Swift Journals

Swift Journal of Agricultural Research
Vol 4(1) pp. 001-020 October, 2018
http://www.swiftjournals.org/sjar
ISSN: 2986-982X

Original Research Article

De Novo Transcriptome Assembly and Comparison of Walnut (Juglans regia L.) cv. Qingxiang Organs: Roots, Stems and Leaves

***Benzhong Fu[1,2], Lulu Zou[1]**

[1]College of Life Science and Technology, Hubei Engineering University.
[2]Hubei Key Laboratory of Quality Control of Characteristic Fruits and Vegetables
Address:272# Jiaotong Avenue, Xiaogan city,Hubei province, 432000, China.

Accepted 8th September, 2018.

ABSTRACT

Due to its medicinal benefits to human health, such as improving the cardiovascular system, reduce problems in metabolic syndrome and anti-cancer, Persian walnut (Juglans regia L.) has become one of most important stone fruits all around the world. China is presently the largest walnuts producer in the global. However, there are still few genomic and transcriptomic data of commercial walnut cultivars, especially about their adult organs and tissues, including the roots, stems and leaves. Also, there are no data that is available about walnut cultivar Qingxiang, which is one of the major cultivars that are being planted in China. In this study, Illumina RNA sequencing (RNA-Seq) was performed on samples collected from J. regia cultivar Qingxiang roots, stems and leaves. The data obtained were used for de novo assembly and the characterization of the transcriptome of three organs. In total, 85,232,619 paired-end reads were generated with 17,215,394,747 nucleotides and GC percentage was 46.18%. These were assembled into 4,336,022 contigs with 94,500 transcripts and 51,807 unigenes. The unigenes were annotated by querying against the NCBI non-redundant and UniProt databases, 28,209 unigenes (54.45%) were homologous to the existing database sequences. Gene Ontology terms were doled out to 23,451 unigenes, 8,292 loci were coordinated to 25 Clusters of Eukaryotic Orthologous Groups arrangements, and 5,885 unigenes were characterized into 116 Kyoto Encyclopedia of Genes and Genomes pathways. The gathered unigenes（more than 1 kb）also contained 6,244 potential straightforward grouping repeats. The put together transcriptome was utilized to distinguish unigenes with organ-particular differential articulation designs. In total, 942 unigenes exhibited organ-specific expression. Our data analyses implied that walnut different organs expressed specific genes in many molecular processes and functions, particular, the plant-pathogen interaction pathways were demonstrated. The existing genomic transcriptomic data from China major walnut cultivars are scarce. The walnut organ-based transcriptome resources developed in this study will enable the analysis of organ development, and accelerate marker-assisted breeding based on SSRs and understanding the disease resistance in walnut. Furthermore, this study provides an enhanced insight into the organ response of probably pathogen infection.

Keywords: Walnut; organ; transcriptome; Illumina; RNA-seq; unigene.

**Corresponding Author:* Benzhong Fu *, [1]College of Life Science and Technology, Hubei Engineering University, [2]Hubei Key Laboratory of Quality Control of Characteristic Fruits and Vegetables. Address:272# Jiaotong Avenue, Xiaogan city,Hubei province, 432000, China.
Email: benzhongf@yahoo.com
Tel: +086-13886387573

INTRODUCTION

Persian walnut (or English walnut, Juglans regia L.) belongs to the family Juglandaceae, order Fagales, and is a diploid species (2n=32). It is an economically important stone fruit that are being cultivated extensively for its nuts or timber production globally1. They are distributed in temperate regions of Asia, Europe, North and South America. Meanwhile, these different planting areas developed abundant of cultivars. In recent years, it is globally popular and valued for its nutritional and health-promoting attributes2. Walnut kernels serve as a valuable source of dietary oil rich in omega-3 fatty acids, and play antioxidant and radical scavenging activities3-5.

China is the world's largest producer and consumer of walnuts6. In 2016, China's walnut production reached 251.37 million tons, accounting for 50% of total global walnut production. Due to the government strong inciting policy, China has been the largest walnuts commercially produced in the world, with harvest area 487,007 ha and 1,785,879 tons production (http://www.fao.org/faostat/en/#data/QC). The United States and European countries are ranked the next largest producer of walnut. Among of the various cultivars planted in China commercially, Qingxiang is a major cultivar planted in Xinjiang, Shandong, Hebei and Hubei provinces, due to its early fruiting and high production and disease resistance quality.

The next generation sequencing innovation has empowered quality disclosure, examination of gene content, and correlation of gene articulation in genomic level, and applied in numerous natural fruit trees, including stone fruit. Transcriptome examination is basic to decipher the useful components of the genome and uncover the sub-atomic constituents and pathways of tissues and organs. RNA-Seq investigation has encouraged transcriptome portrayal in many plant species lacking sequenced genomes7-9. Transcriptome sequencing can be utilized for all-inclusive assurance of supreme transcript levels, recognizable proof of transcripts, and depiction of transcript structure. Transcriptome sequencing can also identify genetic variations such as, simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs)10. Four tropical tree species, three organs (leaves, stems, and roots), from the Amazonian rainforest were analyzed using 454 pyrosequencing for de novo assembly, gene functional annotation and marker discovery11.

The transcriptomic method has also been used to study the gene expression changes in different cultivars and organs12-14. For example, in wild tomato species Solanum pimpinellifolium, using high-throughout RNA sequencing to analyze the transcriptome of ovaries and fruit tissues15. In maize and corn smut (Ustilago maydis) host-pathogen interactions, organ-specific gene expression in plant tumors development was documented by transcriptome profiling16. The transcriptome of Arabidopsis thaliana leaves, hypocotyls, and roots displayed organ-specific changes in response to spaceflight, the differential expression of genes involved with touch, cell wall remodeling, root hairs, and cell expansion17. In Arabidopsis, an organ-specific regulation was assigned to a number of genes reacting to a cytokinin signal, cytokinin-regulated transcriptome showed different response patterns in roots and shoots18.Furthermore, many new methods were developed, such as Single-molecule FISH (smFISH), RNA sequences probing of targets (SPOTs)19, high-resolution, spatially transcriptomics and functional profiling in Arabidopsis thaliana20.

The availability of genomic data for molecular breeding could possibly lead to more rapid genetic improvement of walnut, particularly when combining traits for environmental adaptation with other important traits like disease resistance and nutritional quality.

Substantial research efforts have been made to study the walnut genetics, genomics and transcriptomics21-25. Such as, J. regia cv. Chandler bacterial artificial chromosome (BAC) libraries26, SNP discovery in walnut (J. regia L.)27, walnut bacterial artificial chromosome (BAC) clone-based physical map28. Most recently, a J. regia genome sequence was obtained from the cultivar ‘Chandler’ to discover target genes and additional unknown genes. The 667-Mbp genome was assembled using SOAPdenovo2 and MaSuRCA, with an N50 scaffold size of 464 955 bp (based on a genome size of 606 Mbp), 221 640 contigs and a GC content of 37%. Annotation with MAKER-P and other genomic resources yielded 32 498 gene models29. Utilizing Illumina digital quality articulation profiling, distinguished 4568 differentially expressed genes (DEGs) (DEGs) amongst red and green walnut leaf and 3038 DEGs amongst red and green walnut peel at the ageing stage30. The walnut genomic and transcriptomic data provides an important tools and methods to accelerate breeding and to facilitate the genetic dissection of complex traits29.

However, there is no report about the leaf, stem and leaf organ transcriptomics of walnut yet.

The objective of this study was to identify, the groups of walnut genes involved in different organs to provide information for understanding organ development and organ-specific plant-pathogen interactions.

In this study, we used RNA-Seq technology Illumina HiSeq TM2000 sequencing platform to develop the J. regia cultivar Qingxiang transcriptome dataset. De novo transcriptome sequencing was performed on RNA from three different J. regia organs, roots, stem and leaves. We assembled 94,500 transcripts, and identified 51807 unigenes, annotated 28,209 unigenes, and mapped 5,885 unigenes to the 116 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. We also identified 942 tissue-specific candidate unigenes and 6244 SSRs.

In addition, the frequency of the most abundant sequences in each of the libraries was compared across all libraries to identify genes that are potentially and differentially expressed in differential tissues. This novel dataset will be a vital asset in the further hereditary characterization of J. regia and will be especially important in marker helped breeding and examination of qualities identified with disease resistant. To our understanding, this is the main investigation that portrayal the total transcriptome of walnut cultivar Qingxiang. We rely on this new dataset that it will be a valuable asset for future hereditary and genomic considers on this species and cultivar.

RESULTS AND DISCUSSION

Sequence Analysis and Assemble

To obtain a global overview of the walnut J. regia L. transcriptome profile, a mixed cDNA samples representing root (T1), stem (T2) and leaf (T3) tissues of walnut, was prepared and sequenced using the Illumina sequencing platform. Each sequenced sample yielded 31,381,741bp, 26,466,420 bp and 27,384,458 bp independent reads respectively. After stringent

quality assessment and data filtering, 85,232,619 reads (17.22 Gb) with 100% Cycle Q20 and 82.26% Q30 bases (those with a base quality greater than 30) were selected as high quality reads for further analysis (Table 1).

Using the Trinity de novo assembly program, constructed whole length transcript without gaps. The sequences were assembled into 4336022 contigs, 94500 transcripts, with N50 length of 48bp and 1831bp respectively, and with a mean length of 50.54bp and 1108.37bp respectively.

Among of them, there were 38450 transcripts that are longer than 1 kb and 15602 transcripts longer than 2 kb. The transcripts were subjected to cluster and assembly analyses. A total of 51807 unigenes was obtained, with N50 length of 1493bp and mean length of 815.77bp, among which 13371 genes (25.81%) were greater than 1kb.

Before the removal of any low quality sequences or sequences that were too short, 62.51% of the transcripts had lengths in the 300-2000 nt range. These results demonstrated the effectiveness of Illumina sequencing in a rapid way of capturing a large portion of the transcriptome.

