Macrothink Institute™

Journal of Applied Biotechnology
ISSN 2327-0640
2017, Vol. 5, No. 2

Transcriptome Analysis, Marker Discovery and Pigment Biosynthesis of Red-leaf *Juglans regia*

Xin Chen

Shandong Institute of Pomology, Shandong Provincial Key Laboratory of Fruit Tree Biotechnology Breeding, Tai'an, Shandong, China

Li Xu

Shandong Institute of Pomology, Shandong Provincial Key Laboratory of Fruit Tree Biotechnology Breeding, Tai'an, Shandong, China

Xiao-juan Zong

Shandong Institute of Pomology, Shandong Provincial Key Laboratory of Fruit Tree Biotechnology Breeding, Tai'an, Shandong, China

Hai-rong Wei

Shandong Institute of Pomology, Shandong Provincial Key Laboratory of Fruit Tree Biotechnology Breeding, Tai'an, Shandong, China

Jia-wei Wang

Shandong Institute of Pomology, Shandong Provincial Key Laboratory of Fruit Tree Biotechnology Breeding, Tai'an, Shandong, China

Dong-zi Zhu

Shandong Institute of Pomology, Shandong Provincial Key Laboratory of Fruit Tree Biotechnology Breeding, Tai'an, Shandong, China

Macrothink Institute™

Journal of Applied Biotechnology
ISSN 2327-0640
2017, Vol. 5, No. 2

Yue Tan

Shandong Institute of Pomology, Shandong Provincial Key Laboratory of Fruit Tree Biotechnology Breeding, Tai'an, Shandong, China

Mao-run Fu

College of Food Science and Engineering, Qilu University of Technology, Jinan, China

Qing-zhong Liu (Corresponding author)

Shandong Institute of Pomology, No. 66 Longtan Road, Taian, China

Received: March 1, 2017 Accepted: May 28, 2017 Published: June 4, 2017

Doi: 10.5296/jab.v5i2.11345 URL: http://doi.org/10.5296/jab.v5i2.11345

Abstract

Red-leaf trait rarely occurs in *Juglans regia*, and the genetic mechanism underlying this phenomenon is still unknown. In this study, we attempted to provide insight into the comprehensive transcriptome of red-leaf *J. regia* by RNA-Seq using Illumina SeqTM2000 platform. A total of 33,488,602 high-quality reads (3.35G cleans bases) were obtained and assembled into 53,782 unigenes. A total of 3,683 unigenes were annotated by using basic local alignment search tool to search against protein databases, All the matched unigenes were categorized by gene ontology analysis, and 3,466 were assigned to metabolism, among which 74 were mapped to anthocyanin, carotenoid, and betalain biosynthetic pathways by Kyoto Encyclopedia of Genes and Genomes analysis. Approximately 656 transcription factors were isolated including MYB, NAC, and bHLH. Additionally, a total of 13,981 simple sequence repeats, 41,088 single nucleotide polymorphisms, and 5,860 insertions and deletions were determined from *J. regia* transcriptome. Therefore, the current *J. regia* transcriptome provides deep insight into the molecular basis of red-leaf breeding of *J. regia*.

Keywords: *Juglans regia*, Red-leaf, Marker discovery, Pigment biosynthesis

1. Introduction

Walnut (*Juglans regia*) belongs to the *Juglans* genus of Juglandaceae family and is a native to the region from Balkans eastward to the Himalayas and southwest China; this plant is now widely spread in Asia, America, and southern and eastern Europe (Nael Abu Taha, 2011). *J. regia* is a deciduous tree; its fruits are nutritional nuts rich in unsaturated fatty acids, tocopherols, and phytosterols (Amaral et al., 2008), and its leaves have been used as traditional medicine for the treatment of diabetes mellitus, skin inflammations, toothache,

venous insufficiency, and ulcers (Hosseini et al., 2014b; Paudel et al., 2013). *J. regia* leaves reportedly have antioxidative, antimicrobial, antihypertensive, antihelmintic, and hypoglycemic effects (Hosseini et al., 2014a; Nael Abu Taha, 2011; Pereira et al., 2007; Qureshi et al., 2014; Zhao et al., 2014).

Different colors are determined by different kinds of plant pigments. Generally, green plants contain abundant chlorophyll, as well as carotenoids and anthocyanin. Among these components, chlorophyll is predominant and is responsible for the green color of plant leaves. However, red-colored leaves naturally exist in some plants, such as *Gossypium hirsutum*, *Cotinus coggygria*, *Brassica oleracea*, and *Prunus cerasifera* (Cai et al., 2014; Wang et al., 2013). The content of each compound determines the different colors of plant leaves. A high amount of chlorophyll makes leaves green, and a low amount makes them red. The red pigmentation of *G. hirsutum* L. leaf, crabapple leaf, and *Jatropha curcas* (L.) new leaf were reportedly induced by anthocyanin accumulation (Ranjan et al., 2014; Tian et al., 2015). In the red leaves of *Sorghum bicolor*, carotenoids are the predominant compounds, followed by flavonoids and phenolic acids; a small amount of chlorophyll (a and b) was also present in *S. bicolor* red leaves (Abugri et al., 2013). The ever-red leaf trait rarely occurs in the *Juglans* genus, thereby providing us a novel material for leaf color study. The red coloration of leaf is an attractive feature for ornamental value, and the compounds contributing to red coloration in leaves have biological and ecological importance. For example, anthocyanin can protect plants against pathogens and insects and can attract insect pollinators (Mouradov and Spangenberg, 2014); carotenoids participate in photosynthesis, photomorphogenesis, and photoprotection. However, information on the molecular mechanism underlying pigment biosynthesis in red leaves of *J. regia* is lacking. In 2012, Wu *et al*., performed the genome sequencing of *J. regia* by bacterial artificial chromosome (BAC) technology, yielding 31.2 Mb bases which accounted for only 5.1% of the whole genome(Wu et al., 2012). And the whole-genome of *J. regia* was sequenced by the end of 2014, however the data was still not available for public. More recently, next-generation sequencing (NGS) is a cost-efficient way to apply in gene discovery and analysis of gene expression for non-model species without reference genome.

Generally, pigment biosynthesis in plants is controlled by structural genes in pigment biosynthetic pathways and regulatory genes, e.g., transcription factors (TFs). The aim of this study was to explore molecular markers and pigment biosynthesis of the red-leaf trait by NGS method.

