters of orthologous groups (COG), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases were performed, and the best alignments were used to infer the unigene's directionality. Where the outcomes from the various databases conflicted with one another, priority was given in the following order: NR, Swiss-Prot, and COG. Where no alignment was possible, the software tool ESTScan (v2.0) (Iseli *et al*. 1999) was used to assign directionality.

Functional annotation was assigned using the protein (NR and Swiss-Prot), GO, COG, and KEGG databases. The BLASTX was employed to identify related sequences in the protein databases based on an E-value of less than 10^{-5}. The COG database was an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept (Tatusov *et al*. 2000). The aim of GO analysis was to standardize the representation of genes and their derivatives by insisting on a controlled vocabulary and a strictly defined concept (Wand *et al*. 2010; Xia *et al*. 2011). The annotations acquired from NR and Swiss-Prot were processed through the Blast2 GO program (v2.5.0) (Conesa *et al*. 2005) to obtain the relevant GO terms, and these were then analyzed using the WEGO software (v2.0) (Ye *et al*. 2006) to assign a GO functional classification and to illustrate the distribution of gene functions. The KEGG pathway annotations were performed using Blastall software (v2.2.26+x64-linux) against the KEGG databases.

Differential expression analysis

To estimate the expression of each transcript in different tissue samples, the Fragments Per Kilobase Million fragments (FPKM) method was used to calculate the amount of gene expression and according to the amount of gene expression (FPKM value) calculated differentially and expressed in multiple different samples. The differential gene expression analysis of two assigned libraries was performed using the edgeR package (Audic and Claverie 1997). Then, through the digital gene expression spectrum difference gene detection method, the p-value of the difference test was calibrated for multiple hypotheses, and the threshold value of the p-value was determined by means of the false

discovery rate (FDR). The smaller the FDR and the larger the ratio meant a larger difference of the expression level between the two samples. Those with FDR ≤ 0.001 and ratio > 2.0 were chosen. On the basis of the differential expression analysis, the differentially expressed genes were also annotated using GO and KEGG. Finally, the key differentially expressed genes of the resin biosynthesis pathway were found by means of GO and KEGG pathway annotation.

QF-PCR analysis

The cDNA were extracted from *P. kesiya* var. *langbianensis* samples, which were used as the template DNA. Quantitative real-time polymerase chain reaction (PCR) was tested with StepOne PCR (ABI, USA). Actin (Wang *et al*. 2015b) was selected as the reference gene. Gene-specific primers for the nine selected genes (DXS, DXR, MCT, CMK, MECPS, HDR, GPPS, FPPS, and GGPPS) in the resin biosynthesis pathway were designed using the software Primer 5.0 (Premier, Canada), The optional parameters for the 20 μL PCR reaction system were as below: SybrGreen qPCR Master Mix × 10, upstream and downstream primer (10 μM/L) 0.4 μL, template (cDNA) 2 μL, and ddH20 7.2 μL. The reaction system was used: denaturation program (95 °C for 3 min), amplification and quantification program repeated 45 times (95 °C for 7 s, 57 °C for 10 s, and 72 °C for 15 s with a single fluorescence measurement), melting curve program (60 °C to 95 °C with a heating rate of 0.1 °C /s and a continuous fluorescence measurement), and finally a cooling step to 40 °C.

The dissolution curve was analyzed after the amplification. The samples were repeated with 3 biological repetitions and 3 technical repetitions, and the standard curve was used to test the amplification efficiency of each primer, which began with five concentration mixed samples as the template. Then, the real-time fluorescent qPCR amplification was performed and the standard curve was analyzed using the CFX96TM manager software (v3.1) to calculate the primer amplification efficiency and coefficient R^2 correlation. The software adopted the comparative Ct value method to calculate relative gene expression quantity F, where $F = 2^{-(\Delta\Delta Ct)}$ and ΔΔCt = (Ct value of the test target gene group - Ct value of the test group reference gene) - (Ct value of the control group target gene - Ct value of the control group reference gene).

Data availability

The raw Illumina data generated in this study were deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA412687.

RESULTS AND DISCUSSION

Transcriptome Sequencing and de Novo Assembly

An estimated 141,234,042 raw pair-end reads with a read length of 100 bp were generated. After a trimming process was performed to remove the adaptors, primer sequences, poly A tails, and short and low quality sequences, 133,263,316 high-quality reads were recovered, ranging from 64 to 68 million for each tissue (Table 1). The assembly of high-resin yield of *P. kesiya* var. *langbianensis* reads resulted in 129,501 contigs with mean sizes of 349 bp. After further clustering by the TGICL software (v2.1) (Pertea *et al*. 2003), 75,921 unigenes were obtained of approximately 45.9 Mb in total length and 605

bp in size (Table 2). Meanwhile, the assembly of low-resin yield of *P. kesiya var. langbianensis* reads resulted in 181,342 contigs with mean sizes of 293 bp, where 84,276 unigenes were obtained measuring approximately 51.8 Mb in total length and with a mean size of 615 bp (Table 2). Finally, 68,881 unigenes were obtained measuring approximately 56.5 Mb in total length and 821 bp in size in all of the high-resin yield and low-resin yield samples of *P. kesiya* var. *langbianensis* (Table 2), and the size distributions of the unigenes were compiled (Fig. 1).

Table 1. Transcriptome Sequencing of *P. kesiya* var. *langbianensis*

Samples	Raw Reads	Clean Reads	Total Clean Nucleotides	Q20 (%)	N(%)	GC(%)
High-resin yield of *P. kesiya var. langbianensis*	72,575,174	68,546,492	6,169,184,280	98.57	0.01	45.74
Low-resin yield of *P. kesiya var. langbianensis*	68,658,868	64,689,824	5,822,084,160	97.91	0.00	44.26

* Total Reads and Total Nucleotides are actually clean reads and clean nucleotides; Total Nucleotides should be more than contract provision; Q20 percentage is proportion of nucleotides with quality value larger than 20; N percentage is proportion of unknown nucleotides in clean reads; GC percentage is proportion of guanidine and cytosine nucleotides among total nucleotides.

Table 2. De Novo Assembly of *P. kesiya* var. *langbianensis* Transcriptome

	Samples	Total No.	Total Length (nt)	Mean Length (nt)	N50	Total Consensus Sequences	Distinct Clusters	Distinct Single-tons
Contigs	High-resin yield of *P. kesiya var. langbianensis*	129,501	45,258,086	349	700	-	-	-
	Low-resin yield of *P. kesiya var. langbianensis*	181,342	53,212,177	293	487	-	-	-
Unigenes	High-resin yield of *P. kesiya var. langbianensis*	75,921	45,935,640	605	1,016	75,921	20,629	55,292
	Low-resin yield of *P. kesiya var. langbianensis*	84,276	51,830,711	615	1,209	84,276	13,989	60,287
	All	68,881	56,537,842	821	1,402	68,881	23,253	45,628