The length distributions of assembled unigenes are shown in Fig.1, which revealed that more than 22376 unigenes (43.19%) are greater than 500 bp. An overview of the assembled contigs transcripts and unigenes is presented in Table 2.

Unigenes ORF prediction was completed by etorf (http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html). If several ORFs existed in one unigene, adopt the longest one as its ORF sequence. ORFs length distribution presented in Table3.

Unigenes Annotation

The analysis of assembled unigenes indicated that of the 51,807 unigenes, 28,124 (54.29%) had significant matches in the Nr database while 18,762 (36.22%) unigenes had similarity to proteins in the Swiss-Prot database. Altogether, 28,209 (54.45%) unigenes were successfully annotated in the Nr, Swiss-Prot, KEGG, COG and CO databases listed in Table 4.

The 45.71% of unmapped unigenes that can be assigned a putative function might be mainly due to the short sequence reads generated by the sequencing technology and the relatively short sequences of the resulting unigenes, most of which probably lack the conserved functional domains. Another possible reason is that some of these unigenes might be non-coding RNAs. Meanwhile, some walnut genes have no hits to the syntenic region or even other regions of walnut related suggesting they might be cultivar-specific genes. Nr species distribution except Juglans presented in Fig.2.

Most importantly, these outcomes showed the unwavering quality of Illumina combined end sequencing and de novo gathering. Gene Ontology (GO) examination was completed, which gives a dynamic, controlled vocabulary and various levelled connections for the portrayal of data on molecular capacity, cell segment and biological process, permitting a reasonable explanation of gene product.. There were 28,124 unigenes annotated in Nr, among which 23,451 unigenes were assigned with one or more GO terms (total: 238,564), with 82,988 (34.78%) for cellular components, 123567 (51.80%) for biological processes, and 32,009 (13.42%) for molecular functions (Fig.3).

Furthermore, all unigenes were subjected to a search against the COG database for practical expectation and classification. Overall, 12,076 of the 51,807 successions demonstrating a hit with the Nr database could be allocated to COG classification (Fig.4). COG annotated putative proteins were functionally classified into at least 22 protein families involved in replication, recombination and repair, transcription, signal transduction mechanisms and so on.

The assemblage for the general capacity forecast (2,328; 19.28%) described the biggest group, trailed by replication, recombination and repair (1188;9.84%) ; transcription (1155;9.56%), signal transduction mechanisms (1001;8.30%); Post-translational modification, protein turnover, chaperones (782;6.48%); Carbohydrate transport and metabolism (690;5.71%); Translation, ribosomal structure and biogenesis (645;5.34%); Amino acid transport and metabolism (602;4.99%); Energy production and conversion (477;3.95%); Secondary metabolites biosynthesis, transport and catabolism (434;3.59%); Inorganic ion transport and metabolism (423;3.5%); Lipid transport and metabolism (353;2.95%); Cell cycle control, cell division, chromosome partitioning (284;2.35%); Cell wall/membrane/envelope biogenesis (275;2.28%); Coenzyme transport and metabolism (237;1.96%); Cytoskeleton (164;1.36%); Intracellular trafficking, secretion, and vesicular transport (146;1.21%); RNA processing and modification (139;1.15%); Nucleotide transport and metabolism (127;1.05%);Chromatin structure and dynamics (90;0.75%); (10;0.08%); and whereas only a few unigenes were assigned to cell motility and nuclear structure (10 and 1 unigenes, respectively), the results showed that there is no unigenes assigned to extracellular structure. In addition, 167 unigenes (1.38%) were assigned to defense mechanisms and 358(2.96%) unigenes function are still unknown.

To further demonstrate the usefulness of walnut unigenes generated in the present study, we identified biochemical pathways represented by the unigene collection. The annotations of walnut unigenes were fed into the KEGG Pathway Tools, which is an alternative approach to categorize genes functions with the emphasis on biochemical pathways. This process predicted a total of 116 pathways represented by a total of 5,885 unigenes. The summary of the sequences involved in these pathways is shown in Fig.5.

These predicted pathways represented the majority of biochemical pathways for Plant hormone signal transduction, Protein processing in the endoplasmic reticulum, oxidative phosphorylation, RNA transport and RNA degradation.

The enzymes catalyzing almost all the steps in plant metabolic pathways include the Citrate cycle (TCA cycle), glycolysis, gluconeogenesis, the pentose phosphate pathway, and the various important secondary metabolite biosynthesis pathways includes the carotenoid biosynthesis and flavonoid and anthocyanin biosynthesis, and this could be represented by unigenes derived from the walnut dataset.

Moreover, 164 unigenes involved in Plant-pathogen interactions which account for 2.79% of annotated unigenes were also found in the unigene collection.

Differentially expressed genes analysis of rots, stems and leaves

The cDNA samples representing roots (T1), stems (T2) and leaves (T3) organs of walnut, was prepared and sequenced using the Illumina sequencing platform. Each sequenced sample yielded 31,381,741-bp, 26,466,420-bp and 27,384,458-bp independent reads respectively. The differentially expressed genes are listed in Table 5 and Fig.6.

Among all the three assemblies, 942 unigenes were different in expression. These comparisons indicate that patterns of gene expression are quite different in the various

tissues examined, yielding a diverse array of transcript sequences.

T1 vs T2

The comparison of roots (T1) and stems (T2), roots as control, 4604 differentially expressed genes, of them, 4018 genes had known functions, 2251 (48.9%) were up-regulated (Fig.7).
In biology, a couple of genes are coordinated to network, to fulfill special biological functions. The KEGG pathway analysis facilitates our understanding of DEGs functions in different organs. The differentially expressed genes that are mainly related to cell part, organelle, binding, catalytic activity, cellular process and metabolic process are shown in Table 6. We selected 20 DEGs for further analysis (Fig. 8). In Fig.8, the dots that are closest to the right-up corner of the plot indicate more valuable reference conversely, less valuable. Therefore, comparing the roots, stems organ, the genes related with photosynthesis (antenna proteins), plant hormones signal transduction, and plant-pathogen interactions are more significant.

GO classification of DEGs

GO database could be used in many species to describe and determine genes or proteins. They were divided into three categories; cellular component, molecular function and biological process (Fig. 9).

COG annotation of DEGs

COG database was constructed based on bacteria, algae and eukaryotes evolutionary relationships. The COG annotation of DEGs was presented in Fig.10.

T1 vs T3

Comparison of roots and leaves 5994 differentially expressed genes,out of them 5249 genes had known functions while, 2911 (48.6%) were up-regulated.
Differentially expressed genes that are mainly related to cell part, organelle, binding, catalytic activity, cellular process and metabolic process were shown in Fig.11. We selected 20 DEGs for further analysis (Fig.12). In Fig.12, compare to roots, leaf organ, the genes related with carbon fixation in photosynthetic organisms and starch and sucrose metabolism, are probably more significant.
The COG annotation of DEGs was presented in Fig. 13.
Among plant-pathogen interaction pathways, the DEGs involved in up- regulated genes such as bacterial EF-Tu and NOS pathways are related to EFR and NO mediated hypersensitive response(Fig.14).

T2 vs T3

Comparison of stems and leaves, 5628 differentially expressed genes out of them, 4930 genes had known functions and 2612 (46.4%) were up-regulated.
Differentially expressed genes that are mainly related to cell part, organelle, binding, catalytic activity, cellular process and metabolic process are shown in Fig.15. We selected 20 DEGs for further analysis (Fig.16). In Fig.12, compare to roots, leaf organ, the genes related with carbon fixation in photosynthetic organisms and starch and sucrose metabolism are probably more significant.
The COG annotation of DEGs was presented in Fig.17.
Among plant-pathogen interactions pathways, the DEGs involved in up-regulated genes, such as bacterial EF-Tu and NOS are related to EFR and NO mediated hypersenstive response(Fig.18).
Among plant-pathogen interaction pathways, the DEGs involved in up-regulated genes, besides bacterial EF-Tu and NOS are related to EFR and NO mediated hypersensitive response. The pathways connected to FLS2 and WRKY are related to defense-related gene induction (Fig.18).

SSRs type and statistic

To explore SSR profiles in the unigenes of walnut, the 13,371(≧1 kb, 25.81%) unigene sequences were submitted to an online service to search for SSRs. Among them, 1,162unigene sequences (24.69%) containing more than one SSR, of which mono-nucleotide repeat motif was the most abundant accounting for 44.20%, followed by di-nucleotide repeat motif (39.67%), tri-nucleotide(14.86%), tetra-nucleotide(1.29%), and penta-nucleotide (0.11%) repeat units (Table 7).
In total, 6,244 SSRs were obtained from 4,707unigenes (35.20%)

The distribution of different repeat type classes

Definement of microsatellites (unit size / minimum number of repeats):(1/10) (2/6) (3/5) (4/5) (5/5) (6/5).In total,6,244 SSRs were obtained from 4,707unigenes (35.20%) with 1,162unigene sequences (24.69%) containing more than one SSR, of which mono-nucleotide repeat motif was the most abundant accounting for 44.20%, followed by di-nucleotide repeat motif (39.67%), tri-nucleotide (14.86%), tetra-nucleotide(1.29%) and penta-nucleotide (0.11%) repeat units.

MATERIALS AND METHODS

Plant materials preparation

A widely-cultivated J. regia cultivar Qingxiang was used for transcriptome analysis. Qingxiang nuts were planted in May 2013 and they were grown under natural conditions until they were ready for collection. Roots (T1), Stems (T2) and leaves (T3) were collected 12 months after planting. Tissues were snap-frozen in nitrogen upon harvest and were stored at -80℃ until further processing.

RNA Extraction

Total RNA was isolated from the samples of roots, stems and leaves using Trizol Reagent (Invitrogen, Carlsbad, CA, USA), then treated with DNase I (Fermentas, Pittsburgh, PA, USA) according to the manufacturers' instructions. RNA quality was checked on 1.5% agarose gels, and concentration was measured with NanoDrop ND-2000C (ThermoFisher Scientific, Wilmington, DE, USA).