2. Methods

2.1 Sample Collection and Preparation

The red leaves of *J. regia* were collected from National Clonal Plant Germplasm Repository, Taian, Shandong Province, China. Total RNA was isolated from *J. regia* leaves using Trizol (Invitrogen) and was treated with DNase I to remove the genomic DNA. RNA degradation and contamination were checked on 1% agarose gels. RNA purity was assessed using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). RNA concentration was measured using Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA,

USA), and RNA integrity (RIN) was quantified using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

2.2 Transcriptome Sample Preparation for Sequencing

A total amount of 3 μg RNA per sample was used for sample preparation. All prepared RNA samples had RIN values above 8. Sequencing libraries were generated using Illumina TruSeq™ RNA Sample Preparation Kit (Illumina, San Diego, USA) following manufacturer's recommendations, and index codes were added to attribute sequences to each sample. The mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was performed using divalent cations under elevated temperature in Illumina proprietary fragmentation buffer. First strand cDNA was synthesized using random oligonucleotides and SuperScript II. Second strand cDNA synthesis was subsequently performed using DNA polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities, and the enzymes were removed. After adenylation of the 3' ends of DNA fragments, Illumina PE adapter oligonucleotides were ligated to prepare for hybridization. To select cDNA fragments that are preferentially 200 base pair (bp) in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). DNA fragments with ligated adapter molecules on both ends were selectively enriched using Illumina PCR Primer Cocktail in a 10-cycle PCR reaction. Products were purified (AMPure XP system) and quantified using the Agilent high sensitivity DNA assay on the Agilent Bioanalyzer 2100 system.

2.3 Clustering and Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq 2000 platform, and 90 bp paired-end reads were generated.

2.4 Analysis of Illumina Sequencing Data

Raw reads were initially processed through in-house perl scripts, and the clean reads were obtained by removing reads containing adapter and poly-N and low quality reads. All downstream analyses were based on high-quality clean reads. The transcriptome assembly was accomplished using Trinity software (Grabherr et al., 2011). Gene function was annotated after sequences were searched against the following databases: NCBI non-redundant protein sequences (Nr), NCBI non-redundant nucleotide sequences (Nt), Protein family (Pfam), Clusters of Orthologous Groups of proteins (KOG), Swiss-Prot (A manually annotated and reviewed protein sequence database), Kyoto Encyclopedia of Genes and Genomes (KEGG) Ortholog database (KO), and Gene Ontology (GO). To discover single nucleotide polymorphisms (SNPs), Picard-tools v1.41 (http://picard.sourceforge.net) and Samtools v0.1.18 (http://sourceforge.net/projects/samtools/files/samtools/) were used to sort and remove duplicated reads and merge the bam alignment results of each sample. GATK2 software was used to perform SNP calling, and only SNPs with distances higher than 5 were retained with the GATK standard filter method (McKenna et al., 2010). Simple sequence

repeats (SSRs) of the transcriptome were identified using microsatellite identification tool. Gene expression patterns were estimated by RNA-Seq by Expectation-Maximization (RSEM). First, clean data were mapped back to the assembled transcriptome, and the read counts for each gene were achieved from the mapping results. Finally, the read counts were normalized into Reads Per Kilo bases per Million mapped Reads (RPKM).

3. Results and Discussions

3.1 Assembled transcriptome

To obtain an overview of comprehensive transcripts of *J. regia* red leaves, a cDNA library of *J. regia* red leaves was generated and pair end sequenced using Illumina HisSeqTM2000 platform, thereby generating 34,634,887 raw reads after base calling. Generally, the error rate of each sequenced base should be below 1%, and error rate would increase with increasing sequence read length. The error rate distribution is shown in Figure 1. After removing adapter-related sequences, low quality reads and reads containing N, 33,488,602 clean reads were obtained with 3.35 G clean bases in total, and the clean reads were used for bioinformatics analysis. All clean reads were assembled into 100,605 transcripts using Trinity software, with mean length of 1,200 bp. The longest transcript of each gene was defined as a unigene, and 53,782 unigenes (including 42,966,265 nucleotides) were obtained with a mean length of 799 bp (Table 1). The entire transcriptome (SRA accession Number SRP054989) assembled was used as reference sequences for further analysis. The length distributions of all transcripts and unigenes are shown in Additional File 1.

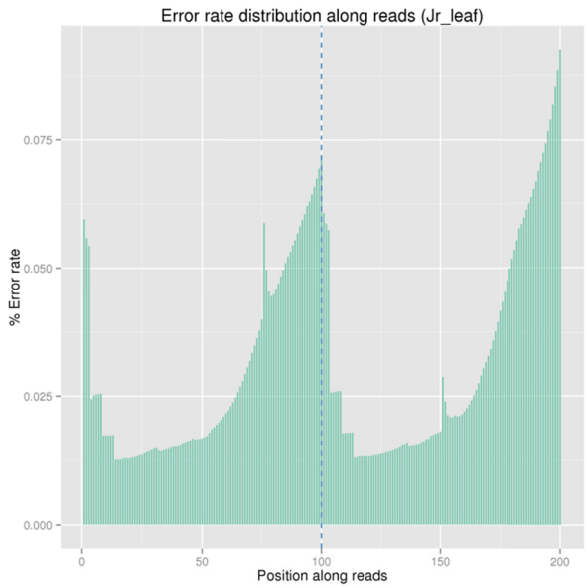

Figure 1. Error rate distribution along reads of *J. regia* transcriptome

Table 1. Summary of *J. regia* Transcriptome and de novo assembly

	Raw Reads	Clean Reads	Clean Bases	Error (%)			
Numbers/Rate	34634887	33488602	3.35G	0.03			
	Min length	Mean length	Median length	Max length	N50	N90	Total nucleotides
Transcripts	201	1200	813	14850	1984	516	120691294
Unigenes	201	799	398	14850	1532	294	42966265

3.2 Gene Function Annotation

For functional annotation of *J. regia* transcriptome, all unigenes were searched against seven public databases. Using this approach, 28,351 unigenes (52.71% of all unigenes) returned a significant basic local alignment search tool result. Among these unigenes, 3,683 (6.84%) were annotated in all databases (Table 2). Approximately 47.29% of the total unigenes had no homology in these databases, thereby suggesting that these transcripts may be unique to *J. regia*.

Table 2. Statistics of annotation percentages of *J. regia* unigenes in public databases

	Number of Unigenes	Percentage (%)
Annotated in NR	26315	48.92
Annotated in NT	14573	27.09
Annotated in KO	7505	13.95
Annotated in SwissProt	18289	34
Annotated in PFAM	17257	32.08
Annotated in GO	19934	37.06
Annotated in KOG	8969	16.67
Annotated in all Databases	3683	6.84

Annotated in at least one Database	28351	52.71
Total Unigenes	53782	100

GO is a classification system that can describe the biological process (BP), cellular component (CC), and molecular function (MF) of genes. For classification of gene functions, GO assignment was performed. A total of 19,934 unigenes were assigned to one or more GO terms, and all these 19,934 unigenes were categorized into 65 functional groups, which are distributed to three main categories, as follows: BP (53,093), MF (25,579), and CC (37,311) (Figure 2). However, the categories with no more than 20 unigenes are not shown in Figure 1. Within the BP category, cellular (12290 unigenes, 23.15%), metabolic (11688 unigenes, 22.01%), and single-organism (5790 unigenes, 10.90%) processes were the most enriched. Under the CC category, cell (7693 unigenes, 20.62%), cell part (7679, 20.58%), and organelle (5279 unigenes, 14.15%) were the most highly represented GO terms. In the MF category, the majority of unigenes were involved in binding (11,647 unigenes, 45.53%), catalytic activity (10,107 unigenes, 39.51%), and transporter activity (1,438 unigenes, 5.62%).