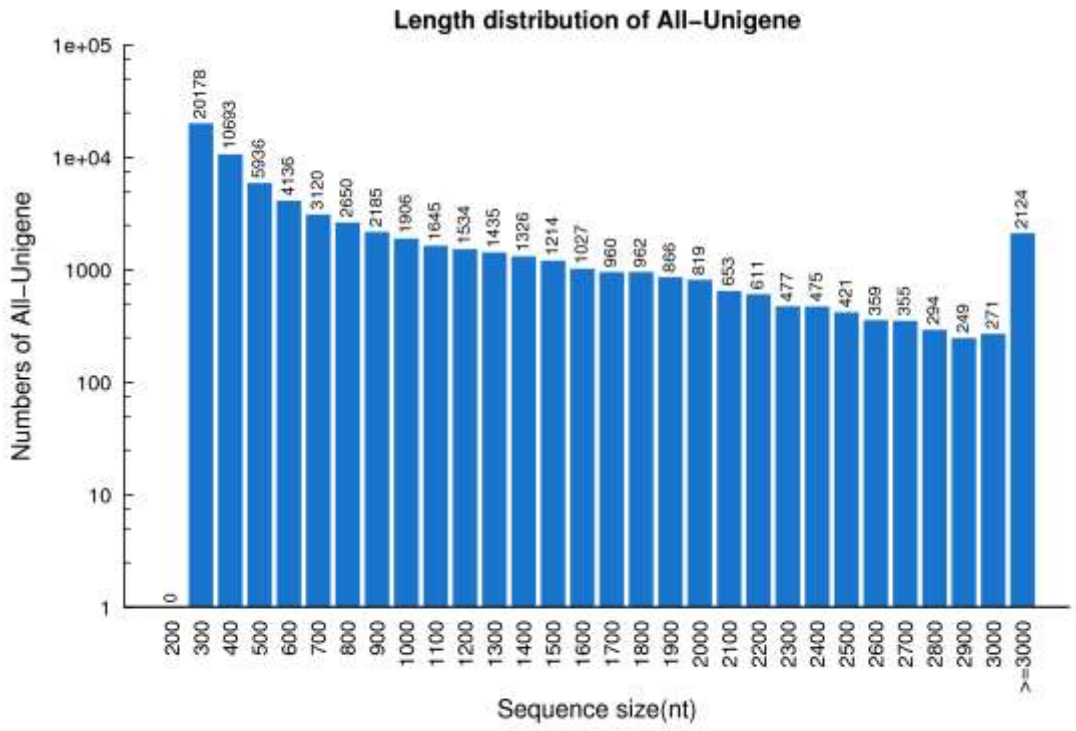


Fig. 1. The size distributions of *P. kesiya* var. *langbianensis* for all unigenes

Functional Annotation

The functional annotation of the assembled transcriptome provided insight into their molecular functions and the biological processes in an organism. The *P. kesiya* var. *langbianensis* transcriptomes were annotated based on protein similarity using BLAST searches against several public databases.

A total of 48,035 (70.0%) *P. kesiya* var. *langbianensis* unigenes were matched to known genes in the public databases, and the statistical summary for this annotation is shown in Table 3.

Table 3. Functional Annotation of *P. kesiya* var. *langbianensis* Transcriptome

Sequence	NR	NT	Swiss-Prot	KEGG	COG	GO	ALL
All unigenes of *P. kesiya var. langbianensis*	37,757	42,869	24,093	21,433	13,528	23,491	48,035

To annotate the *P. kesiya* var. *langbianensis* transcriptome, unigene sequences were first searched against the NCBI NR protein database using a cut-off E-value of 10^{-5}. A total of 37,757 unigenes were aligned to the NCBI NR protein database: 51.9% had an E-value < 1e-50, 43.3% had best hits with *Picea sitchensis*, and 12.9% were matched to *Vitis vinifera*. Most of the matched unigenes (85.9%) had greater than 40% similarity with the public genes (Fig. S1).

The GO terms were subsequently assigned to *P. kesiya var. langbianensis* unigenes based on their sequence matches to known protein sequences in the NR database.

A total of 23,491 unigenes (34.1%) were assigned to 56 functional groups using GO assignments (Fig. 2). Within the cellular component domain, the three most matched categories were "cell part," "cell," and "organelle". In the molecular process domain, the three most common categories were "catalytic activity," "binding," and "transporter activity". In the biological process domain, the three most enriched categories were "cellular process," "metabolic process," and "single-organism process". This provided sufficient data for further study on the process of primary and secondary metabolism.

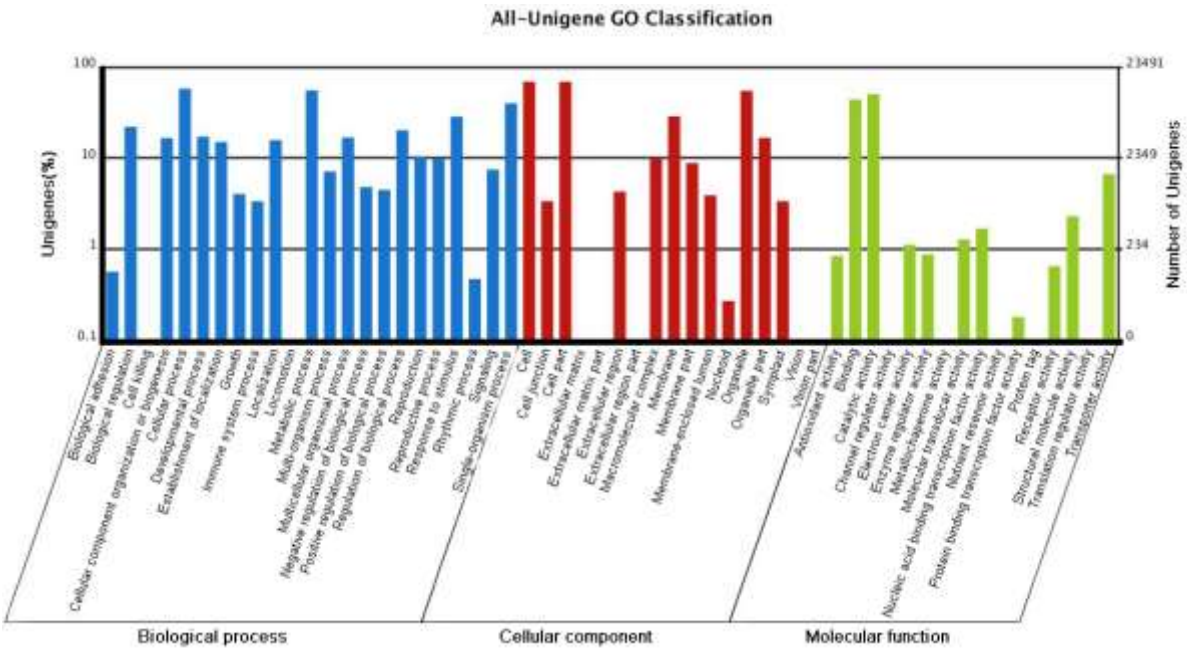


Fig. 2. GO annotation of *P. kesiya* var. *langbianensis* transcriptome

To further elucidate the functionality of the *P. kesiya* var. *langbianensis* transcriptome, the annotated unigenes were categorized into different functional groups based on the COG database (Fig. 3). Next, 13,528 sequences were classified into 25 COG categories, among which "General function prediction only" represented the largest group, followed by "Replication, recombination, and repair," and "Transcription," while "Extracellular structures" and "nuclear structure" were the smallest groups.

To further identify the active biochemical pathways in the tissue of *P. kesiya* var. *langbianensis*, the *P. kesiya* var. *langbianensis* unigenes were mapped to the reference canonical pathways in KEGG. The KEGG is thought to provide a basic platform for the systematic analysis of gene function in terms of the network of gene products. A total of 21,433 unigenes (31.1%) were annotated based on a BLASTX search of the KEGG database, and 128 biosynthesis pathways were predicted. Of these 128 KEGG pathways, the metabolic pathway was the largest, containing 4,467 members (unigene products). Other pathways included biosynthesis of secondary metabolites (2,209 members), plant-pathogen interaction (1,713 members), plant hormone signal transduction (1,012 members), RNA transport (865 members), and spliceosome (823 members). In this way, the analyses further provided sufficient data for the study of the synthetic metabolism of terpenoids from *P. kesiya* var. *langbianensis*.

