Plant Omics Journal POJ

POJ 8(4):361-371 (2015) ISSN:1836-3644

Transcriptomics and comparative analysis of three *Juglans* species; *J. regia*, *J. sigillata* and *J. cathayensis*

Tao Wu[†], Liangjun Xiao[†], Shaoyu Chen, Delu Ning*

Institute of Economic Forest, Yunnan Academy of Forestry, Kunming 650201, China

***Corresponding author: ningdelu@163.com**

†T. Wu and L. Xiao contributed equally to this work.

Abstract

Walnut (*Juglans*) has been globally cultivated for its valuable nut, which has abundant polyunsaturated fatty acids and proteins. In China, only the Persian or English walnut (*J. regia*) and Yunnan or iron walnut (*J. sigillata*) are commercially cultivated for nut production, and Chinese butternut (*J. cathayensis*) is commonly used as rootstock and potential breeding material. Only few genomic resources are available for these non-model plants, particularly the last two species. Hence, we present the sequencing, *de novo* assembly and annotation of transcriptomes from fresh leaves of the three *Juglans* species by RNA-seq technology and bioinformatics analysis to discover a collection of SSR and SNP markers, to be used by researchers for further genetic improvements. In total, 59035134 (7.38 G bp), 43949544 (5.5 G bp) and 58609226 (7.32 G bp) high quality clean reads were generated from cDNA libraries of *J. regia*, *J. sigillata* and *J. cathayensis*, respectively. A total of 192360 unigenes longer than 200 bp were *de novo* assembled, 92858 (48.20%) unigenes were annotated, and 32110 CDSs (16.70%) were deduced. The potential function of each unigene was classified based on COG and GO database. 5683 differentially expressed genes (DEGs) were enriched in KEGG pathways. A total of 41141 SSRs and 206355 SNPs were identified as potential molecular markers. The raw reads of transcriptome from *J. regia* (accession number SRR1767234), *J. sigillata* (accession number SRR1767236) and *J. cathayensis* (accession number SRR1767237) were deposited in NCBI database. Our transcriptome data enrich the genomic resource of *Juglans* species and will be essential to accelerate the process of molecular research and breeding.

Keyword: Walnut; *Juglans*; Transcriptome; High-throughput sequencing; Differential expression genes.

Abbreviations: BLAST_Basic Local Alignment Search Tool; CDS_Coding Region Sequence; COG_Cluster of Orthologous Groups of proteins; GO_Gene Ontology; KEGG_Kyoto Encyclopedia of Gene and Genomes; DEGs_differentially expressed genes; SSR_simple sequence repeat; SNP_single nucleotide polymorphism

Introduction

All species of walnuts (*Juglans*) produce nuts, but the Persian or English walnut (*J. regia*) is the most widespread species cultivated for nut production. It is globally popular and valued for its nutritional, health and sensory attributes (Martínez et al., 2010). China is considered to be one of the main walnut production countries (Britton et al., 2008). Annual world walnut production totaled 3418559 metric tons in 2012, of which, China produced 1700000 tons, followed by the Iran (450000 T), USA (425820 T), Turkey (194298 T), Mexico (110605 T), Ukraine (96900 T), India (40000 T), Chile (38000 T), France (36425 T), and Romania (30546 T) (FAOSTAT, updated 24 December 2014, http://faostat.fao.org/faostat/).

In China, walnut is distributed from 21°29' to 44°54' north latitude and from 75°15' to 124°21' east longitude, and there are six species in the walnut genus (*Juglans*). They are *J. regia*, *J. sigillata*, *J. cathayensis*, *J. mandshurica*, *J. cordiformis* and *J. hopeiensis*, and all species produce nuts, but only *J. regia* and *J. sigillata* are commercially cultivated for nut production. Only less than 17 authorized or approved cultivars of *J. sigillata* have been popularized, including ‘Yangpao’, ‘Santai’, ‘Niangqing’, while the others are still wild. Nevertheless, the Chinese walnut or Chinese butternut, *J. cathayensis*, is a vigorous tree, commonly used as rootstock for *J. regia* to provide tolerance of biotic/abiotic stresses in regions of the Yangtze River in China. Because of climatic diversity, high heterozygosity and sexual propagation, very rich genetic materials exist among Chinese walnut population (McGranahan and Leslie, 1990, 2009). The considerable variations between *J. regia* and *J. sigillata*, particularly in nut size and shape, led taxonomists to describe other additional species that have not been widely accepted but that illustrate some of the diversity (Dode, 1909). Undoubtedly, the best genotypes selected for nut improvement and the expansion of commercial growth range are probably possible based on germplasm evaluation and breeding.