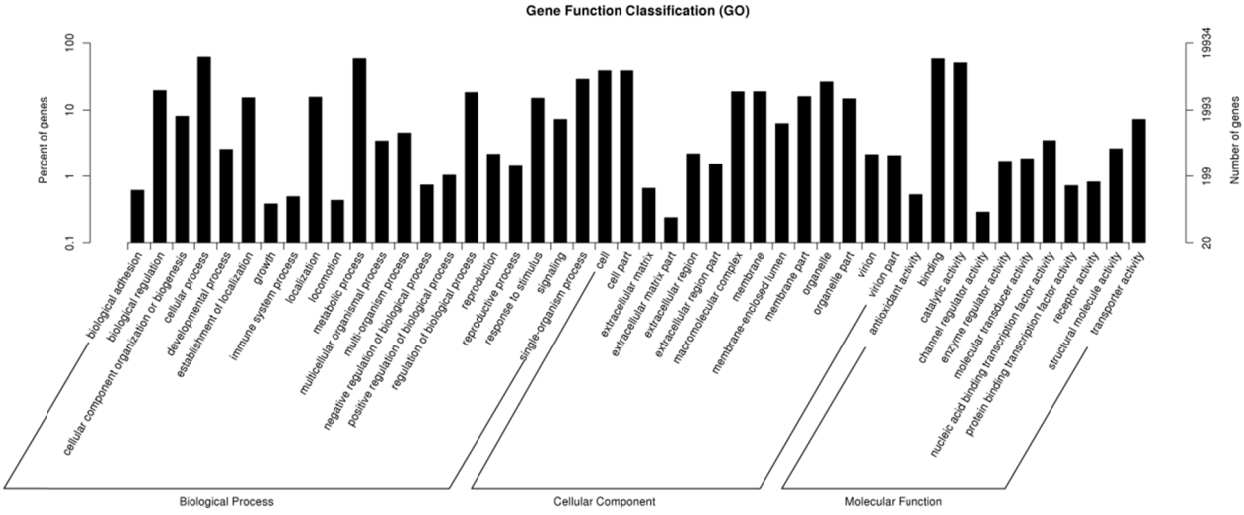

Figure 2. GO Classification of *J. regia* unigenes

The three main categories of GO terms were BP, CC, and MF. The horizontal axis was a GO term, and the vertical axis represented the number and percentage of annotated unigenes. The categories with numbers below 20 are not shown in this Figure.

EuKaryotic Ortholog Groups (KOG) analysis was used to further evaluate the functions of annotated unigenes specific to eukaryons. A total of 8,969 unigenes were annotated into 26

KOG groups, the cluster of general function prediction only (1,610 unigenes, 17.95%) was the largest group, followed by post-translational modification, protein turnover, chaperon (1,207 unigenes, 12.90%), and signal transduction (822 unigenes, 9.16%). The unnamed protein (1 unigenes, 0.00%), cell motility (3 unigenes, 0.00%), and extracellular structures (30 unigenes, 0.33%) represented the smallest groups (Figure 3).

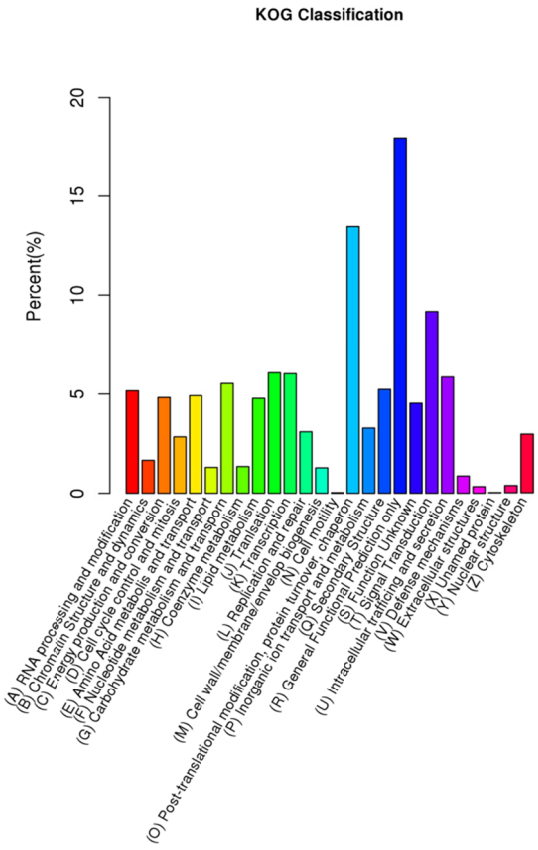

Figure 3. KOG Classification of *J. regia* unigenes

These 8,969 unigenes of *J. regia* were assigned into 26 KOG categories. The x-axis comprised KOG categories, whereas the y-axis comprised percentage of assigned unigenes.

After KO annotation analysis, we further classified the unigenes based on the KEGG metabolic pathways. KEGG pathways cover the six main groups, such as metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases. In total, 7,505 unigenes were mapped to 241 KEGG pathways. In addition, we created a statistical data of unigenes involved in the second hierarchy pathways without human diseases category, and the categories showing the greatest representation by unigenes were metabolism (3,466 unigenes) and organismal systems (1,235 unigenes) (Figure 4). Among all the related secondary metabolism pathways, biosyntheses of anthocyanin (2 unigenes), flavone and flavonol (10 unigenes), flavonoid (32 unigenes), and

carotenoid (28 unigenes) were associated with color formation and variation.

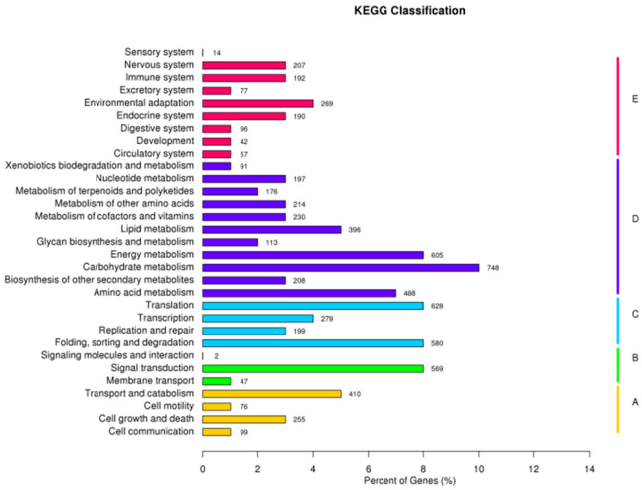

Figure 4. KEGG Classifications of *J. regia* unigenes

A: Cellular Processes; B: Environmental Information Processing; C: Genetic Information Processing; D: Metabolism; E: Organismal Systems.