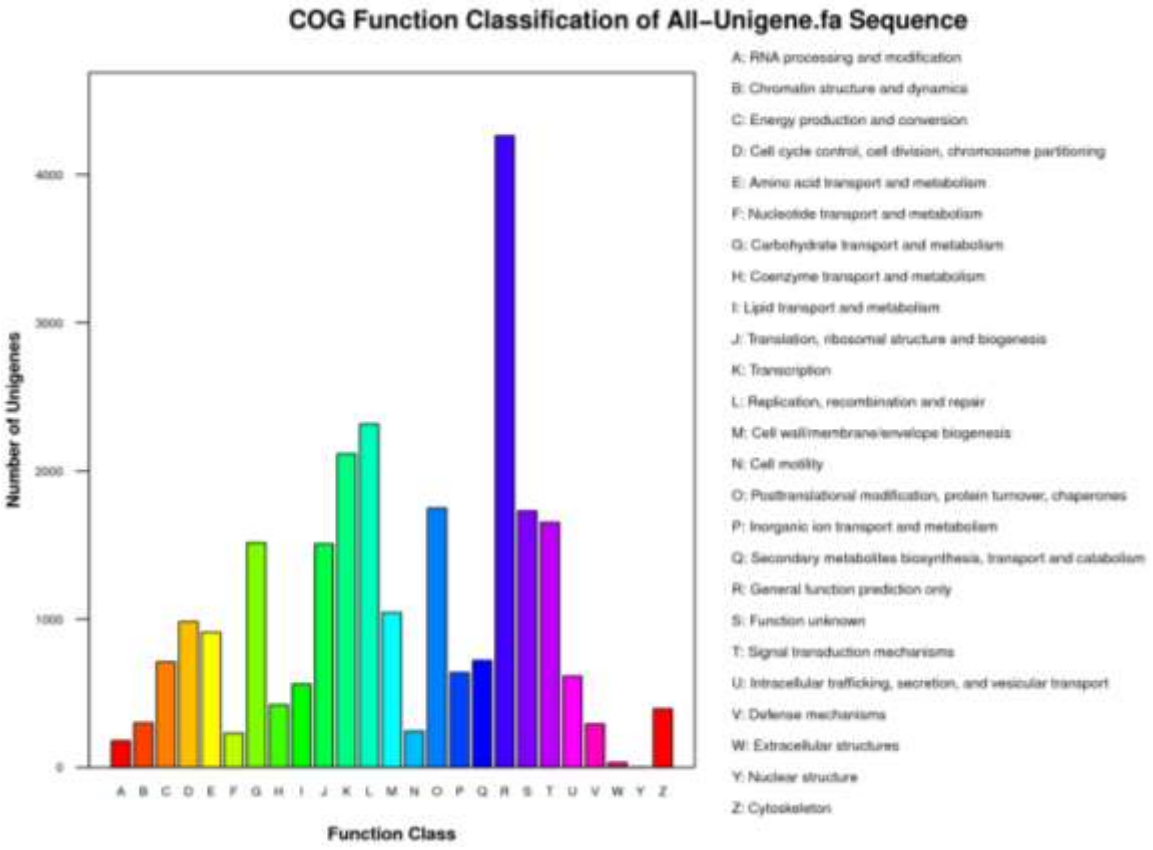


Fig. 3. COG annotation of *P. kesiya* var. *langbianensis* transcriptome

Differential Gene Expression in Resin Biosynthetic Pathway

The combination of transcriptomics and metabolomics was able to provide precise information for identifying the key functional genes in resin production. The main constituents of pine resin are bicyclic terpenes, monocyclic terpenes, and sesquiterpenes. All of the terpenoids are derived by repetitive fusion of C5 units based on the isopentane skeleton. The biosynthesis of terpenes may be divided in two ways, where MEP is the mainstream in resin biosynthesis (Cordoba *et al*. 2009) (Fig. 4). Firstly, pyruvate is reacted with glyceraldehyde 3-phosphate under the catalyzer of 1-deoxy-D-xylulose-5-phosphase (DXS) to produce 1-Deoxy-D-xylulose 5-phosphate (DXP). Then, DXP is converted to 2-C methylerythritlol 4-phosyhate (MEP) by 1-deoxy-D-xylulose-5-phosphate reductase (DXR). In the next steps, MEP is catalyzed to a mixture of isoprenyl diphosphate (IPP) and dimethyl allyl diphosphate (DMAPP) *via* the enzymes CMT, CMK, MECPS, HDS, and HDR, and IPP and DMAPP are the precursors of all terpenoid compounds. The IPP is isomerized by isopentenyl diphosphateisomerase (IDI) to DMAPP. Next, IPP synthesizes geranyl diphosphate (GPP) under the enzyme GPPS, which formulates the monoterpene (C10) skeleton and combines another IPP unit with 1'-4 head-to-tail additions to generate farnesyl diphosphate (FPP), forming the prototype for sesquiterpenes (C15). Subsequently, geranyl geranyl diphosphate (GGPP) is derived by GGPPS and FGPPS, which evolve to originate sesquiterpenes (C15) and diterpenes (C20). The next stage occurs during a series of cyclizations driven by TPS. The pentanol isoprene units C5H8 can produce very different terpenoid metabolite constituents (Gershenzon and Kreis 1999; Monson 2013; Lange and Rios-Estepa 2014). In the pine resin biosynthesis pathway, the TPS, such as

DXS, DXR, HDR, IPPI, IPPS, GPPS, FPPS, and GGPPS, were able to participate in the resin terpene production progress.

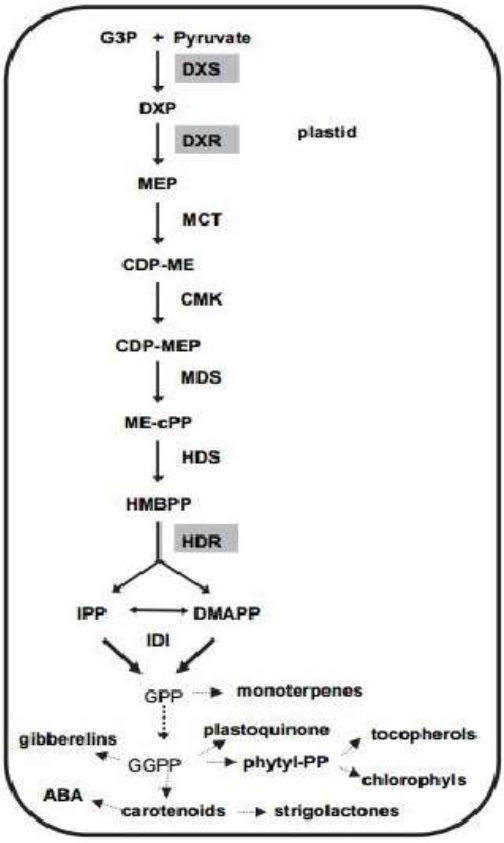


Fig. 4. The terpene biosynthesis pathway

Fig. 5. The distribution of differentially expressed genes (DEGs) in high-resin yield samples of *P. kesiya* var. *langbianensis vs.* low-resin yield samples of *P. kesiya* var. *langbianensis*. And the red bar represents up-regulated genes, and the green bar is for down-regulated genes.