Library Preparation for Transcriptome Sequencing

A total amount of 2 μg RNA per sample was used as input material for the mRNA sample preparations. Sequencing libraries were generated using NEBNext mRNA Library Prep Master Mix Set for Illumina（NEB, E6110 Ipswich, MA, USA)）and NEBNext Multiplex Oligos for Illumina（NEB, E7500 Ipswich, MA, USA) following manufacturer's

recommendations and index codes were added to attribute sequences to each sample.

Briefly, mRNA was purified from total RNA using NEBNext Poly（A）mRNA Magnetic Isolation Module（NEB, E7490）according to the manufacturer's instructions. Fragmentation was done utilizing divalent cations under raised temperature in NEBNext First-Strand Synthesis Reaction Buffer. First strand of cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase.

The second strand of cDNA synthesis was performed using DNA polymerase I and RNase H. The remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 30 ends of DNA fragments, NEBNext adaptor with hairpin loop structure was ligated to prepare for hybridization. The library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, MA, USA). Then 3 μL USER Enzyme (NEB) was used with size-selected, adaptor-ligated cDNA at 37 ℃ for 15 min followed by 5 min at 95℃ before PCR. PCR was performed by Phusion High-Fidelity DNA polymerase, universal PCR primers and index (X) Primer.

After amplification by PCR, fragments were separated using electrophoresis and purified on 1.8% agarose gel. Subsequent, QPCR conducted by quantification using the Library Quantification Kit-Illumina GA Universal (Kapa, KK4824). Finally, PCR products were purified (AMPure XP system, Beckman Coulter, Inc., Brea, CA, USA) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

CLUSTERING AND SEQUENCING

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v4-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq TM2000 platform and paired-end reads were generated.

Transcriptome Assembly

The raw sequence data were generated by the Illumina analysis pipeline. Raw data (raw reads) of fastq format were first processed through in-house perl scripts. In this step, clean data (clean reads) were obtained by removing reads containing adapter, reads containing ploy-N and low quality read from raw data. Only reads with a perfect match or one mismatch were further analyzed and annotated based on the reference genome. After stringent quality assessment and data filtering, clean reads with 100% Cycle and 82% Q30 bases were selected as high quality reads for further analysis. Sequence assembly was conducted by Trinity (http://trinityrnaseq.sourceforge.net/)(http://www.nature.com/nbt/journal/v29/n7/abs/nbt.1883.html).

Gene Functional Annotation

Gene function was annotated using NCBI BLAST 2.2.28+ 31 with an E-value threshold of 10-5 based on the following databases: NCBI non-redundant protein (Nr) sequences32; NCBI non-redundant nucleotide (Nt) sequences; protein family (Pfam) was assigned using the HMMER3.0 package; eukaryotic or Clusters of Orthologous Groups of proteins (KOG/COG); Swiss-Prot (a manually annotated and reviewed protein sequence database)33; KEGG pathways were assigned to the assembled sequences using the online KEGG Automatic Annotation Server (KAAS)(http://www.genome.jp/kegg/kaas/).The bi-directional best hit (BBH) method was used to obtain KEGG Orthology (KO) assignment. The output of KEGG analysis includes KO assignments and KEGG pathways that are populated with the KO assignments34.

Open reading frames (ORFs) were predicted using the ''getorf'' program of EMBOSS software package,with the longest ORF extracted for each unigene（http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html).

To annotate the assembled sequences with GO terms describing biological processes, molecular functions and cellular components, the Swiss-Prot BLAST results were imported into Blast2GO v2.5 (BioBam, Valencia, Spain), a software package that retrieves GO terms, allowing gene functions to be determined and compared35. These GO expressions are relegated to inquiry groupings, delivering a wide review of gatherings of qualities classified in the transcriptome for every one of three ontology vocabularies, biological procedures, molecular capacities and cell composition. The acquired explanation was advanced and refined utilizing ANNEX. The information introduced herein represent GO examination at level 2, representing general utilitarian classes. The unigenes sequences were also aligned to the COG database to predict and classify their possible functions36.

Analysis of Differential Expression and Organ-Specific Loci

Three mRNA libraries were generated from separate organs using Illumina sequencing. Reads for each sequenced tag were mapped to the assembled loci using Bowtie (mismatch #2 bp, other parameters as default), and the number of clean mapped reads for each locus was counted. The DEGseq package was used to identify differentially expressed genes37,38. The three different libraries were compared pairwise using a greater than two-fold difference as the criterion for differential expression. Significant differential expression between organs was defined by p-value, 0.001, FDR, 0.01, and log2.2.

The differential expression analysis between the organs was used to identify candidate loci with organ-specific expressions, and to determine functionally enriched loci, as described above. Organ-specific loci were selected based on the read counts from roots, stems and leaves samples of walnut.

Identification of SSRs

SSRs were detected using the MIcroSAtellite Identification Tool (MISA, http://pgrc.inpk-gatersleben.de/misa/) Perl script. The assembled unigene sequences that are more than 1kb length were screened for mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeat motifs with a minimum repeat number of 10, 6, 5, 5, 5 and 5, respectively. A maximum distance of 100 nucleotides was allowed between two SSRs.

Accession numbers

Sequence data from this article can be found in the NCBI Sequence Read Archive (SUB4285453) under BioProject accession number (PRJNA480600).

CONCLUSIONS

This work presents the de novo transcriptome sequencing analysis of mixed RNA from walnut cv.Qingxiang (roots, stems and leaves) using the Illumina platform. In total, 85,232,619 paired-end reads were generated with 17,215,394,747 nucleotides and GC percentage was 46.18%. These were assembled into 4,336,022 contigs with 94,500 transcripts and 51,807 unigenes. For three organs, 6,338,412,972 (6.33G bp), 5,345,746,994 (5.34G bp) and 5,531,234,781 (5.53 G bp) high quality clean reads were generated from cDNA libraries of roots, stems and leaves, respectively.

Unigenes were annotated by querying against the NCBI non-redundant and UniProt databases 28,209 unigenes (54.45%) were homologous to existing database sequences. Gene Ontology terms were doled out to 23,451 unigenes, 8,292 loci were coordinated to 25 Clusters of Eukaryotic Orthologous Groups orders, and 5,885 unigenes were characterized into 116 Kyoto Encyclopedia of Genes and Genomes pathways.. Besides, 5885 differentially expressed genes (DEGs) were enriched in KEGG pathways. The assembled unigenes（13,371unigenes more than 1 kb）also contained 6,244 potential simple sequence repeats.

The raw reads of transcriptome from J. regia (accession number) were deposited in NCBI database. The assembled transcriptome was used to identify unigenes with organ-specific differential expression patterns. In total, 942 unigenes exhibited organ-specific expression. Our data analyses implied that walnut have different organs that expressed specific genes in many molecular processes and functions. Particularly, the plant-pathogen interaction pathways were demonstrated.

The existing genomic transcriptomic data from China major walnut cultivars are scarce. The walnut organ-based transcriptome resources developed in this study will enable the analysis of organ development, accelerate the marker-assisted breeding based on SSRs and understand the disease resistance in walnut. Our transcriptome data enrich the genomic resource of Juglans species and will be essential to accelerate the process of molecular research and breeding. Furthermore, the dataset will improve our understanding of the molecular mechanisms of organ development, disease resistant and other biochemical processes in walnut.

The existing genomic transcriptomic data from China major walnut cultivars are scarce. The walnut organ-based transcriptome resources developed in this study will enable the analysis of organ development, accelerate the marker-assisted breeding based on SSRs and understand the disease resistance in walnut. Our transcriptome data enrich the genomic resource of Juglans species and will be essential to accelerate the process of molecular research and breeding. Furthermore, the dataset will improve our understanding of the molecular mechanisms of organ development, disease resistant and other biochemical processes in walnut.