3.3 SSR Discovery in The J. regia Transcriptome Assembly

SSRs are short tandem repeat sequences (1–6 bp) and are distributed throughout the genome. For the past 20 years, SSRs have been widely applied for plant genetic analysis because of their highly polymorphic characterization. From the 53,782 unigenes that were examined in *J. regia* transcriptome assembly, 11,315 unigenes containing putative SSRs were identified with the occurrence frequency of 21.04%, among which 2,148 unigenes had more than one SSR marker signature. A total of 13,981 SSRs were discovered, of which 815 SSRs were of compound formation (Table 3). Repeat motifs among these SSRs consisted of 69 types, and the number of repeat unit size varied from 5 to 24. Table 4 shows the density of different types of SSRs. The single base repeats were the most abundant motifs (49.29%), followed by dibase type SSRs (37.42%). In recent years, several SSR markers have been developed by screening expressed sequencing tags (ESTs). Thirty-five and 30 highly transferable and polymorphic expressed sequencing tags-simple sequence repeats (EST-SSRs) and 706 EST-SSRs were identified in *J. regia* by Qi *et al*. (2010), Zhang *et al*. (2010), and Zhang *et al*. (2013), respectively (Qi JX, 2011; Zhang R, 2010; Zhang et al., 2013). Therefore, compared with the previous results, our study obtained more SSRs. Moreover, the new SSRs would be further identified by primer designing and PCR reactions.

Table 3. Summary of SSR Analysis in the *J. regia* Transcriptome

Total number of sequences examined	**53782**
Total size of examined sequences (bp)	42966265
Total number of identified SSRs	13981
Number of SSR containing sequences	11315
Number of sequences containing more than 1 SSR	2148
Number of SSRs present in compound formation	815

Table 4. Frequency of different SSR types in the *J. regia* transcriptome

SSRs type	**Repeat Number**									**Total**
	5	6	7	8	9	10	11	>11, ≤20	>20	
P1	-	-	-	-	-	2037	1241	3419	194	6891
P2	-	1254	924	950	1150	792	154	7	-	5231
P3	881	457	314	20	-	-	-	3	1	1676
P4	118	29	2	1	-	-	-	1	-	151
P5	20	-	-	-	-	-	-	1	-	21
P6	5	4	-	-	1	-	-	1	-	11
Total	1024	1744	1240	971	1151	2829	1395	3432	195	13981

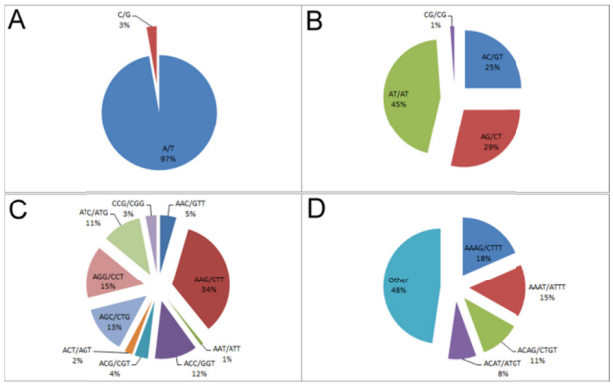

Figure 5. Percentages of Different Motifs among Single-nucleotide (A), Dinucleotide (B), Tribucleotide (C), and Tetranucleotide (D) in the *J. regia* Transcriptome

3.4 Identification of SNP and InDel Polymorphisms

Although SSR exploration has been conducted in *J. regia*, SSR frequency was lower than SNP (Kaya et al., 2013). A total of 41,088 SNPs and 5,860 insertion/deletion polymorphisms (InDels) were detected. Of the 5,860 InDels, 4616 were insertions and 1,244 were deletions. Of the 41,088 SNPs, 8,424 were transitions and 32,652 were transversions (with a ratio of approximately 1:4).

3.5 Mining of Genes are Putatively Related to Leaf Red-pigment Trait

Generally, chlorophyll makes leaf green, and flavonoids, flavone and flavonol, anthocyanin, betalain, and carotenoid biosynthesis are related to red color pigment (Tanaka et al., 2008). In *J. regia* transcriptome, nearly all genes involved in chlorophyll biosynthesis were found. The flavonoid compounds were the major pigments responsible for colors of flower and non-green leaves, and the ratio of different anthocyanidins resulted in different colors. To date, we are not sure which pigments were involved in red mutation because of the lack of pigment content determination in *J. regia* leaves, which may be the focus of further study. Therefore, we analyzed related genes in the biosynthesis of the pigments mentioned above. From *J. regia* transcriptome, genes involved in these pathways were analyzed and searched by standard gene names and synonyms (Additional File 2).

A total of 74 unigenes (including 50 genes) were assigned to the biosynthetic pathways above by mapping to KEGG pathways. Anthocyanins, a class of flavonoids, are synthesized in the cytosol and stored in the vacuole. The colors of anthocyanins ranged from red to blue depending on pH, metal ions, and co-pigments (typically flavones and flavonols). In a previous study by Zhao *et al*., 17 compounds isolated and identified from the leaves of *J. regia* were collected from Tianjin, China (Zhao et al., 2014). Among these unigenes, all genes except for four annotated in flavonoid biosynthetic pathway were found (Additional file 3A). Additionally, in the anthocyanin biosynthesis of the downstream, all genes were identified

except for the last gene encoding UDP glucose: flavonoid 3-Oglucosyltransferase (3B). Ten genes were in the flavone and flavonol pathways (3C). Carotenoids are terpenoid compounds with colors ranging from yellow to red and are synthesized in chloroplasts. All genes involved in carotene biosynthesis and most genes involved in lutein and astaxanthin (carotenoid) biosynthesis were identified (Additional File 3D). Betalains are compounds responsible for violet and yellow color in plants, but the biosynthetic pathway has not been elucidated well. From the *J. regia* transcriptome, we found the key enzyme annotated as '4,5-DOPA extradiol dioxygenase' (Additional File 2E). Unlike anthocyanin, betalains are more stable, and the color does not depend on the pH (Tanaka et al., 2008). We speculate that flavonoids and carotenoid may be responsible for the red color of *J. regia* leaves.