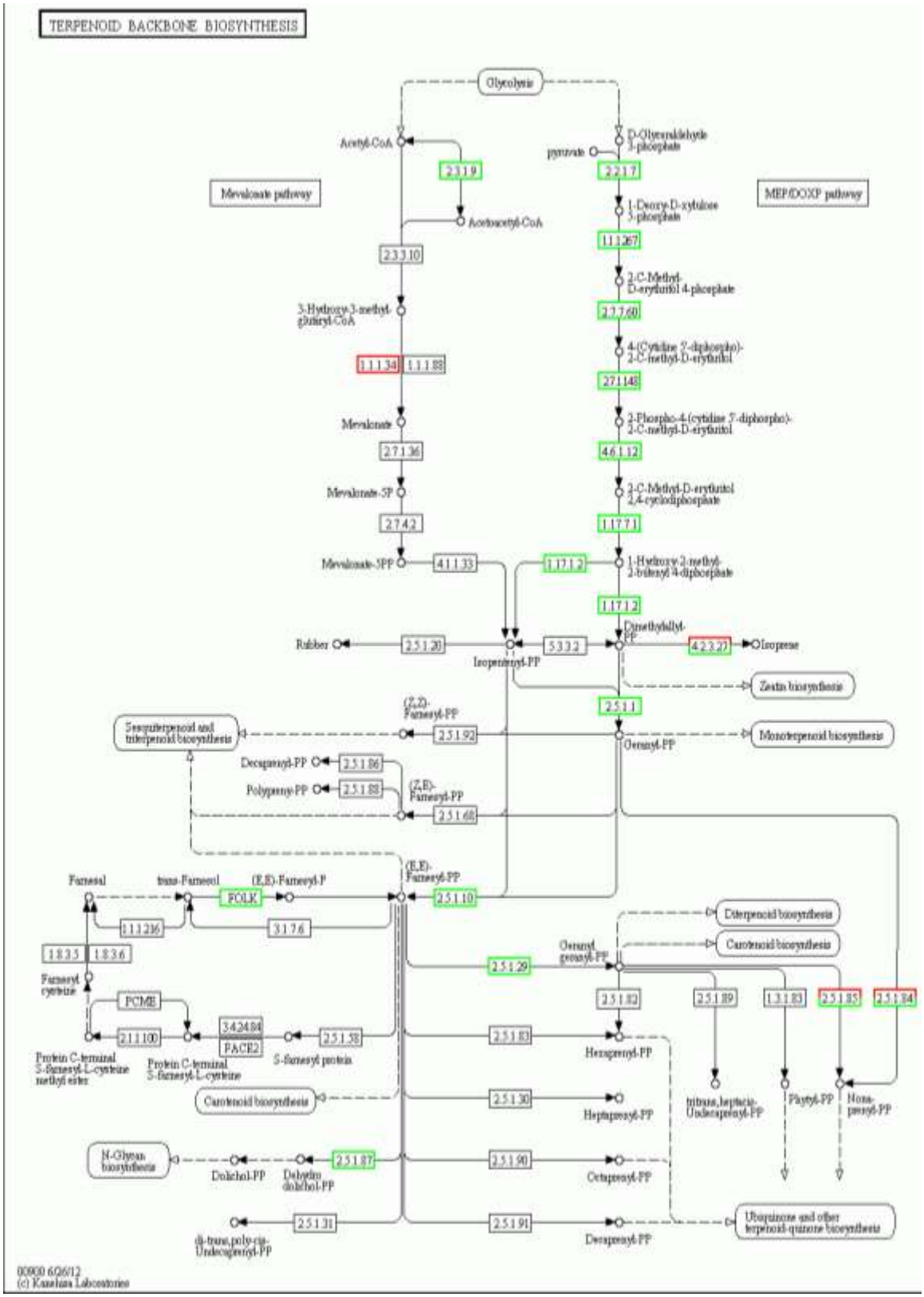


Fig. 6. Biosynthesis pathway of terpenes

Through the screening of differentially expressed genes, the genes that were differentially expressed between the two samples were obtained. When low-resin yield samples of *P. kesiya* var. *langbianensis* were compared with high-yield samples of *P. kesiya* var. *langbianensis*, the number of unigenes whose expression was up-regulated is 15340, while those with down-regulated expressions numbered 6853 (Fig. 5).

GO functional analysis provides GO functional classification annotation for DEGs as well as GO functional enrichment analysis for differentially expressed genes. Meanwhile, different genes usually cooperate with each other to exercise their biological functions. Pathway-based analysis can help to further understand genes biological functions and KEGG is the major public pathway-related database. Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways in differentially expressed genes comparing with the whole genome background.

According to the results GO classification and KEGG pathway analysis of the differentially expressed genes in transcriptome, differentially expressed genes of annotation from the metabolism of terpenoids were obtained (Fig. 6). In the differential expression analysis, the red represented the up-regulated genes and green represented the down-regulated genes; compared with mevalonate pathway (MVA) and MEP pathways, the MEP pathway was closer to the resin biosynthetic pathway, and the key enzymes DXS, DXR, MCT, CMK, MECPS, HDR, GPPS, FPPS, and GGPPS in the low-resin yield samples of *P. kesiya* var. *langbianensis* were down-regulated, which indicated that the yield of *P. kesiya* var. *langbianensis* was regulated by the key genes in the MEP pathway.

Besides, gene numbers and expressive quantities identified the differential genes according to the enzyme (EC) number from the transcriptome gene annotation (Table 4). This method selected and confirmed the gene sequences of differentially expressed genes by the qPCR verification.

The Detection of Gene Amplification Efficiency and the Establishment of Standard Curve

Partial gene sequences of nine genes encoding enzymes of MEP pathways were designed from conserved gene regions by using the software Primer 5.0 (Table 5). The amplicons for all gene fragments ranged from 100 to 210 bp.

The amplification curve and standard curve of the target genes are shown in Fig. S2, while the amplification efficiency 'E' and correlation coefficient 'R^2' are shown in Table 6.

In relation to the detection of five concentration gradients, the correlation of the Ct value and the number of the initial template dilution was good, the amplification efficiency of the reference gene and target gene was between 90% and 105%, and the dissolution curve presented a single peak.

The primers had good primer specificity, no primer dimer or heterosexual amplification, and the amplification efficiency conformed to the requirements of the comparative value method to calculate the relative expression.

Table 4. Variation of Expression Genes in MEP Metabolic Pathway of *P. kesiya* var*. langbianensis*

Gene	Gene Name	EC Number	Unigene Number	Gene Expression Up/Down	Species
DXS	1-deoxy-D-xylulose 5-phosphate synthas	2.2.1.7	Unigene7409	-1.98	*Pinus abies*
DXS	1-deoxy-D-xylulose 5-phosphate synthas	2.2.1.7	Unigene14118	-4.53	*Pinus abies*
DXR	1-deoxy-D-xylulose 5-phosphate reductoisomerase	1.1.1.267	Unigene15527	-1.94	*Pinus taeda*
DXR	1-deoxy-D-xylulose 5-phosphate reductoisomerase	1.1.1.267	Unigene7204	-1.88	*Pinus taeda*
MCT	2-C-Methyl-D-erythritol 4-phosphate cytidyltransferase	2.7.7.60	Unigene20090	-1.81	*Pinus sitchensis*
CMK	4-cytidine-2-C-methyl erythritol kinase	2.7.1.148	Unigene13142	-1.42	*Ginkgo biloba*
MECPS	2-C-methyl-D-erythritol 2, 4-cyclodiphosphate synthase	4.6.1.12	Unigene1610	-2.65	*Pinus sitchensis*
HDR	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	1.17.1.2	Unigene20413	-2.25	*Pinus densiflora*
GPPS	geranyl diphosphate synthase	2.5.1.1	Unigene27479	-4.44	*Pinus abies*
FPPS	farnesyl diphosphate synthase	2.5.1.10	Unigene21386	-0.63	*Pinus abies*
GGPPS	geranylgeranyl diphosphate synthase	2.5.1.29	Unigene23201	-0.95	*Pinus abies*
GGPPS	geranylgeranyl diphosphate synthase	2.5.1.29	Unigene10371	-1.16	*Pinus abies*
GGPPS	geranylgeranyl diphosphate synthase	2.5.1.29	Unigene1892	-0.03	*Pinus sitchensis*
GGPPS	geranylgeranyl diphosphate synthase	2.5.1.29	Unigene23597	-0.77	*Pinus abies*