Although walnut has been cultivated for centuries, walnut breeding starts recently and only a few systemic molecular studies on walnut have been reported (McGranahan and Leslie, 1990, 2009; Britton et al., 2008). Because of its commercial value, far more gene sequences are available for *J. regia* than other members of the same genus, even though, the number of nucleotide sequences of *J. regia* is still smaller than other crops (Dandekar et al., 2005). As of December 2014 GenBank (National Center for Biotechnology Information, NCBI), the public repository for DNA sequence data in the USA, listed 6287 nucleotide sequences for *J. regia*, including nuclear and chloroplast genes, and expressed sequence tags (ESTs). The derived information is sporadic and limited, and insufficient to use, especially to picture a global transcriptome profile for genetic improvements on the important agronomic and economic traits, such as improved climate adaptation (late budbreak, low chilling requirement or winter hardiness), early fruiting and high productivity (lateral fruitfulness), disease tolerance (blight and anthracnose). This situation is merely a question with the increasing availability of low cost, high-throughput sequencing technologies (Metzker, 2010; Grabherr et al., 2011).

In this study, total RNAs, extracted from the leaves of three *Juglans* species, *J. regia*, *J. sigillata* and *J. cathayensis*, were sequenced using RNA-seq technology, and the assembled sequences were analyzed. In addition, a collection of cDNA-derived SSR and SNP markers were developed and characterized. The data obtained here might be beneficial for cloning genes of interest and genetic improvements of *Juglans*.

Results and Discussions

Illumina sequencing and de novo assembly

The cDNA libraries of *J. regia* (JRE), *J. sigillata* (JSI) and *J. cathayensis* (JCA) were generated using the mRNA-Seq procedure for transcriptome sequencing on an Illumina Hiseq™ 2000 platform. Using Illumina paired-end sequencing technology, each sequencing feature can yield 2 × (100 ± 25 bp) independent reads from both ends of a 200 ~ 300 bp cDNA fragment. After data filtering, a total of 59035134, 43949544 and 58609226 high quality clean reads from cDNA libraries of JRE, JSI and JCA (a total of 7.38×10^9, 5.50×10^9 and 7.32×10^9 base pairs) were obtained, respectively (Table 1). A total of 113043, 115036 and 123567 transcripts longer than 200 bp were generated with a mean length of 1103 bp, 1066 bp and 1063 bp and a N50 of 1886 bp, 1814 bp and 1869 bp for JRE, JSI and JCA, respectively. To remove any redundancies, transcripts were clustered using the TGICL (Pehtea et al., 2003). The sequences not extended on either end were defined as unigenes. A total of 60710, 62415 and 69235 unigenes longer than 200 bp were generated by the clustering, with a mean length of 712 bp, 700 bp and 674 bp and a N50 of 1300 bp, 1283 bp and 1204 bp for JRE, JSI and JCA, respectively (Table 1). The raw reads of transcriptome from JRE, JSI and JCA were deposited in NCBI database (Accession numbers: SRR1767234 for JRE, SRR1767236 for JSI, and SRR1767237 for JCA).

On average, this assembly produced a substantial number of large unigenes: 22708 unigenes (35.41%) longer than 500 bp, 12333 unigenes (19.23%) longer than 1000 bp, and 4837 unigenes (7.54%) longer than 2000 bp (Fig. 1). These results suggested that the number of unigenes assembled from JRE, JSI and JCA were close to each other. The genome size of one of three walnut species (*J. regia*, C=0.62 pg, approximately 606 Mb) was recently described (Bennett and Leitch, 2012), but the percentages of the transcribed genomes remain unknown. Thus, it is difficult to predict the depth of coverage of the walnut transcriptome by our *de novo* assembled sequences.