3.6 Identification of TFs

In previous studies, a set of TFs, including R2R3-MYB, basic helix-loop-helix (bHLH), and WD40-repeat protein, were identified as key regulators for transcription expression in pigment biosynthetic pathways (Grotewold, 2006; Jung et al., 2009; Yang et al., 2015; Zhu et al., 2014). A total of 656 unigenes annotated as TFs were found (Additional File 3), and the TF families with the largest number of unigenes were WRKY (64 unigenes), NAC (50 unigenes), bHLH (50 unigenes), MYB (49 unigenes), and heat shock factors (45 unigenes) (Table 5). TFs from MYB family were the most widespread regulators of anthocyanin biosynthesis. Tian *et al*. demonstrated that *McMYB10* can positively regulate anthocyanin biosynthesis via flavonoid 3'-hydroxylase in ever-red leaf crabapple (Tian et al., 2015). In 2015, Albert *et al*. isolated two R2R3-MYB genes, *RED LEAF* and *RED V*, which were confirmed to be anthocyanin regulators (Albert et al., 2015). Zhou *et al*. found that overexpression of *PpMYB10.4* induces anthocyanin accumulation in tobacco and peach leaves (Zhou et al., 2014). Other evidences support that carotenoids biosynthesis were mainly regulated at the transcriptional level (Sandmann et al., 2006). In tomato, *SlNAC4* functioned as a positive regulator of carotenoid accumulation (Zhu et al., 2014). Moreover, several unigenes encoding phytochrome-interacting factors, a class of bHLH TFs that can increase carotenoids accumulation (Rodriguez-Villalon et al., 2009), were also identified in our *J. regia* transcriptome. The identified TFs potentially involved in pigment biosynthesis regulation can be explored deeply in the future.

Table 5. Statistics of transcription factors in red-leaf *J. regia*

Transcription factors	Number
WRKY	64
bHLH	50
NAC	50

MYB	49
Heat shock	45
GATA	10
MADS-box	10
AP2/ERF	23
WD-repeat	21
GRAS	15
AP2 domain containing	15
R2R3-MYB	13
HD-domain	6

3.7 Analysis of Gene Expression Level

The *J. regia* transcriptome assembled by Trinity was used for analyzing reference sequences, and of all 33,488,602 clean reads, 28,493,056 (85.08%) were mapped to the reference transcriptome. Mapping was performed through RSEM software with two mismatches as the bowtie parameter (Li and Dewey, 2011). RPKM is the most common method to estimate gene expression level. Furthermore, all the read count numbers of each sample mapping to each gene were normalized to RPKM to calculate gene RPKM values. The 49,340 unigenes had an estimated RPKM value of over 0.3. In particular, 26 unigenes had RPKM values of over 1000, the largest group comprised 28631 unigenes with RPKM values between 1 and 100, followed by the group with RPKM values between 0.3 and 1 (Table 6). Additionally, RPKM density distribution of *J. regia* is shown in Figure 6. For the candidate genes involved in pigment biosynthesis, some key enzymes in anthocyanin biosynthesis, such as flavonol 3-O-glucosyltransferase, flavonol 3-O-methyltransferase, flavonoid 3',5'-hydroxylase, flavonol synthase, and chalcone synthase, had high RPKM values (up to 300 to 700). Most genes in the carotenoid pathway had RPKM values between 20 and 100. Only one gene, annotated as 4,5-DOPA dioxygenase for betalain accumulation, had an RPKM value of 36.89, thereby suggesting that all these three pathways were highly expressed in *J. regia* red leaves (Additional File 2). Chlorophyll biosynthesis interruption was not the cause of red-leaf trait, but rather the high ratio of anthocyanin, carotenoid, and betalain biosyntheses.

Table 6. The read counts_RPKM of *J. regia* genes

RPKM value	Number
<0.3	4442 (8.26%)
≥0.3 and <1	19764 (36.75%)
≥1 and <100	28631 (53.23%)
≥100 and <1000	919 (1.7%)
≥1000	26 (0.07%)
Total	53782

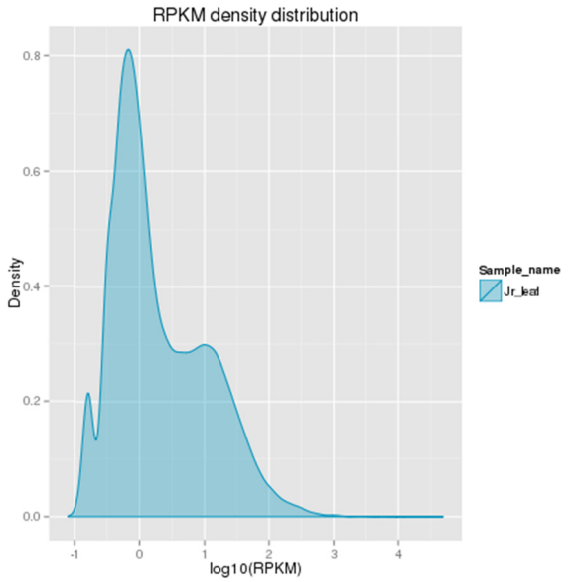

Figure 6. The RPKM density distribution of *J. regia* transcriptome

4. Conclusions

High throughput SSRs, SNPs, and InDel discoveries have been developed with the advent of NGS technology. NGS is an efficient approach for marker discovery and gene mining and for determination of the metabolism pathways involved, especially in species without a reference genome (Strickler et al., 2012). The current study generated 53,782 unigenes, 13,981 SSRs, 41,088 SNPs, and 5,860 InDels. Marker-assisted selection has played an important role in accelerating agriculture breeding. Furthermore, based on the transcriptome data, five pigments related to biosynthetic pathways were identified, and most genes involved in

anthocyanin and carotenoid biosynthesis were identified, thereby showing that anthocyanins and carotenoids may play essential roles in the development of the red-leaf trait in *J. regia*. The pigments of plants are controlled by both structural genes and regulatory genes. Some candidate transcription factors were collected. The current study is the first comprehensive report on the large scale transcriptome sequencing analysis of *J. regia* for gene annotation and molecular marker discovery. Considering the reports on the resistance to insects and pathogen of red-leaf, this study provides resources for red-leaf trees breeding.

Acknowledge

This work was supported by Taian Good Cultivar Engineering Project-Fruit Germplasm Resources Collection and New Germplasm Innovation program（2015-09）to XC，Shandong Good Cultivar Engineering Project-Fruit Germplasm Resources Collection and New Germplasm Innovation program（2014-96）to QS and National Science & Technology Infrastructure Program(2015-048) of QL.