Table 5. The Gene qPCR Primer Sequence and Source Information

Gene	Primer Sequence (5′ to 3′)	Annealing Temperature (°C)	Amplification Product Length (bp)	Unigene Number
Actin	F:ATTGCTGACCGTATGAGCA R:TGAATACTAGCTAGCCTCA	-	-	-
DXS	F:CGACCTTTGCCAATCTCGA R:CTGCAGCAATTGACGATCGA	59.0 59.0	115	4118
DXR	F:CACTGTTCCCTACAAGCACCTC R:TGTCCAATGTCTGAGTTCCGAT	58.9 59.4	136	15527
MCT	F:TAGACAAGGACGCCGATGA R:GCTGCCCTAATAATGGAAGATAC	57.0 57.4	144	20090
CMK	F:TTCAAGCAAGCGAAGAGGAT R:CATCTCCTTTTTCTTCTCTCACG	57.7 58.0	196	13142
MECPS	F:CCTTCACAACATTCACCTCCCT R:GCGTTTCAGTCGCAATCACA	60.5 60.2	210	1610
HDR	F:ACCTGTCCGTGGGTGTCTAAG R:CATATCTGGCCTCTGTAATGTTCT	58.9 57.6	163	20413
GPPS	F:TGCATCTCCATGGGCATTTG R:CATTGCCTTGGAATGCAGGT	58.9 59.1	161	27479
FPPS	F:CGGAGGAAAGCTGAATCGTG R:GCCAAGCACACATCCAAGAA	59.0 59.0	106	21386
GGPPS	F:GCAAGCGAGTTCGTCCTATT R:TCCTCGGCGTATATCATCATT	57.3 57.4	161	10371

Table 6. Amplification Efficiency and Correlation Coefficient of Target Genes of *P. kesiya* var. *langbianensis*

Gene	Amplification Efficiency (E%)	Correlation coefficient (R^2)
Actin	90.6	0.998
DXS	101.5	0.999
DXR	98.6	0.999
MCT	102.9	0.999
CMK	98.3	0.999
MCECPS	100.9	0.999
HDR	99.9	0.999
GPPS	101.2	0.999
FPPS	99.4	0.999
GGPPS	103.4	0.999

QF-PCR Validation of Target Genes Expression

The expression of the target genes with high-resin yield from *P. kesiya* var. *langbianensis* was controlled, the relative gene expressions in the high-resin yield and low-resin yield samples of *P. kesiya* var. *langbianensis* were calculated using the formula $F = 2^{-(\Delta\Delta Ct)}$, and the quantitative Ct values should be generally between 15 and 35. The changes in gene expression were analyzed (Fig. 7). The results showed that the expressions of the

differential genes were all down-regulated in the low-resin yield samples of *P. kesiya var. langbianensis*, which was consistent with the differential expression of target genes in the transcript. This indicated that the sequencing results of the transcriptome were accurate and viable.

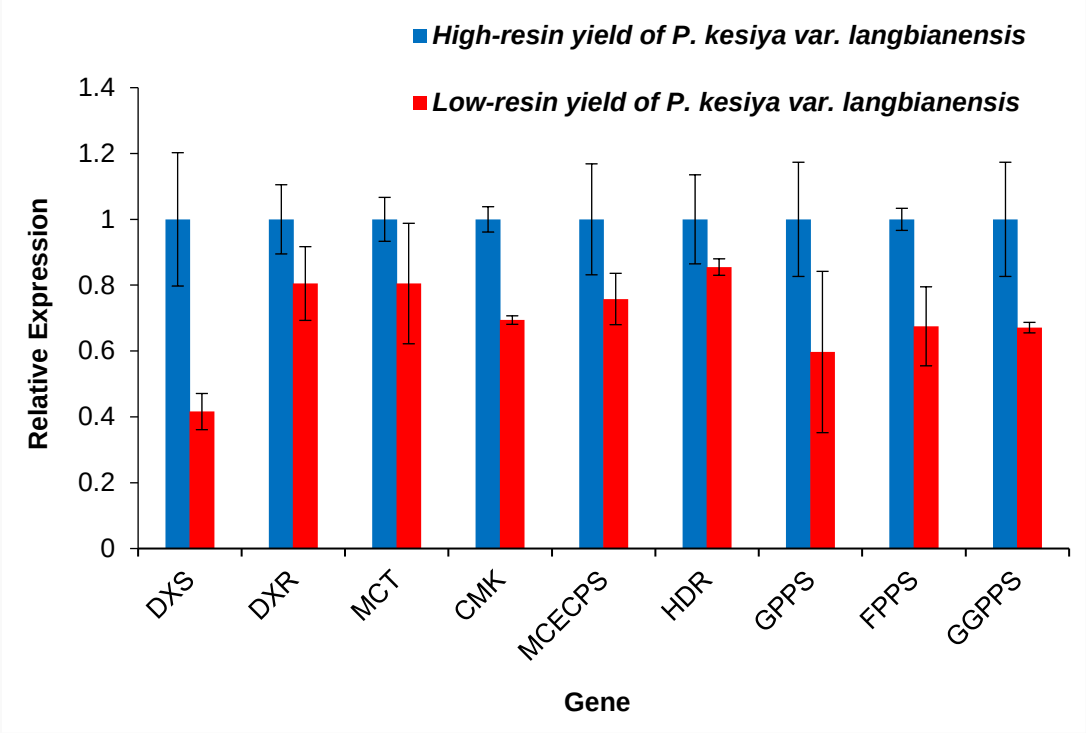


Fig. 7. The expressions of the target genes in *P. kesiya* var. *langbianensis* for high-resin yield and low-resin yield samples

In the MEP pathway, DXS, DXR, and HDR are important regulatory nodes for terpenoids. DXS is the first enzyme in the MEP pathway, which can regulate the primary metabolism and secondary metabolism of plants by controlling the supply of isoprene precursor in plastids, DXR is the key reductase enzyme to form MEP, and HDR is the last one to synthesis all the key enzyme of terpenoid precursors IPP and DMAPP, so all of them are regarded as the rate-limiting enzymes in the MEP pathway. The expression of DXS, DXR, and HDR was lower than that of the high-resin yield samples of *P. kesiya* var. *langbianensis*, indicating that the expression of DXS, DXR, and HDR was positively correlated with the yield of *P. kesiya* var. *langbianensis*. All terpenes, including the main components of the resin yield of *P. kesiya var. langbianensis*, are based on IPP and DMAPP as substrates. A variety of precursors of terpenoids are produced under the action of isoprene transferases, and monoterpenes, sesquiterpenoids, and diterpenoids are produced by terpene synthase. The GPPS, FPPS, and GGPPS are the synthases that form monoterpenes, sesquiterpenes, and diterpenoids as direct precursors (GPP, FPP, GGPP), respectively. According to the results of qPCR, GPPS, FPPS, and GGPPS were down-regulated in low-resin yield samples of *P. kesiya var. langbianensis*, which was consistent with the results of transcription groups. This indicated that the ability to synthesize monoterpenes, sesquiterpenes, and diterpenoids was decreased in low-resin yield samples of *P. kesiya* var. *langbianensis*, which led to the decrease in total synthesis of

monoterpenes, sesquiterpenes, and sesquiterpenes in low-resin yield samples of *P. kesiya* var. *langbianensis*. Thus, in summary, accordingly choosing nine differentially expressed genes of MEP, gene expression was significantly down-regulated in low-resin yield samples of *P. kesiya* var. *langbianensis.* Furthermore, DXS, DXR, and HDR, as the rate-limiting enzymes, and GPPS, FPPS, GGPPS, as the key enzymes for the synthesis of monoterpenes, sesquiterpenes, and diterpenoids in the MEP pathway, respectively, their levels of expression directly affected the yield of *P. kesiya var. langbianensis.* This study has the potential to accelerate the understanding of the molecular mechanism of pine resin production for the better application of molecular breeding in the future.