REFERENCES

1. Vavilov, N.I. Origin of Cultivated Plants (translated by D. Love). Cambridge,UK: Cambridge University Press. (1992).
2. Martínez M.L., Labuckas D.O., Lamarque A.L. & Maestri D.M.Walnut (Juglans regia L.): genetic resources, chemistry, by-products. J. Sci. Food Agr.90,1959-1967(2010).
3. Crews, C., Hough, P., Godward, J., Brereton, P., Lees, M., Guiet, S. & Winkelmann, W. Study of the main constituents of some authentic walnut oils. J. Agric. Food. Chem. 53, 4853–4860(2005).
4. Korkina, L. Phenylpropanoids as naturally occurring antioxidants: from plant defense to human health. Cell. Mol. Biol. 53, 15–25(2007).
5. Zhang, Z.J., Liao, L.P., Moore, J., Wu, T. & Wang, Z.T. Antioxidant phenolic compounds from walnut kernels (Juglans regia L.). Food Chem.113, 160–165(2009).
6. Britton M.T., Leslie C.A., Caboni E., Dandekar A.M. & McGranahan G.H. Persian Walnut. In: Chittaranjan K and Timothy CH (ed) Compendium of transgenic crop plants: transgenic temperate fruits and nuts. Wiley-Blackwell, Massachusetts. (2008).
7. Feng C.et al. Codon usage patterns in Chinese bayberry (Myricarubra) based on RNA-Seq data. BMC Genomics 14,732(2013).
8. He C.Y., Cui K., Zhang J.G., Duan A.G. & Zeng Y.F. Next-generation sequencing-based mRNA and microRNA expression profiling analysis revealed pathways involved in the rapid growth of developing culms in Moso bamboo.BMC Plant Biol.13,119(2013).
9. Chow K.S., Ghazali A.K., Hoh C.C. & Mohd-Zainuddin Z. RNA sequencingread depth requirement for optimal transcriptome coverage in Heveabrasiliensis. BMC Res. Notes 7, 69(2014).
10. Mutz K.O., Heilkenbrinker A., Lonne M., Walter J.G. & Stahl F. Transcriptome analysis using next-generation sequencing.Curr. Opin. Biotechnol.24, 22–30(2013).
11. Brousseau, L., Tinaut, A., Duret, C., Lang, T., Garnier-Gere, P. & Scotti, I. High-throughput transcriptome sequencing and preliminary functional analysis in four Neotropical tree species. BMC Genomics, 15(1),https://doi.org/10.1186/1471-2164-15-238(2014).
12. Jiang, Z., He, F., & Zhang, Z. Large-scale transcriptome analysis reveals arabidopsis metabolic pathways are frequently influenced by different pathogens. Plant Molecular Biology 94(4–5), 453–467. https://doi.org/10.1007/s11103-017-0617-5(2017).
13. Gustavo,R.A. et al.Transcriptomics insights into the genetic regulation of root apical meristem exhaustion and determinate primary root growth in Pachycereus pringlei (Cactaceae).Scientific Reports 8,8529.DOI:10.1038/s41598-018-26897-1(2018).
14. Yeonhwa J. et al. Peach RNA viromes in six different peach cultivars. Scientific Reports 8,1844.DOI:10.1038/s41598-018 - 20256-w. (2018).
15. Pattison, R. J., Csukasi, F., Zheng, Y., Fei, Z., van der Knaap, E., & Catalá, C. Comprehensive Tissue-Specific Transcriptome Analysis Reveals Distinct Regulatory Programs during Early Tomato Fruit Development. Plant Physiology 168(4), 1684–1701. https://doi.org/10.1104/pp.15.00287(2015).
16. Skibbe, D. S., Doehlemann, G., Fernandes, J., & Walbot, V. Maize tumors caused by ustilago maydis require organ-specific genes in host and pathogen. Science 328(5974), 89–92. https://doi.org/10.1126/science.1185775(2010).
17. Paul, A. L., Zupanska, A. K., Schultz, E. R., & Ferl, R. J. Organ-specific remodeling of the Arabidopsis transcriptome in response to spaceflight. BMC Plant Biology 13(1),https://doi.org/10.1186/1471-2229-13-112 (2013).
18. Brenner, W. G. & Schmülling, T. Transcript profiling of cytokinin action in Arabidopsis roots and shoots discovers largely similar but also organ-specific responses. BMC Plant Biology 12,https://doi.org/10.1186/1471-2229-12-112(2012).
19. Eng, C. H. L., Shah, S., Thomassie, J. & Cai, L. Profiling the transcriptome with RNA SPOTs. Nature Methods 14(12), 1153–1155. https://doi.org/10.1038/nmeth.4500(2017).
20. Giacomello, S.et al. Spatially resolved transcriptome profiling in model plant species. Nature Plants 3, https://doi.org/10.1038/nplants.2017.61(2017).
21. Aradhya, M.K., Potter, D.& Simon, C.J. Cladistic biogeography of Juglans (Juglandaceae) based on chloroplast DNA intergenic spacer sequences. In Darwin's Harvest: New Approaches to the Origins, Evolution and Conservation of Crops (Motley, T.J., Zerega, N. and Cross, H.,eds). New York: Columbia University Press, pp. 143–170.). (2006).
22. Aradhya, M., Woeste, K. & Velasco, D. Genetic diversity, structure and differentiation in cultivated walnut (Juglans regia L.). Acta Hortic. 861, 127–132(2010).

23. Wu, T., Xiao, L.J., Chen, S.Y. & Ning,D.L.Transcriptomicsand comparative analysis of three Juglans species; J. regia, J. sigillata and J. cathayensis. Plant Omics Journal 8(4),361-371 (2015).
24. Bai, W.N., Yan, P.C., Zhang, B.W., Woeste, K. E., Lin, K.& Zhang, D.Y. Demographically idiosyncratic responses to climate change and rapid Pleistocene diversification of the walnut genus Juglans (Juglandaceae) revealed by whole-genome sequences. New Phytologist,https://doi.org/10.1111/nph.14917(2017).
25. Dong, W.et al. Phylogenetic Resolution in Juglans Based on Complete Chloroplast Genomes and Nuclear DNA Sequences. Frontiers in Plant Science 8, 1148. http://doi.org/10.3389/fpls.2017.01148(2017).
26. Wu, J.et al. Characterizing the walnut genome through analyses of BAC end sequences. Plant Molecular Biology 78(1–2), 95–107. https://doi.org/10.1007/s11103-011-9849-y(2012).
27. You, F. M.et al.Genome-wide SNP discovery in walnut with an AGSNP pipeline updated for SNP discovery in allogamous organisms. BMC Genomics 13(1), https://doi.org/10.1186/1471-2164-13-354(2012).
28. Luo, M.C.et al. Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. BMC Genomics 16(1), https://doi.org/10.1186/ s12864-015-1906-5(2015).
29. Martínez-García, P.J.et al. The walnut (Juglans regia) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. The Plant Journal For Cell and Molecular Biology 87(5), 507–532. https://doi.org/10.1111/tpj.13207(2016).
30. Li, Y.et al. Comparative transcriptome analysis of genes involved in anthocyanin biosynthesis in red and green walnut (Juglans regia L.). Molecules 23(1), https://doi.org/10.3390/molecules23010025(2018).
31. Stephen F. et al. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Res.25(17), 3389(1997).
32. Deng, Y.Y. et al. Integrated nr Database in Protein Annotation System and Its Localization. Computer Engineering 32(5),71-74(2006).
33. Apweiler R. et al. UniProt: the Universal Protein Knowledgebase [EB/OL]. Nucleic Acids Res. http:// nar.Oxfordjournals.org/cgi/ content/full/ 32/supp1_1/ d115, 32(Database Issue):115(2004).
34. Minoru K. et al. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32,277-280(2004).
35. Ashburner, M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature Genetics 25(1), 25-9. (2000).
36. Roman L. et al. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28(1),33-6(2000).
37. Anders S. & Huber W. Differential expression analysis for sequence count data. Genome Biol. 11, R106. (2010).
38. Ning L. et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29,1035-1043(2013).

ACKNOWLEDGEMENTS

This work was supported by Basic Research for Application of Yunnan Province (2011FB066) and National Natural Science Foundation of China (31200488). We thank Hubei Linyunong Agricultural Science and Technology Co. Ltd. to afford the walnut cultivar Qingxiang.

AUTHOR CONTRIBUTIONS

B.Z. Fu designed the research, analyzed the data and wrote the manuscript; L.L. Zou performed the lab experiments; L.H. Wang and G.Y. Li helped in preparation of the manuscripts and lab experiments. Competing Interests: The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

Sequence data from this article can be found in the NCBI Sequence Read Archive (SUB4285453) under BioProject accession number (PRJNA480600).

Table 1. Evaluation of sequencing data of walnut differential organs.

Samples	Total reads	Base (bp)	GC (%)	N (%)	Q20%	CycleQ20%	Q30 (%)
T1(Roots)	31,381,741	6,338,412,972	46.16	0.07	89.24	100.00	82.36
T2(Stems)	26,466,420	5,345,746,994	46.17	0.07	89.13	100.00	82.20
T3(leaves)	27,384,458	5,531,234,781	46.22	0.07	89.13	100.00	82.22
Total/average	85,232,619	17,215,394,747	46.18%	0.07	89.17	100.00	82.26%

Table 2. An overview of the assembled contigs, transcripts and unigenes.

Length range	Contigs (account)	Transcripts (account)	Unigenes (account)
0-300	4289194(98.92%)	19828(20.98%)	16422(31.70%)
300-500	19801(0.46%)	17767(18.80%)	13009(25.11%)
500-1000	12926(0.30%)	18455(19.53%)	9005(17.38%)
1000-2000	9295(0.21%)	22848(24.18%)	8332(16.08%)
2000+	4806(0.11%)	15602(16.51%)	5039(9.73%)
Total number	4336022	94500	51807
Total length	219123587	104740886	42262833
N_{50} length	48	1831	1493
Mean length	50.54	1108.37	815.77

Table 3. ORF length distribution.

Length range	Unigene ORF
0-300	33,139(64.20%)
300-500	4,577(8.87%)
500-1000	6,037(11.70%)
1000-2000	5,839(11.31%)
2000+	2,023(3.92%)
Total number	51,615
Total length	24,979,707
N50 length	1,155
Mean length	483.96

Table 4. Unigenes annotation in Nr, GO, COG, KEGG and Swissprot databases.

Database	Annotated_Number	300<=length<1000	length>=1000
Nr	28124	10661	12888
GO	23451	8229	11653
COG	8292	1984	5579
KEGG	5885	1900	3131
Swissprot	18762	6528	9533
All	28209	10709	12892

Table 5. Comparison of differentially expressed genes among three organs.

Type	number	up	down
T1 vs T2	4,604	2,251	2,353
T1 vs T3	5,994	2,911	3,083
T2 vs T3	5,628	2,612	3,016

Table 6. KEGG annotation of DEGs

Pathway	DEGs with pathway annotation (457)	All genes with pathway annotation (4060)	p_value	corr_p_value	Pathway ID
Phenylpropanoid biosynthesis	35 (7.66%)	89 (2.19%)	3.3614e-12	3.2606e-10	ko00940
Phenylalanine metabolism	30 (6.56%)	83 (2.04%)	1.6221e-09	1.5735e-07	ko00360
Photosynthesis	33 (7.22%)	102 (2.51%)	6.3753e-09	6.1840e-07	ko00195
Photosynthesis - antenna proteins	19 (4.16%)	44 (1.08%)	6.3127e-08	6.1233e-06	ko00196
Porphyrin and chlorophyll metabolism	19 (4.16%)	53 (1.31%)	2.0121e-06	1.9517e-04	ko00860

Plant hormone signal transduction	47 (10.28%)	229 (5.64%)	1.9472e-05	1.8888e-03	ko04075
Pentose and glucuronate interconversions	18 (3.94%)	61 (1.5%)	7.7778e-05	7.5445e-03	ko00040
Zeatin biosynthesis	8 (1.75%)	17 (0.42%)	2.3438e-04	2.2734e-02	ko00908
Tropane, piperidine and pyridine alkaloid biosynthesis	8 (1.75%)	24 (0.59%)	3.4724e-03	3.3682e-01	ko00960
Plant-pathogen interaction	30 (6.56%)	164 (4.04%)	4.3281e-03	4.1982e-01	ko04626

Table 7. Results of microsatellite search.