Abbreviation

SSR: simple sequence repeats

SNPs: single nucleotide polymorphism

Bp: base pair

bHLH: basic helix-loop-helix

ESTs: expressed sequencing tags

GO: gene ontology

InDels: insertions and deletions

KEGG: Kyoto Encyclopedia of Genes and Genomes

KO: KEGG Ortholog

KOG: Clusters of Orthologous Groups of proteins

Nr: NCBI non-redundant protein sequences

Nt: NCBI non-redundant nucleotide sequences

Pfam: Protein family

PIFs: phytochrome-interacting factors

RPKM: Reads Per Kilo bases per Million mapped Reads

TFs: Transcription Factors

4,5-DOPA dioxygenase: 4,5-dihydroxyphenylalanine dioxygenase

References

Abugri, D. A., Tiimob, B. J., Apalangya, V. A., Pritchett, G., & McElhenney, W. H. (2013). Bioactive and nutritive compounds in Sorghum bicolor (Guinea corn) red leaves and their health implication. *Food Chem*, *138*, 718-723. https://doi.org/10.1016/j.foodchem.2012.09.149

Albert, N. W., Griffiths, A. G., Cousins, G. R., Verry, I. M., &d Williams, W. M. (2015). Anthocyanin leaf markings are regulated by a family of R2R3-MYB genes in the genus Trifolium. *The New phytologist, 205*, 882-893. https://doi.org/10.1111/nph.13100

Amaral, J. S., Valentao, P., Andrade, P. B., Martins, R. C., & Seabra, R. M. (2008). Do cultivar, geographical location and crop season influence phenolic profile of walnut leaves? *Molecules (Basel, Switzerland)*, *13*, 1321-1332. https://doi.org/10.3390/molecules13061321

Cai, C., Zhang, X., Niu, E., Zhao, L., Li, N., Wang, L., Ding, L., & Guo, W. (2014). GhPSY, a phytoene synthase gene, is related to the red plant phenotype in upland cotton (*Gossypium hirsutum* L.). *Molecular biology reports, 41*, 4941-4952. https://doi.org/10.1007/s11033-014-3360-x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Zeng, Q. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology, 29*, 644-652. https://doi.org/10.1038/nbt.1883

Grotewold, E. (2006). The genetics and biochemistry of floral pigments. A*nnual review of plant biology, 57*, 761-780. https://doi.org/10.1146/annurev.arplant.57.032905.105248

Hosseini, S., Huseini, H. F., Larijani, B., Mohammad, K., Najmizadeh, A., Nourijelyani, K., & Jamshidi, L. (2014a). The hypoglycemic effect of Juglans regia leaves aqueous extract in diabetic patients: A first human trial. *Daru: journal of Faculty of Pharmacy, Tehran University of Medical Sciences, 22*, 19. https://doi.org/10.1186/2008-2231-22-19

Hosseini, S., Jamshidi, L., Mehrzadi, S., Mohammad, K., Najmizadeh, A. R., Alimoradi, H., & Huseini, H. F. (2014b). Effects of Juglans regia L. leaf extract on hyperglycemia and lipid profiles in type two diabetic patients: a randomized double-blind, placebo-controlled clinical trial. *Journal of ethnopharmacology*, *152*, 451-456. https://doi.org/10.1016/j.jep.2014.01.012

Jung, C. S., Griffiths, H. M., De Jong, D. M., Cheng, S., Bodis, M., Kim, T. S., & De Jong, W.S. (2009). The potato developer (D) locus encodes an R2R3 MYB transcription factor that regulates expression of multiple anthocyanin structural genes in tuber skin. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik*, *120*, 45-57. https://doi.org/10.1007/s00122-009-1158-3

Kaya, H. B., Cetin, O., Kaya, H., Sahin, M., Sefer, F., Kahraman, A., & Tanyolac, B. (2013). SNP discovery by illumina-based transcriptome sequencing of the olive and the genetic characterization of Turkish olive genotypes revealed by AFLP, SSR and SNP markers. *PloS one*, *8*, e73674. https://doi.org/10.1371/journal.pone.0073674

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data

with or without a reference genome. *BMC bioinformatics, 12*, 323. https://doi.org/10.1186/1471-2105-12-323

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research, 20*, 1297-1303. https://doi.org/10.1101/gr.107524.110

Mouradov, A., & Spangenberg, G. (2014). Flavonoids: a metabolic network mediating plants adaptation to their real estate. *Frontiers in plant science, 5*, 620. https://doi.org/10.3389/fpls.2014.00620

Nael Abu Taha, M. A. A. -w. (2011). Utility and importance of walnut, Juglans regia Linn: A review. *Afr J Microbiol Res, 5*, 10. https://doi.org/10.5897/AJMR11.610

Paudel, P., Satyal, P., Dosoky, N. S., Maharjan, S., & Setzer, W. N. (2013). Juglans regia and J. nigra, two trees important in traditional medicine: A comparison of leaf essential oil compositions and biological activities. *Natural product communications, 8*, 1481-1486.

Pereira, J. A., Oliveira, I., Sousa, A., Valentao, P., Andrade, P. B., Ferreira, I. C., ... Estevinho, L. (2007). Walnut (*Juglans regia* L.) leaves: phenolic compounds, antibacterial activity and antioxidant potential of different cultivars. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association, 45*, 2287-2295.

Qi, J. X., H.Y., Zhu, Y, & Wu, C. L. (2011). Studies on germplasm of Juglans by EST-SSR markers. *Acta Hortic Sin, 38*, 8.

Qureshi, M. N., Stecher, G., & Bonn, G. K. (2014). Determination of total polyphenolic compounds and flavonoids in *Juglans regia* leaves. *Pakistan journal of pharmaceutical sciences, 27*, 865-869.

Ranjan, S., Singh, R., Singh, M., Pathre, U. V., & Shirke, P. A. (2014). Characterizing photoinhibition and photosynthesis in juvenile-red versus mature-green leaves of Jatropha curcas L. *Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale, 79*, 48-59. https://doi.org/10.1016/j.plaphy.2014.03.007

Rodriguez-Villalon, A., Gas, E., & Rodriguez-Concepcion, M. (2009). Phytoene synthase activity controls the biosynthesis of carotenoids and the supply of their metabolic precursors in dark-grown Arabidopsis seedlings. *The Plant journal : for cell and molecular biology, 60*, 424-435. https://doi.org/10.1111/j.1365-313X.2009.03966.x

Sandmann, G., Romer, S., & Fraser, P. D. (2006). Understanding carotenoid metabolism as a necessity for genetic engineering of crop plants. *Metabolic engineering, 8*, 291-302. https://doi.org/10.1016/j.ymben.2006.01.005

Strickler, S. R., Bombarely, A., & Mueller, L. A. (2012). Designing a transcriptome next-generation sequencing project for a nonmodel plant species. *American journal of botany, 99*, 257-266. https://doi.org/10.3732/ajb.1100292

Tanaka, Y., Sasaki, N., & Ohmiya, A. (2008). Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *The Plant journal : for cell and molecular biology, 54*, 733-749. https://doi.org/10.1111/j.1365-313X.2008.03447.x

Tian, J., Peng, Z., Zhang, J., Song, T., Wan, H., Zhang, M., & Yao, Y. (2015). McMYB10 regulates coloration via activating McF3'H and later structural genes in ever-red leaf crabapple. *Plant biotechnology journal*. https://doi.org/10.1111/pbi.12331

Wang, Y. S., Liu, Z. Y., Li, Y. F., Zhang, Y., Yang, X. F., & Feng, H. (2013). Identification of sequence-related amplified polymorphism markers linked to the red leaf trait in ornamental kale (*Brassica oleracea L.* var. acephala). *Genetics and molecular research : GMR, 12*, 870-877. https://doi.org/10.4238/2013.April.2.3