CONCLUSIONS

1. Functional annotation and differential gene expression in the resin biosynthetic pathway indicated that the yield of *P. kesiya var. Langbianensis* was related to terpenoids metabolism pathway synthase.
2. Using QF-PCR validation, the nine key enzyme genes involved in the resin metabolic pathway were all down-regulated in the low-resin yield samples of *P. kesiya var. Langbianensis*, which was consistent with the differential expression of target genes in the transcript. It was further established that the key enzyme genes of gene levels of expression directly affected the yield of *P. kesiya var. Langbianensis.*

ACKNOWLEGEMENTS

This work was supported by the State Forestry Administration Public Welfare Forestry Project of Breeding and Functional Gene Cloning and Identification of *Pinus Kesiya* (No. 201304105) and the National Natural Science Foundation of China (No. 31660029).

REFERENCES CITED

Audic, S., and Claverie, J. M. (1997). "The significance of digital gene expression profiles," *Genome Research* 7(10), 986. DOI: 10.1101/gr.7.10.986

Burdon, R. D., and Wilcox, P. L. (2007). "Population management: Potential impacts of advances in genomics," *New Forests* 34(2), 187-206. DOI: 10.1007/s11056-007-9047-6

Cai, N. H., Xu, Y. L., and Wang, D. W. (2017). "Identification and characterization of microsatellite markers in *Pinus kesiya* var. *Langbianensis* (Pinaceae)," *Applications in Plant Sciences* 5(2), 1600126. DOI: 10.3732/apps.1600126

Cameron, M., Williams, H. E., and Cannane, A. (2005). "Improved gapped alignment in BLAST," *IEEE/ACM Transactions on Computational Biology & Bioinformatics* 1(3), 116-129. DOI: 10.1109/tcbb.2004.32

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., and Talon, M. (2005). "Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics* 21(18), 3674-3676. DOI: 10.1093/bioinformatics/bti610

Cordoba, E., Salmi, M., and León, P. (2009). "Unraveling the regulatory mechanisms that modulate the MEP pathway in higher plants," *Journal of Experimental Botany* 60(10), 2933-2943. DOI: 10.1093/jxb/erp190

Gen, F. F., and Xiao, F. K. (2015). "Induction and proliferation of embryonic calluses from mature zygotic embryos of *P. kesiya* var. *Langbianensis*," *Journal of Northeast Forestry University* 43(05), 60-63. DOI: 10.13759/j.cnki.dlxb.20150522.022

Gershenzon, J., and Kreis, W. (1999). "Biochemistry of terpenoids: Monoterpenes, sesquiterpenes, diterpenes, sterols, cardiac glycosides, and steroid saponins," in: *Biochemistry of Plant Secondary Metabolism*, M. Wink (ed.), CRC Press, Boca Raton, FL, USA, pp. 222-299.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., and Thompson, D. A. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology* 29(7), 644-652. DOI: 10.1038/nbt.1883

Hall, D. E., Yuen, M. M. S., Jancsik, S., Quesada, A. L., and Dullat, H. K. (2013). "Transcriptome resources and functional characterization of monoterpene synthases for two host species of the mountain pine beetle, lodgepole pine (*Pinus contorta*) and jack pine (*Pinus banksiana*)," *BMC Plant Biology* 13(1), 80. DOI: 10.1186/1471-2229-13-80

Iseli, C., Jongeneel, C. V., and Bucher, P. (1999). "ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences," *Proceedings International Conference on Intelligent Systems for Molecular Biology* 15(2), 138-148.

Jiang, Y. D., Li, S. G., and Yang, Z. Y. (2006). "Effect of soil chemical property on growth of young *Pinus kesiya* var. *langbianensis* planted forests," *Journal of Northeast Forestry University* 34(1), 25-27. DOI: 10.13759/j.cnki.dlxb.2006.01.009

Lange, B. M., and Rios-Estepa, R. (2014). "Kinetic modeling of plant metabolism and its predictive power: Peppermint essential oil biosynthesis as an example," *Methods in Molecular Biology* 1083(1), 287-311. DOI: 10.1007/978-1-62703-661-0 17

Liu, X. Y., Chen, Y. H., and Li, M. (2003). "The study of high-resin yield of *P. kesiya* var. *Langbianensis*," *Journal of Sichuan Forestry Science and Technology* 24(2), 51-52. DOI: 10.16779/j.cnki.1003-5508.2003.02.013

Monson, R. K. (2013). "Metabolic and gene expression controls on the production of biogenic volatile organic compounds," in: *Biology, Controls and Models of Tree Volatile Organic Compound Emissions*, Volume 5, U. Niinemets and R. K. Monson (eds.), Springer, Berlin, pp. 153-179. DOI: 10.1007/978-94-007-6606-8_6

Nie, H. Q. (2017). "Study on the growth of pine resin industry of *Pinus massoniana*," *Journal of Green Science and Technology* 33(5), 71-75. DOI: 10.16663/j.cnki.lskj.2017.05.033

Niu, S. H., Li, Z. X., and Yuan, H. W. (2013). "Transcriptome characterisation of *Pinus tabuliformis* and evolution of genes in the Pinus phylogeny," *BMC Genomics* 14(1), 263. DOI: 10.1186/1471-2164-14-263

Pertea, G., Huang, X., and Liang, F. (2003). "TIGR gene indices clustering tools (TGICL): A software system for fast clustering of large EST datasets," *Bioinformatics* 19(5), 651-652. DOI: 10.1093/bioinformatics/btg034

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). "The COG database: A tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research* 28(1), 33-36. DOI: 10.1093/nar/28.1.33

Wang, J. F., Chen, Y. P., Yao, K., Wilbon, P. A., and Zhang, W. (2012). "Robust antimicrobial compounds and polymers derived from natural resin acids," *Chemical Communications* 48(6), 916-918. DOI: 10.1039/c1cc16432e

Wang, Y., Zhou, X., and Bi, W. (2015a). "Identification and characterization of a 1-deoxy-D xylulose 5-phosphate synthase gene from *Pinus kesiya* var. *Langbianensis*," *Forest Research* 28(6), 833-838. DOI: 10.13275/j.cnki.lykxyj.2015.06.011

Wang, Y., Zhou, X., and Bi, W. (2015b). "Cloning and sequence analysis of 1- hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate reductase gene cDNA from *Pinus kesiya* var. *langbianensis*," *Journal of Guangxi Plants* 35(5), 721-727. DOI: 10.11931 / guihaia.gxzw201409048