Searching item	Numbers
Total number of sequences examined	13,371
Total size of examined sequences (bp)	26,973,526
Total number of identified SSRs	6,244
Number of SSR containing sequences	4,707
Number of sequences containing more than 1 SSR	1,162
Number of SSRs present in compound formation	459
Mono-nucleotide	2,751
Di-nucleotide	2,477
Tri-nucleotide	928
Tetra-nucleotide	81
Penta-nucleotide	7

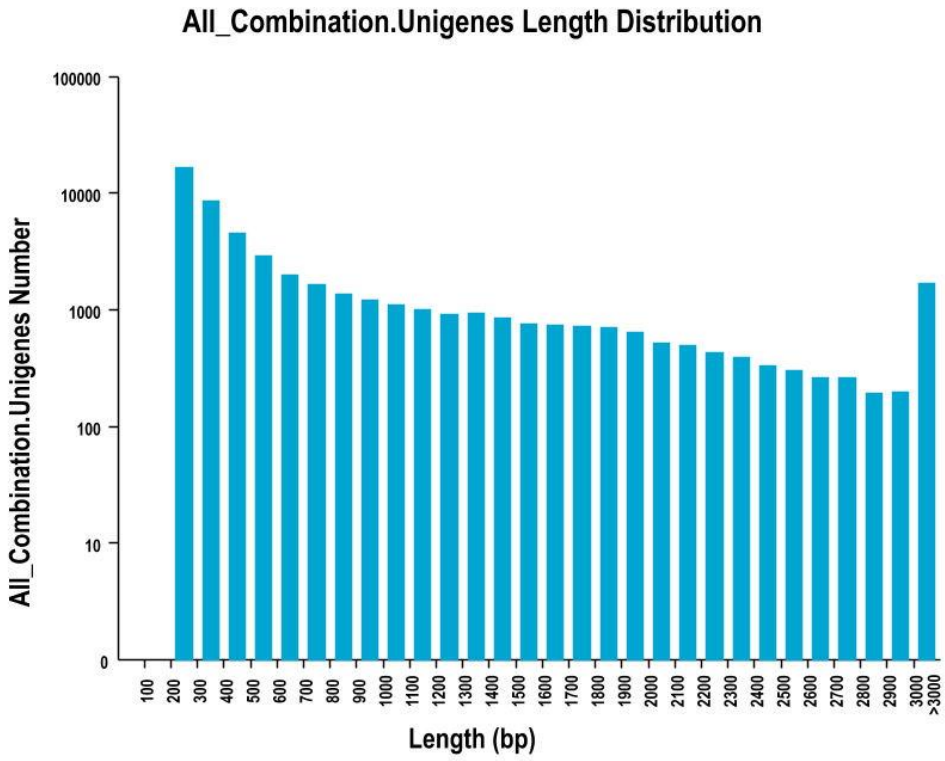

Figure 1. The length distributions of unigenes.

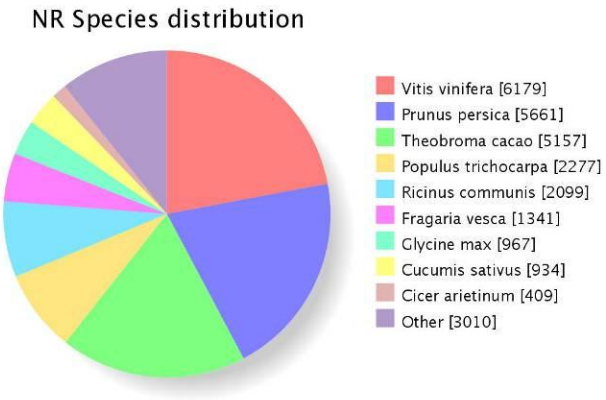

Figure 2. Nr species distribution in other plants.

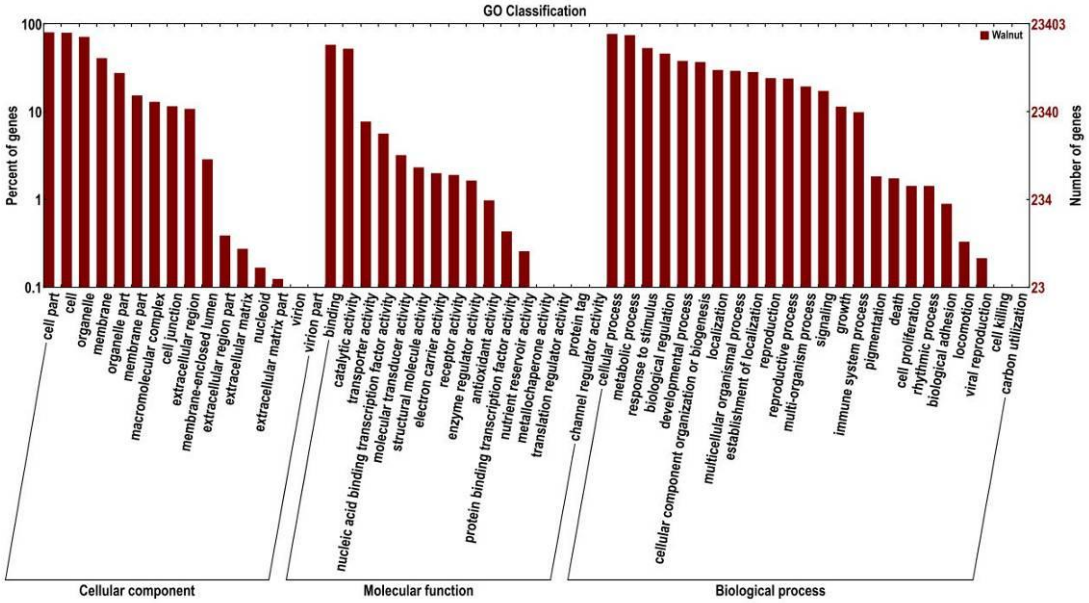

Figure 3. Functional annotation of the assembled sequences based on gene ontology (GO) categorization.

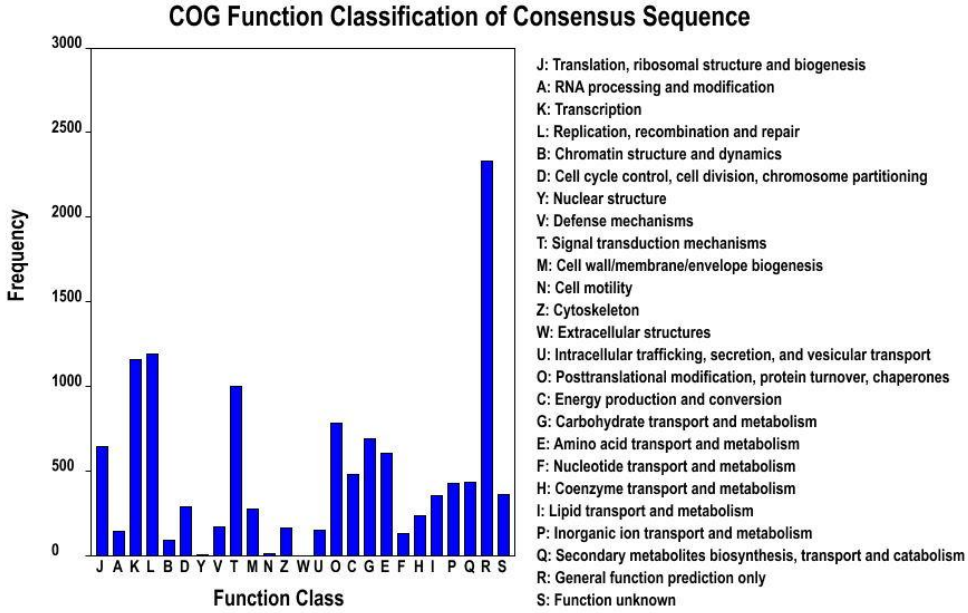

Figure 4. Clusters of orthologous groups (COG) classification.

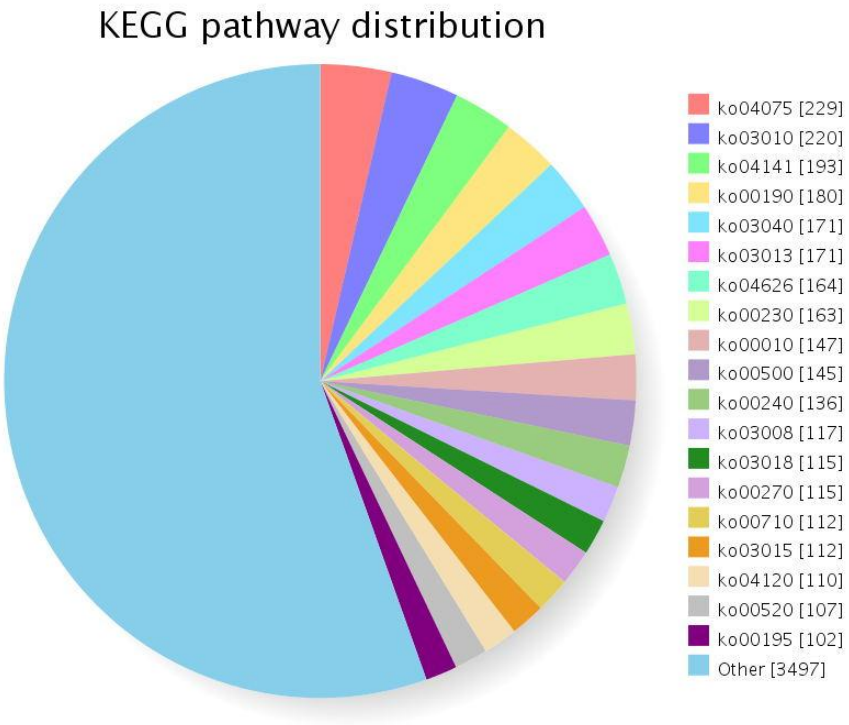

Figure 5. Unigenes KEGG pathway distribution.