Wu, J., Gu, Y. Q., Hu, Y., You, F. M., Dandekar, A. M., Leslie, C. A.,... Luo, M. C. (2012). Characterizing the walnut genome through analyses of BAC end sequences. *Plant molecular biology, 78*, 95-107. https://doi.org/10.1007/s11103-011-9849-y

Yang, Y. N., Yao, G. F., Zheng, D., Zhang, S. L., Wang, C., Zhang, M. Y., & Wu, J. (2015). Expression differences of anthocyanin biosynthesis genes reveal regulation patterns for red pear coloration. *Plant cell reports, 34*, 189-198. https://doi.org/10.1007/s00299-014-1698-0

Zhang R, Z.A., & Wang, X. J, Yu, J. (2010). Development of Juglans regia SSR markers by data mining of the EST database. *Plant Mol Biol Rep, 28*, 9. https://doi.org/10.1007/s11105-010-0192-2

Zhang, Z. Y., Han, J. W., Jin, Q., Wang, Y., Pang, X. M., & Li, Y. Y. (2013). Development and characterization of new microsatellites for walnut (*Juglans regia*). *Genetics and molecular research : GMR*, *12*, 4723-4734. https://doi.org/10.4238/2013.October.18.10

Zhao, M. H., Jiang, Z. T. L. (2014). Flavonoids in Juglans regia L. leaves and evaluation of in vitro antioxidant activity via intracellular and chemical methods. *The Scientific World Journal, 3*03878.

Zhou, Y., Zhou, H., Lin-Wang, K., Vimolmangkang, S., Espley, R. V., Wang, L., Allan, A C., & Han, Y. (2014). Transcriptome analysis and transient transformation suggest an ancient duplicated MYB transcription factor as a candidate gene for leaf red coloration in peach. *BMC plant biology, 14*, 388. https://doi.org/10.1186/s12870-014-0388-y

Zhu, M., Chen, G., Zhou, S., Tu, Y., Wang, Y., Dong, T., & Hu, Z. (2014). A new tomato NAC (NAM/ATAF1/2/CUC2) transcription factor, SlNAC4, functions as a positive regulator of fruit ripening and carotenoid accumulation. *Plant & cell physiology, 55*, 119-135. https://doi.org/10.1093/pcp/pct162

Additional File 1. The lengths distributions of transcripts and unigenes

Transcript length interval	200-500bp	500-1kbp	1k-2kbp	>2kbp	-
Total Number of transcripts	38000	17954	24702	19949	100605
Total Number of Unigenes	31833	8950	7600	5399	53782

Additional File 2. Candidate genes related to the pigmentation of *Juglans regia* red-leaf

Function	Unigene No.	KO	Enzyme	Pathway ID	RPKM
Porphyrin and chlorophyll metabolism	comp14200_c0	K02257	protoheme IX farnesyltransferase	ko00860	22.50
	comp20631_c0	K01749	hydroxymethylbilane synthase		50.42
	comp10637_c0	K02259	cytochrome c oxidase assembly protein subunit 15		11.98
	comp29312_c0	K01845	glutamate-1-semialdehyde 2,1-aminomutase		88.33
	comp25659_c0	K02492	glutamyl-tRNA reductase		134.71
	comp9774_c0	K02492	glutamyl-tRNA reductase		15.85
	comp10356_c0	K02492	glutamyl-tRNA reductase		51.31
	comp32259_c0	K02492	glutamyl-tRNA reductase		21.01
	comp10356_c1	K02492	glutamyl-tRNA reductase		59.13
	comp18909_c0	K00522	ferritin heavy chain		17.34
	comp18909_c1	K00522	ferritin heavy chain		8.11
	comp20893_c0	K00522	ferritin heavy chain		58.53

comp17081_c0	K00522	ferritin heavy chain	72.45
comp21097_c0	K04040	chlorophyll synthase	39.61
comp125945_c0	K04040	chlorophyll synthase	0.56
comp23862_c0	K00228	coproporphyrinogen III oxidase	65.07
comp1031_c0	K03403	magnesium chelatase subunit H	-
comp5337_c0	K03403	magnesium chelatase subunit H	-
comp170589_c0	K03403	magnesium chelatase subunit H	0.44
comp28686_c0	K03403	magnesium chelatase subunit H	302.70
comp17992_c0	K01698	porphobilinogen synthase	70.93
comp11188_c0	K10960	geranylgeranyl reductase	9.43
comp28845_c0	K10960	geranylgeranyl reductase	201.57
comp25852_c1	K08101	phytochromobilin:ferredoxin oxidoreductase	8.75
comp26939_c0	K13600	chlorophyllide a oxygenase	69.39
comp11746_c0	K13600	chlorophyllide a oxygenase	38.42
comp25409_c0	K13606	chlorophyll(ide) b reductase	20.23
comp27579_c0	K01772	ferrochelatase	34.19
comp16739_c0	K01772	ferrochelatase	25.51
comp26150_c0	K01719	uroporphyrinogen-III synthase	14.14
comp30546_c0	K01599	uroporphyrinogen decarboxylase	9.26
comp24500_c0	K01599	uroporphyrinogen decarboxylase	89.53
comp215367_c0	K01599	uroporphyrinogen decarboxylase	-

comp26382_c1	K08099	chlorophyllase	24.21
comp19172_c0	K08099	chlorophyllase	10.94
comp142184_c0	K13071	pheophorbide a oxygenase	0.66
comp22718_c0	K13071	pheophorbide a oxygenase	2.88
comp18526_c1	K13071	pheophorbide a oxygenase	36.04
comp28700_c0	K04035	magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase	402.05
comp107069_c0	K02495	oxygen-independent coproporphyrinogen III oxidase	1.38
comp26649_c1	K00231	oxygen-dependent protoporphyrinogen oxidase	8.23
comp27094_c0	K00231	oxygen-dependent protoporphyrinogen oxidase	28.51
comp33662_c0	K13545	red chlorophyll catabolite reductase	19.23
comp32040_c0	K13545	red chlorophyll catabolite reductase	28.03
comp10014_c0	K03428	magnesium-protoporphyrin O-methyltransferase	78.87
comp18749_c0	K01885	glutamyl-tRNA synthetase	25.20
comp10705_c0	K01885	glutamyl-tRNA synthetase	34.89
comp26677_c0	K03405	magnesium chelatase subunit I	70.86
comp22973_c0	K00510	heme oxygenase	28.21
comp9311_c0	K00218	protochlorophyllide reductase	5.93
comp18794_c0	K00218	protochlorophyllide reductase	195.54
comp14630_c0	K00218	protochlorophyllide reductase	5.15