Xia, Z., Xu, H., Zhai, J., Li, D., and Luo, H. (2011). "RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*," *Plant Molecular Biology* 77(3), 299-308. DOI: 10.1007/s11103-011-9811-z

Ye, J., Fang, L., Zheng, H., Zhang, Y., and Chen, J. (2006). "WEGO: A web tool for plotting GO annotations," *Nucleic Acids Research* 34(Web Server issue), W293-297. DOI: 10.1093/nar/gkl031

Zhao, N., Shi, R., Li, B., and Xiong, Z. (2016). "Correlation between indexes of resin tapping and chemical components of *Pinus kesiya* var. *Langbianensis*," *Journal of West China Forestry Science* 45(3), 63-68. DOI: 10.16473/j.cnki.xblykx1972.2016.03.011

Zimin, A., Stevens, K. A., and Crepeau, M. W. (2014). "Sequencing and assembly of the 22-Gb loblolly pine genome," *Genetics* 196(3), 875-890. DOI: 10.1534/genetics.113.159715

Article submitted: September 25, 2017; Peer review completed: December 16, 2017; Revised version received: December 24, 2017; Accepted: December 27, 2017; Published: January 26, 2018.
DOI: 10.15376/biores.13.1.1852-1871

APPENDIX

SUPPLEMENTARY DATA

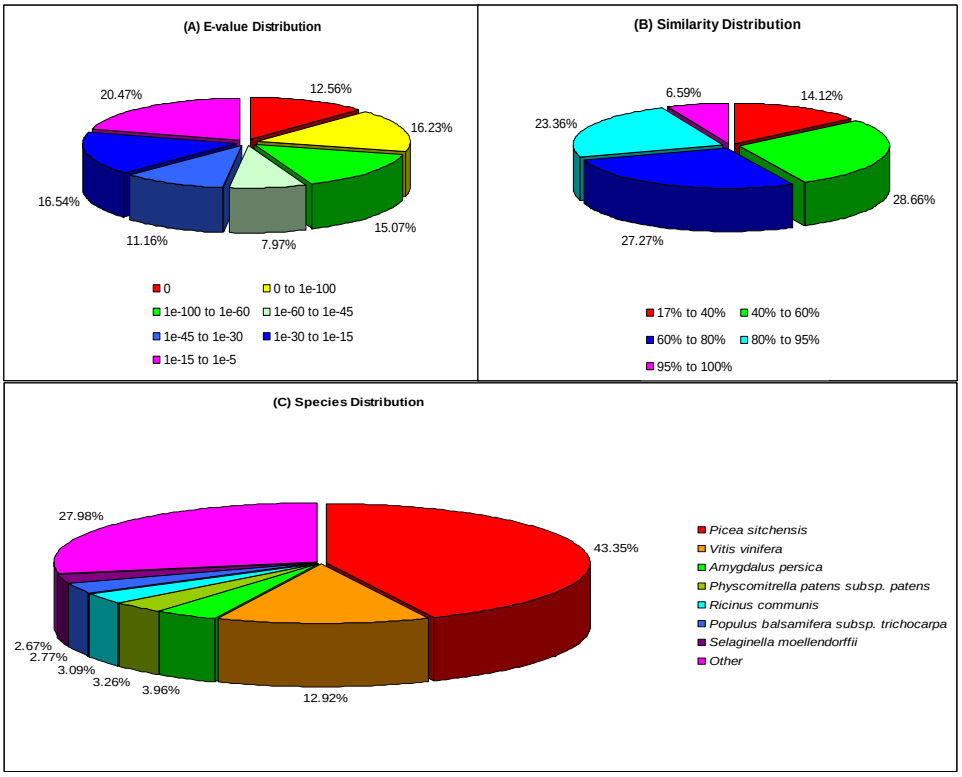
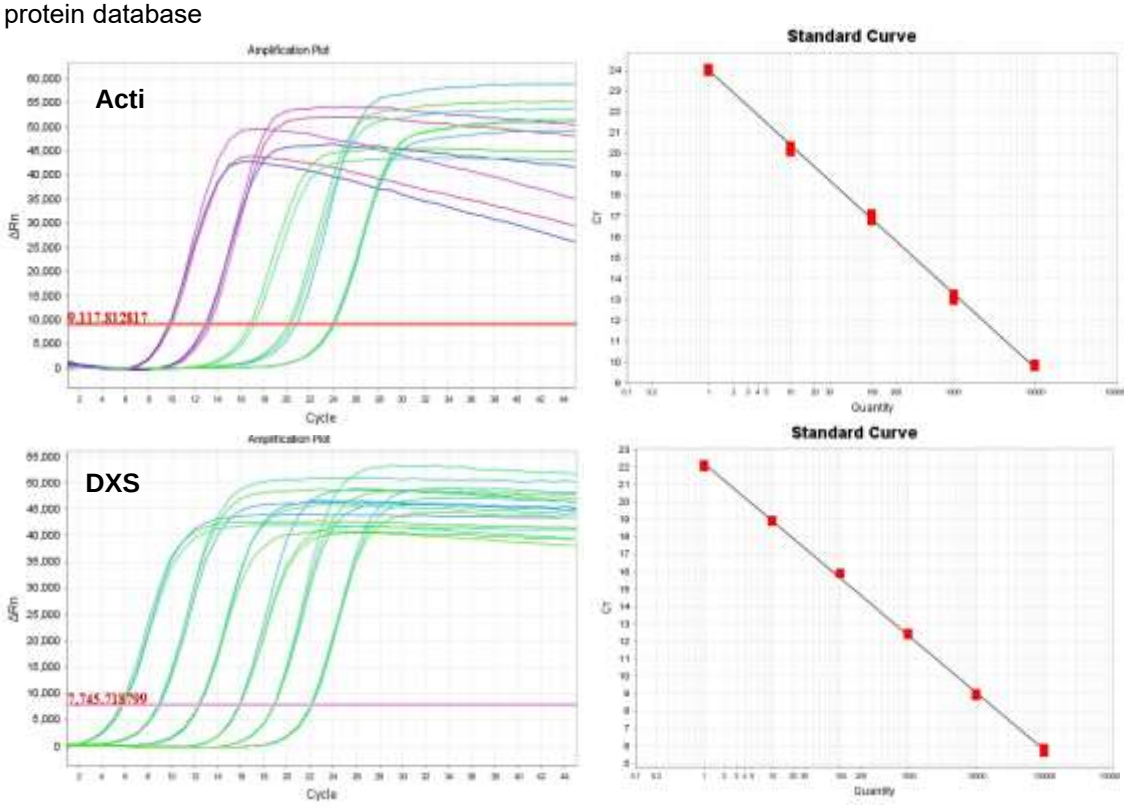
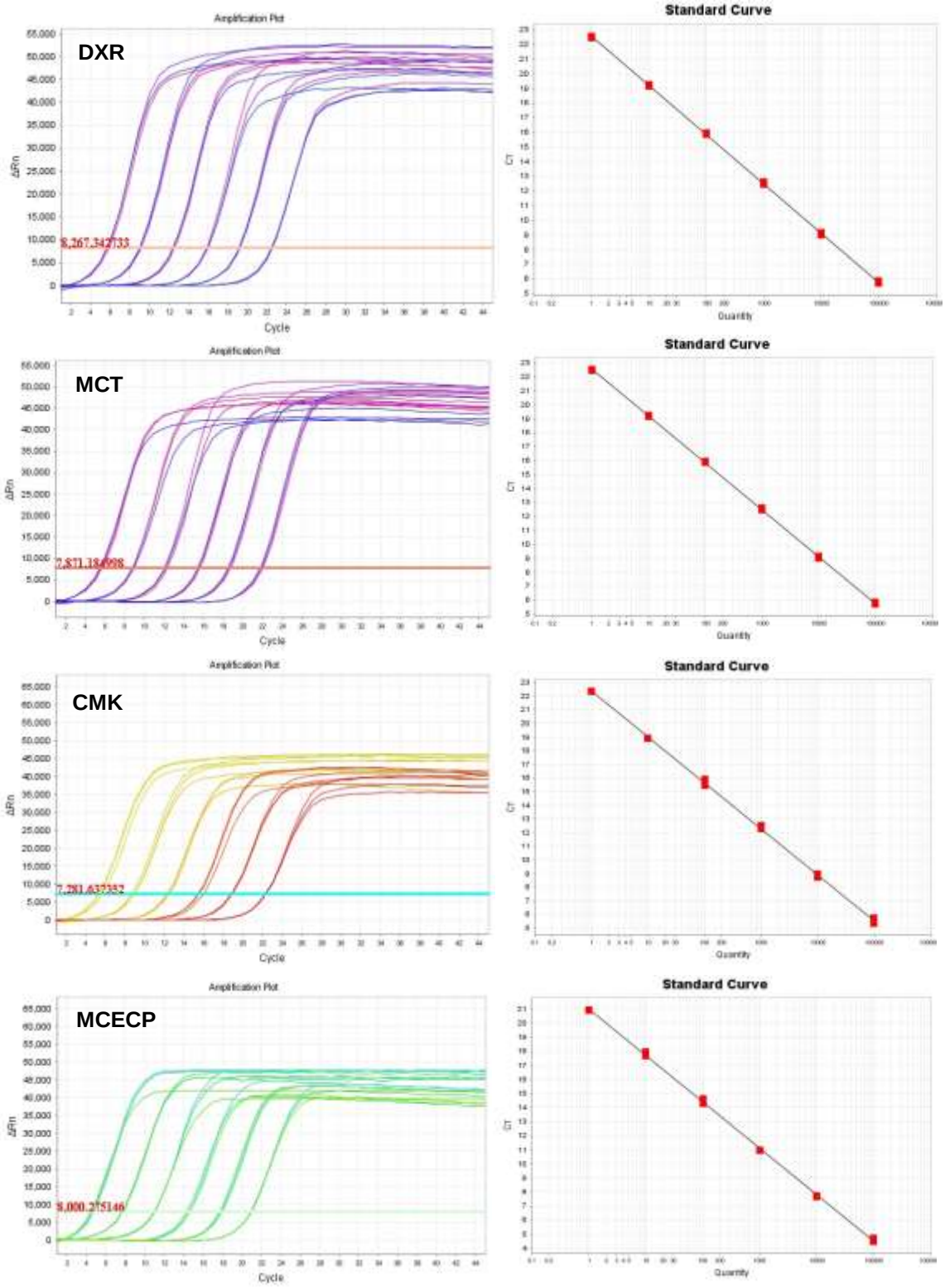