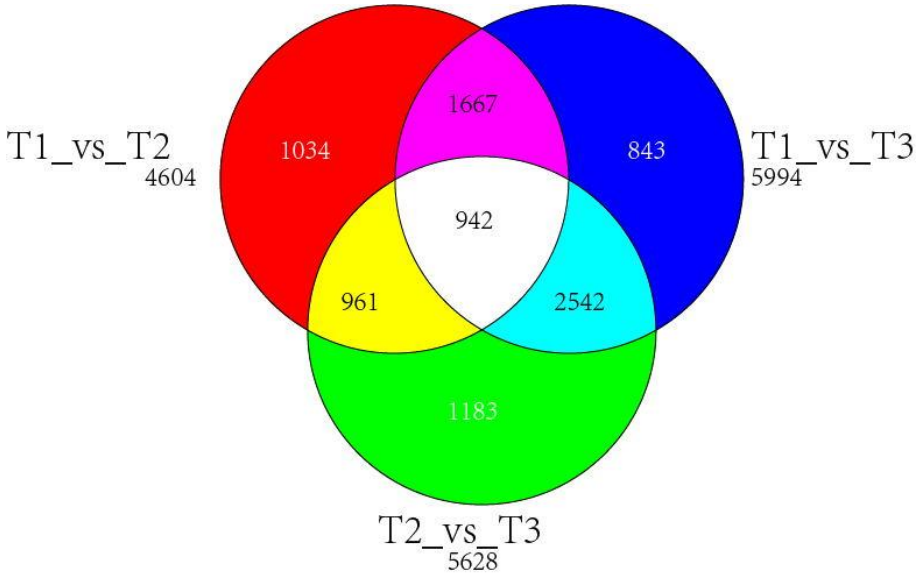

Figure 6. Vinn diagram showing the total number of unigenes from each of the three assemblies (roots, stems and leaves) and the numbers of unigenes shared between each pair of assemblies and all three assemblies.

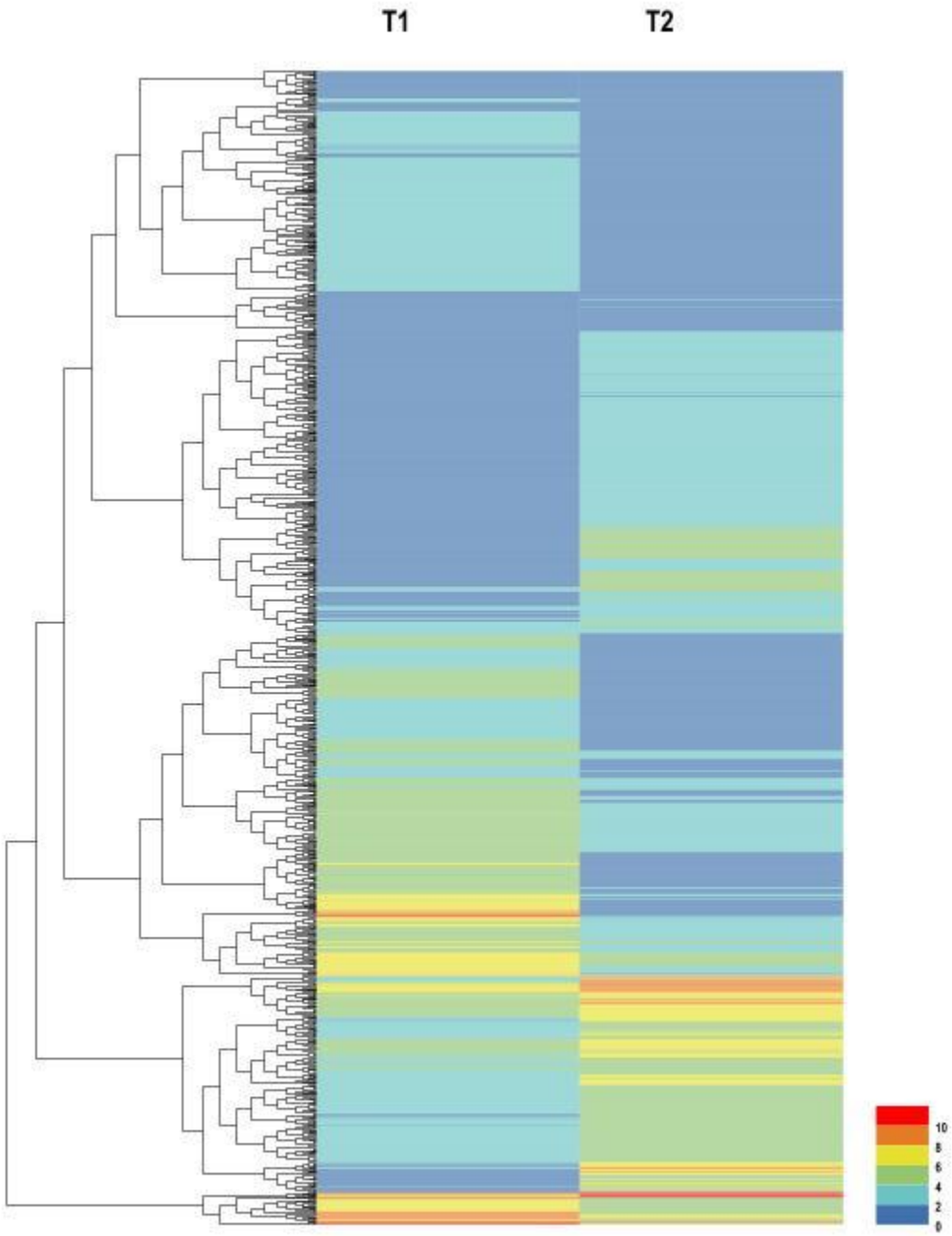

Figure 7. Clustering of differentially expressed genes. (Yellow means high expression, blue means low expression.)

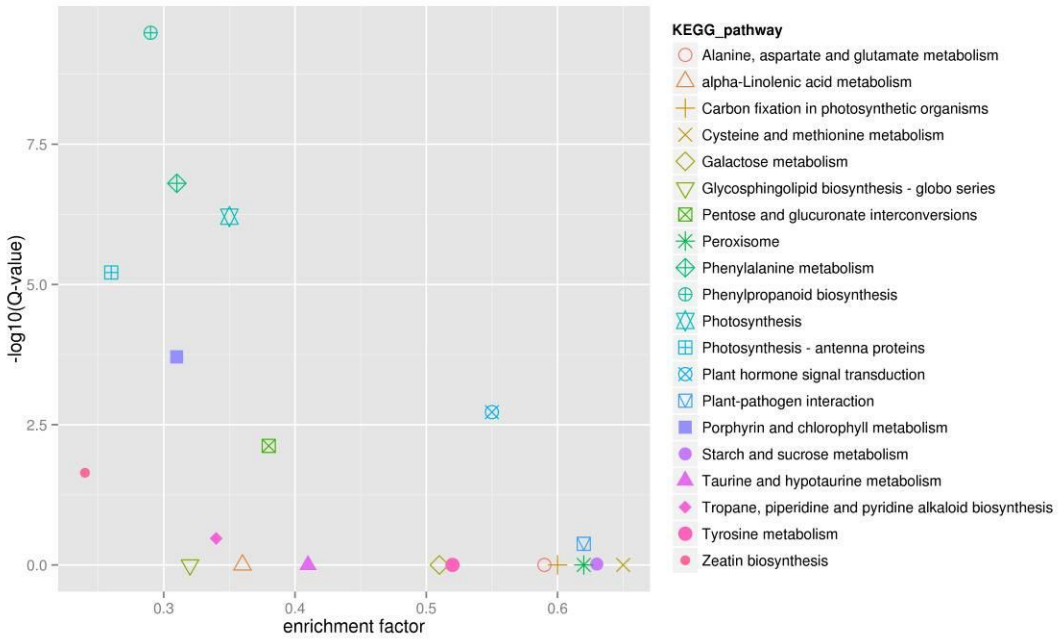

Figure 8. Shown scatter plots of differentially expressed genes of stems compared to the roots.

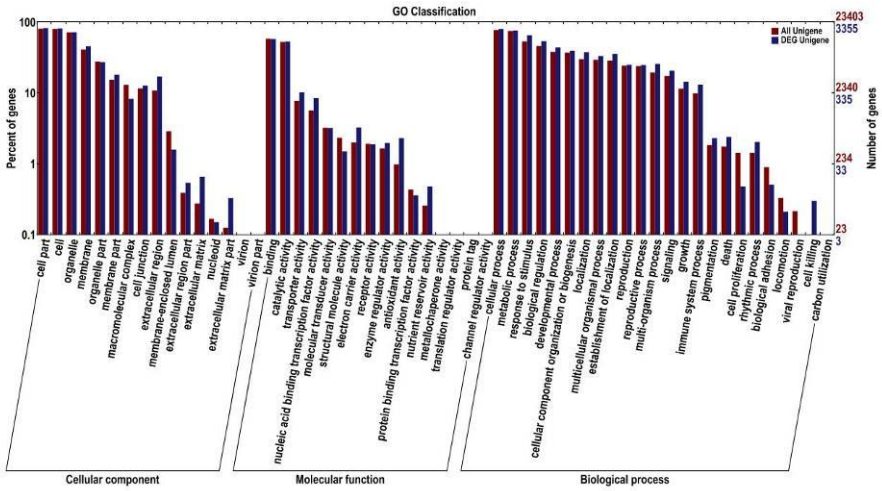

Figure 9. Clustering of GO of DEGs

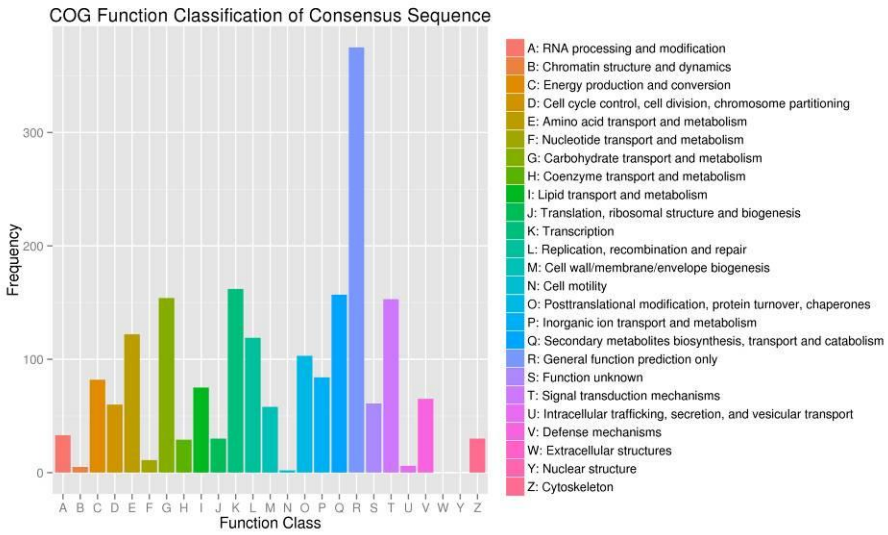

Figure 10. Classfication of COG annotation of DEGs.