	comp14630_c2	K00218	protochlorophyllide reductase		11.28
	comp14630_c1	K00218	protochlorophyllide reductase		15.98
Carotenoid biosynthesis	comp26309_c1	K15746	beta-carotene 3-hydroxylase	ko00906	82.90
	comp31739_c0	K15746	beta-carotene 3-hydroxylase		12.70
	comp10365_c0	K15747	cytochrome P450, family 97, subfamily A (beta-ring hydroxylase)		44.35
	comp25896_c0	K15744	zeta-carotene isomerase		9.58
	comp9456_c0	K09843	(+)-abscisic acid 8'-hydroxylase		4.74
	comp1748_c0	K09843	(+)-abscisic acid 8'-hydroxylase		1.21
	comp25736_c0	K09843	(+)-abscisic acid 8'-hydroxylase		17.98
	comp13144_c0	K09843	(+)-abscisic acid 8'-hydroxylase		1.15
	comp25328_c1	K09843	(+)-abscisic acid 8'-hydroxylase		47.66
	comp27108_c2	K09843	(+)-abscisic acid 8'-hydroxylase		1.25
	comp80215_c0	K09843	(+)-abscisic acid 8'-hydroxylase		25.02
	comp13725_c1	K09843	(+)-abscisic acid 8'-hydroxylase		25.02
	comp25736_c0	K09843	(+)-abscisic acid 8'-hydroxylase		17.98
	comp13144_c0	K09843	(+)-abscisic acid 8'-hydroxylase		1.15
	comp25328_c1	K09843	(+)-abscisic acid 8'-hydroxylase		6.86
	comp27108_c2	K09843	(+)-abscisic acid 8'-hydroxylase		1.25
	comp80215_c0	K09843	(+)-abscisic acid 8'-hydroxylase		-
	comp13725_c1	K09843	(+)-abscisic acid 8'-hydroxylase		25.02
	comp10689_c0	K02291	phytoene synthase		38.17

	comp12167_c0	K02291	phytoene synthase		51.62
	comp14642_c0	K02293	15-cis-phytoene desaturase		18.86
	comp18122_c0	K06444	lycopene epsilon-cyclase		26.42
	comp26813_c0	K09835	prolycopene isomerase		8.56
	comp14129_c0	K09840	9-cis-epoxycarotenoid dioxygenase		11.52
	comp19276_c0	K09840	9-cis-epoxycarotenoid dioxygenase		7.85
	comp22533_c0	K09840	9-cis-epoxycarotenoid dioxygenase		56.00
	comp11553_c1	K06443	lycopene beta-cyclase		20.77
	comp62128_c0	K06443	lycopene beta-cyclase		2.29
	comp16609_c0	K09838	zeaxanthin epoxidase		100.10
	comp30468_c0	K09839	violaxanthin de-epoxidase		25.23
	comp18640_c0	K00514	zeta-carotene desaturase		23.06
	comp16623_c0	K09841	xanthoxin dehydrogenase		68.97
	comp21350_c0	K09841	xanthoxin dehydrogenase		41.89
	comp27445_c0	K09841	xanthoxin dehydrogenase		23.01
Flavonoid biosynthesis	comp87855_c0	K13065	shikimate O-hydroxycinnamoyltransferase	ko00941	-
	comp16724_c0		shikimate O-hydroxycinnamoyltransferase		8.95
	comp26179_c0		shikimate O-hydroxycinnamoyltransferase		32.82
	comp118258_c0		shikimate O-hydroxycinnamoyltransferase		0.85

comp182423_c0		shikimate O-hydroxycinnamoyltransferase	-
comp75175_c0		shikimate O-hydroxycinnamoyltransferase	0.72
comp16079_c0	K08695	anthocyanidin reductase	55.55
comp10118_c0	K13083	cytochrome P450, family 75, subfamily A (flavonoid 3',5'-hydroxylase)	235.43
comp29670_c0	K13082	bifunctional dihydroflavonol 4-reductase/flavanone 4-reductase	70.27
comp21359_c0	K13081	leucoanthocyanidin reductase	5.68
comp23640_c0		leucoanthocyanidin reductase	11.72
comp46728_c0		leucoanthocyanidin reductase	5.55
comp22978_c0		leucoanthocyanidin reductase	26.07
comp29197_c0	K01859	chalcone isomerase	94.91
comp28746_c0	K00475	naringenin 3-dioxygenase	346.57
comp20952_c0	K05277	leucoanthocyanidin dioxygenase	306.62
comp25022_c1	K00660	chalcone synthase	767.76
comp10942_c0		chalcone synthase	226.16
comp93311_c0		chalcone synthase	1.22
comp41264_c0		chalcone synthase	7.06
comp13594_c0	K09754	coumaroylquinate(coumaroylshikimate) 3'-monooxygenase	2.09
comp27087_c0		coumaroylquinate(coumaroylshikimate) 3'-monooxygenase	102.36

	comp41917_c0		coumaroylquinate(coumaroylshikimate) 3'-monooxygenase		4.84
	comp13594_c1		coumaroylquinate(coumaroylshikimate) 3'-monooxygenase		4.03
	comp10217_c0	K00588	caffeoyl-CoA O-methyltransferase		24.06
	comp15919_c0		caffeoyl-CoA O-methyltransferase		147.39
	comp28828_c0	K05278	flavonol synthase/flavanone 3-hydroxylase		236.03
	comp26379_c1		flavonol synthase/flavanone 3-hydroxylase		36.68
	comp271762_c0		flavonol synthase/flavanone 3-hydroxylase		0.31
	comp26299_c0	K00487	trans-cinnamate 4-monooxygenase		293.82
	comp16111_c0		trans-cinnamate 4-monooxygenase		3.69
	comp20629_c0	K05280	flavonoid 3'-monooxygenase		77.20
Anthocyanin biosynthesis	comp13565_c0	K12338	anthocyanin 5-O-glucosyltransferase	ko00942	4.29
	comp18527_c2	K12930	anthocyanidin 3-O-glucosyltransferase		0.36
Flavone and flavonol biosynthesis	comp25776_c1	K10757	flavonol 3-O-glucosyltransferase	ko00944	476.94
	comp23883_c0	K05279	flavonol 3-O-methyltransferase		52.08
	comp27385_c0	K05279	flavonol 3-O-methyltransferase		287.98
	comp27708_c1	K05279	flavonol 3-O-methyltransferase		34.95
	comp20779_c0	K05279	flavonol 3-O-methyltransferase		314.48
	comp27385_c1	K05279	flavonol 3-O-methyltransferase		189.86
	comp33903_c0	K05279	flavonol 3-O-methyltransferase		19.37

	comp75470_c0	K05279	flavonol 3-O-methyltransferase		-
	comp10118_c0	K13083	cytochrome P450, family 75, subfamily A (flavonoid 3',5'-hydroxylase)		235.43
	comp20629_c0	K05280	flavonoid 3'-monooxygenase		77.20
Betalain biosynthesis	comp17877_c0	K15777	4,5-DOPA dioxygenase extradiol	ko00965	36.89
	comp49680_c0	K15777	4,5-DOPA dioxygenase extradiol		3.76