Fig. S1. Alignment between *P. kesiya* var. *langbianensis* unigenes and NCBI non-redundant protein database




Amplification Plot
DXR
8,267.342733
ΔRn
Cycle
Standard Curve
Ct
Quantity
Amplification Plot
MCT
ΔRn
Cycle
Standard Curve
Ct
Quantity
Amplification Plot
CMK
ΔRn
Cycle
Standard Curve
Ct
Quantity
Amplification Plot
MCECP
8,000.275146
ΔRn
Cycle
Standard Curve
Ct
Quantity

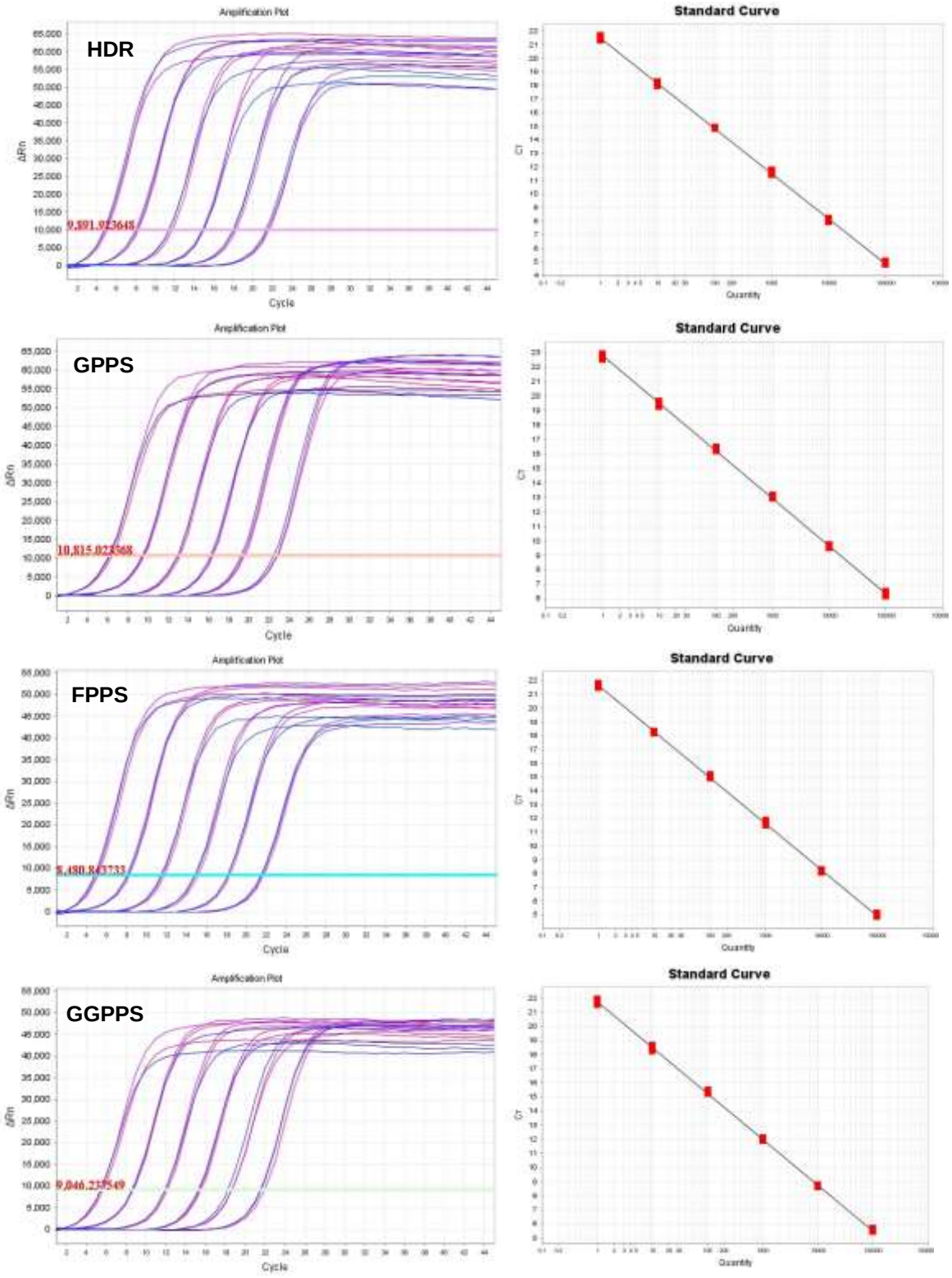


Fig. S2. Amplification curve and the standard curve of genes of *P. kesiya* var. *langbianensis*