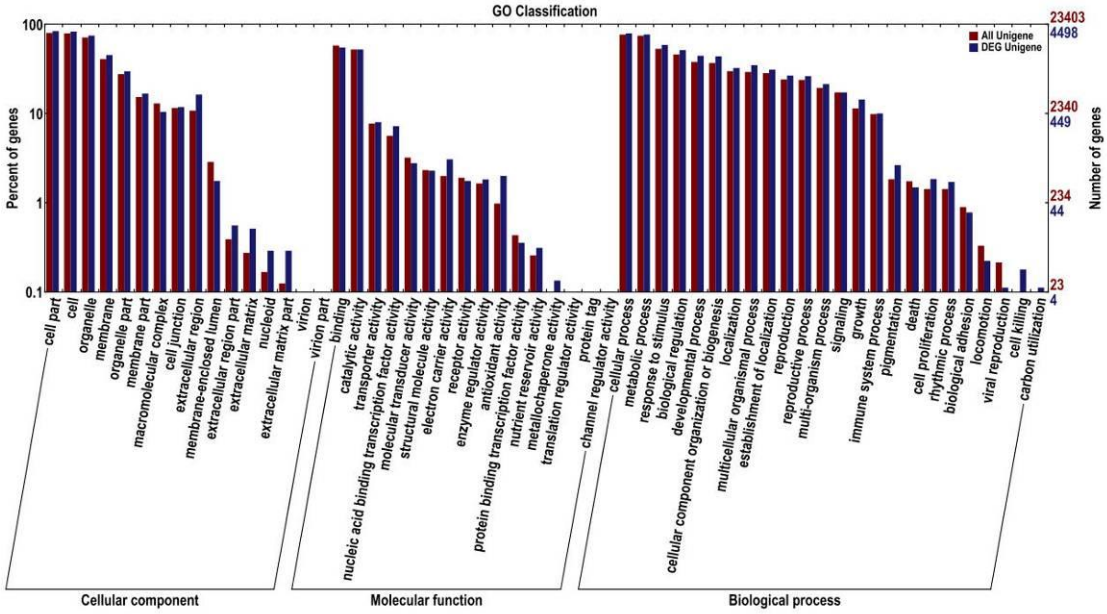

Figure 11. Clustering of GO of DEGs.

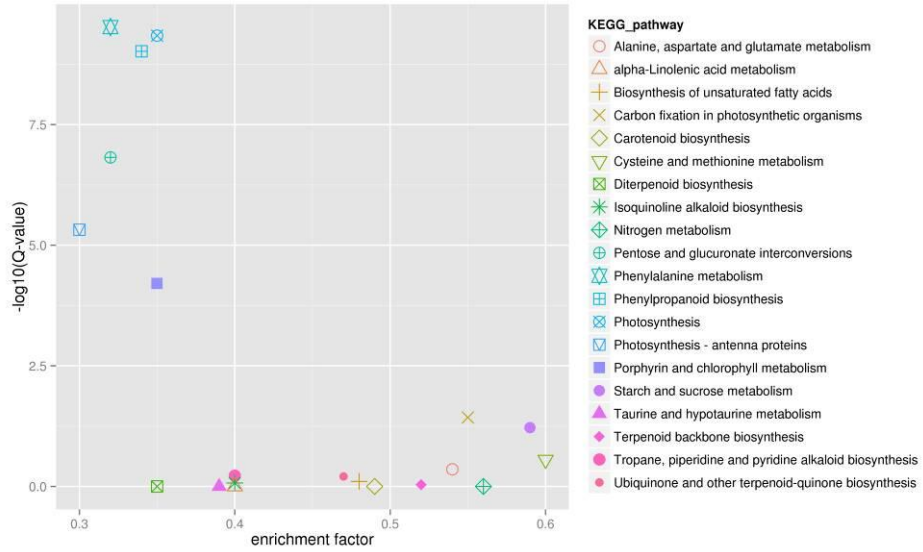

Figure 12. Shown scatter plots of differentially expressed genes of stems compared to roots.

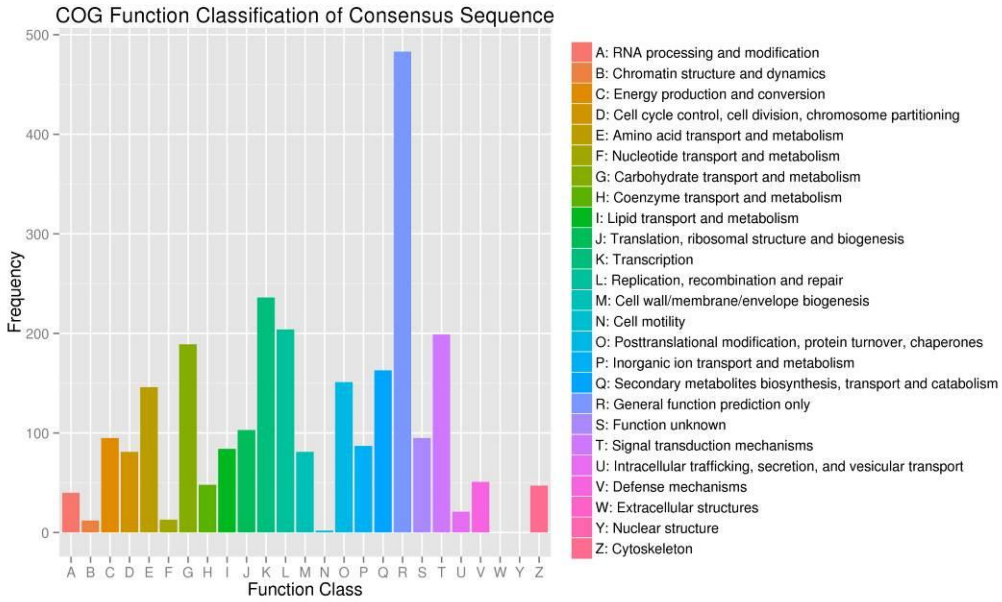

Figure 13. Classification of COG annotation of DEGs.

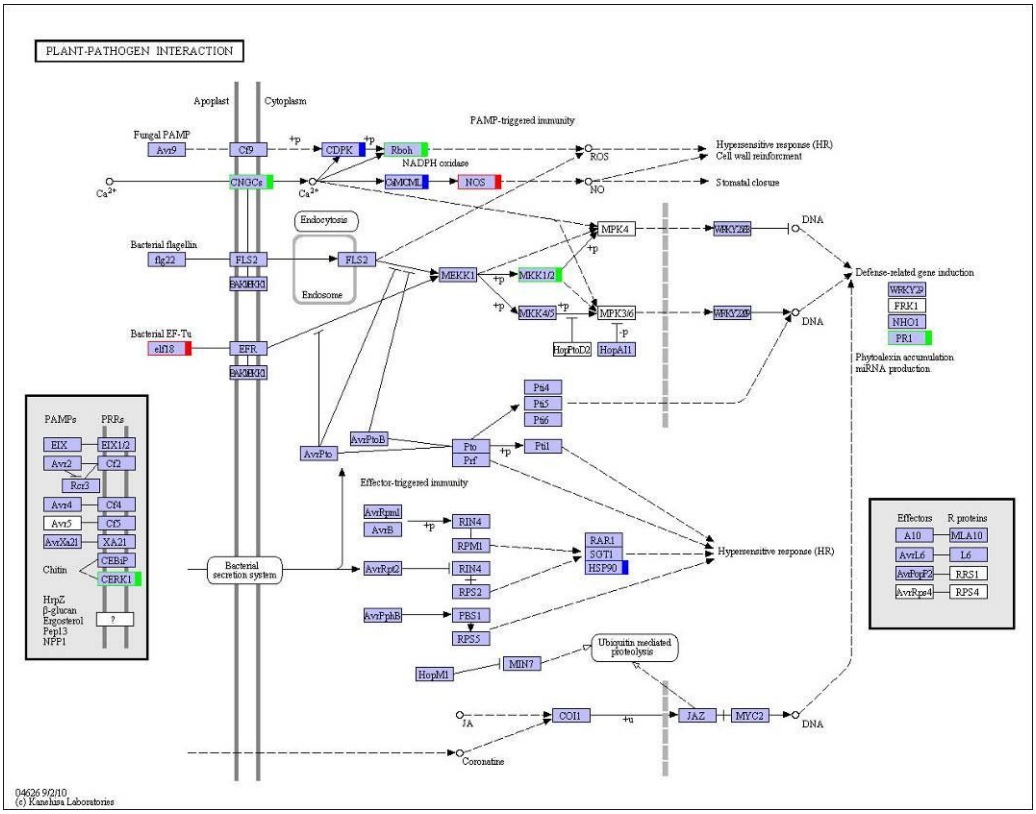

Figure 14. KEGG annotation of DEGs.(The red frame represent up-regulated genes, green frame represent down-regulated genes, blue frame means contain up and down regulated genes at the same time).

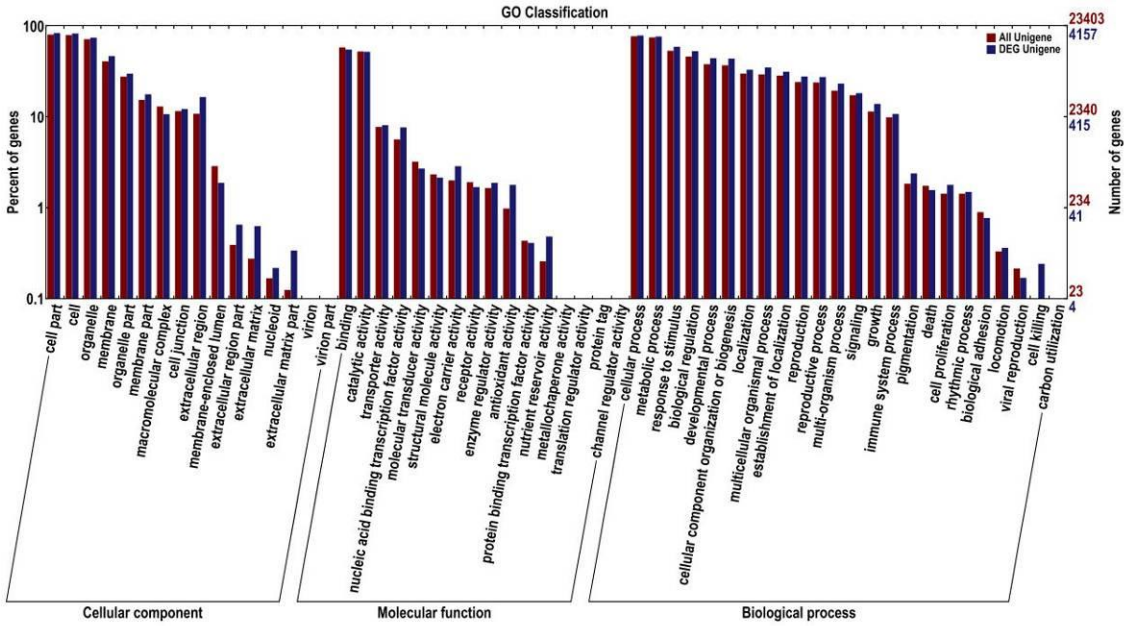

Figure 15. Clustering of GO of DEGs.

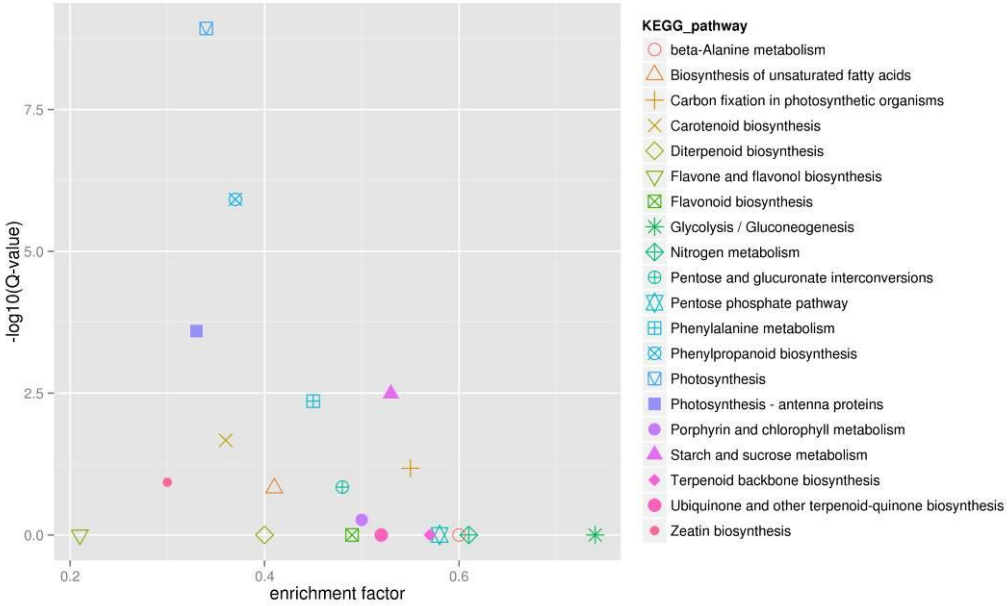

Figure 16. Scatter plot of differentially expressed genes of stems compare to roots.

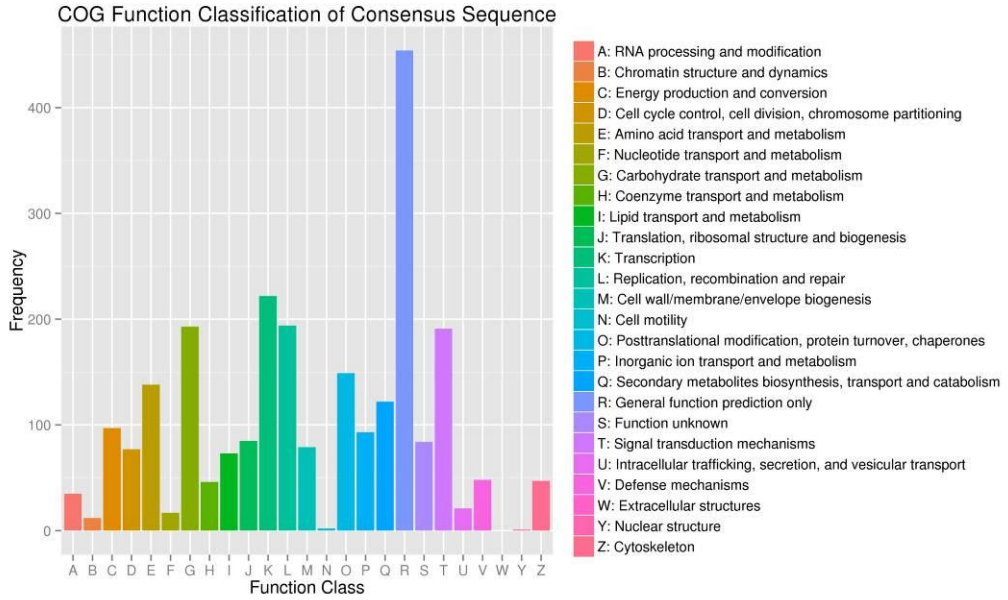

Figure 17. Classification of COG annotation of DEGs.

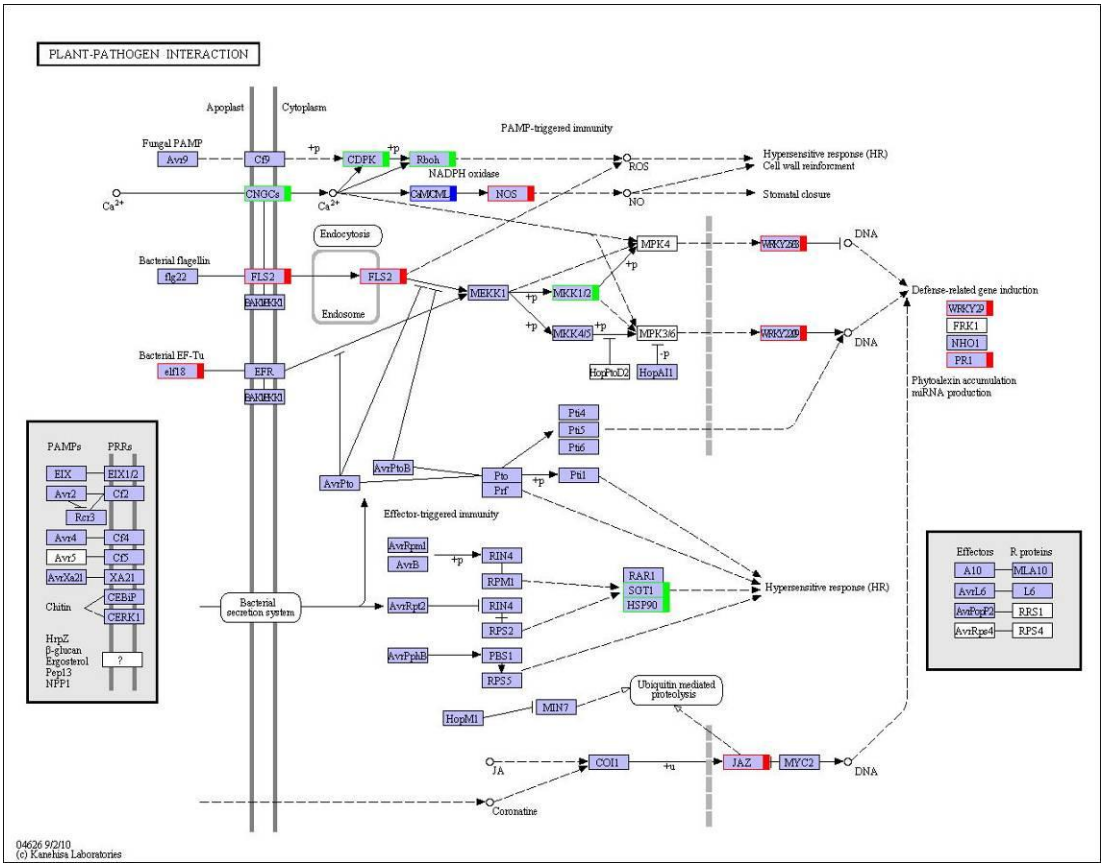

Figure 18. KEGG annotation of DEGs.(The red frame represent up-regulated genes, green frame represent down-regulated genes, blue frame means contain up and down regulated genes at same time.)