One of the most important questions in this type of study is whether the short reads were correctly assembled. However, the presently available publications on *de novo* assembly are very limited, especially for non-model organisms. As no generally accepted protocol for evaluating such an assembly exists, it is therefore a great challenge to make a reliable judgement or validation of an optimal assembly without the reference genomic sequence. Fortunately, many ESTs and genomic survey sequences deposited in NCBI database can serve as reference. An alternative approach to experimental validation is to conduct computational analysis. We selected 49 unigenes from this *de novo* assembly data, which could be matched with nucleotide sequences of *J. regia* in GenBank (as reference sequence) through megablast search. There was 32783 bps (96.0%, 32783/34188) of 49 unigenes producing identical alignments (Unigene sequences available in Supplementary Table S1). Moreover, although a total of 3219 out of 4837 unigenes longer than 2000 bp had significant matches that covered more than 90% of their corresponding subjects, they were classified as putative, hypothetical or predicted protein. Overall, these results suggested that our data assembly was of high quality, but also indicated that the major assembled unigenes have not been sequenced previously or far not well been characterized.

Annotation and CDS prediction

For annotation, homologs of the unigenes were searched in seven databases: NT, NR, Swiss-Prot, KOG, KO, GO and PFAM. A total of 26704 (43.98%) JRE unigenes, 32055 (51.35%) JSI unigenes and 34099 (49.25%) JCA unigenes were annotated to at least one of the seven databases (Table 2). Among them, most unigenes could be annotated to the NR database, followed by the Swiss-Prot, GO, PFAM, NT, KOG and KO database. Based on NR annotation and E-value distribution, 70.58% of the mapped unigenes had high homology (E-value<le^{-30}), and 33.94% showed very strong homology (E-value<le^{-100}) with known protein sequences (Fig. 2A). The similarity distribution showed 77.34% of the annotated sequences have a similarity greater than 60% (Fig. 2B). Averagely, 80.73% of unigenes can be annotated to the top 5 species in the species distribution (Fig. 2C). After BLASTx analysis, 96.33% of the unigenes over 2,000 bp had BLASTx hits, while only 41.28% of the unigenes shorter than 500 bp had homologs. It indicates that the length of query sequence was important for determining the level of significance of the BLASTx match. Longer unigenes were more likely to have BLASTx hits in protein databases.

The coding sequence (CDS) of all unigenes was predicted by ESTScan or BLAST with an E-value threshold of 10^{-5} in the NR and Swiss-Prot protein database. The front and back sequences beyond CDS in the unigene were considered as the 5′ and 3′ UTR sequences. A total of 11682 JCA unigenes (16.87%) and 11266 JSI unigenes (18.05%) contained the 5′ and 3′ UTR sequence, and the full-length open reading frames (ORF), whereas the number of unigenes containing these in JRE transcriptome was only 9162 (15.09%) (Table 1).

Functional classification

In order to classify the potential function of each unigene, all walnuts unigenes were aligned to the KOG database. The result showed that all the JRE, JSI and JCA unigenes covered all 26 KOG functional categories (Fig 3). The top 10 categories were R, O, T, J, C, K, U, A, G, I, respectively. Another alternative approach, GO assignments were used to classify the functions of all unigenes. Based on sequence similarity, 322098 unigenes of all walnuts were assigned to one or more ontologies. Totally, 147316 unigenes were grouped under biological processes, 102789 unigenes under cellular components, 71993 unigenes under molecular functions. Binding (32676 unigenes, 45.39%) and catalytic activities (28793 unigenes, 40.00%) were the most highly represented classes under the molecular function category. For the biological process class, the assignments were mainly given to the cellular process (34592 unigenes, 23.48%) and metabolic process (33192 unigenes, 22.53%). In the cellular components category, the largest proportion of transcripts was involved in the ‘cell’ (47.78%), followed by the ‘cell part’ (33.21%), ‘organelle’ (5.35%) ‘macromolecular complex’

Table 1. Statistics for pyrosequencing of the three *Juglans* species.

Species	*J. regia*	*J. sigillata*	*J. cathayensis*
Total no. of reads	59035134	43949544	58609226
Total size (bp)	7.38×10^9	5.50×10^9	7.32×10^9
Total no. of transcripts	113043	115036	123567
Total size wihtin transcripts (bp)	124693119	122644019	131347815
Average length of transcripts (bp)	1103	1066	1063
N50 of transcripts	1886	1814	1869
N90 of transcripts	447	426	408
Total no. of unigenes	60710	62415	69235
Total size wihtin unigenes (bp)	43208532	43696782	46668510
Average length of unigenes (bp)	712	700	674
N50 of unigenes	1300	1283	1204
N90 of unigenes	273	269	262
Unigenes with full-length ORF	9162 (15.09%)	11266 (18.05%)	11682 (16.87%)
No. of unigenes annotated	26704 (43.98%)	32055 (51.35%)	34099 (49.25%)

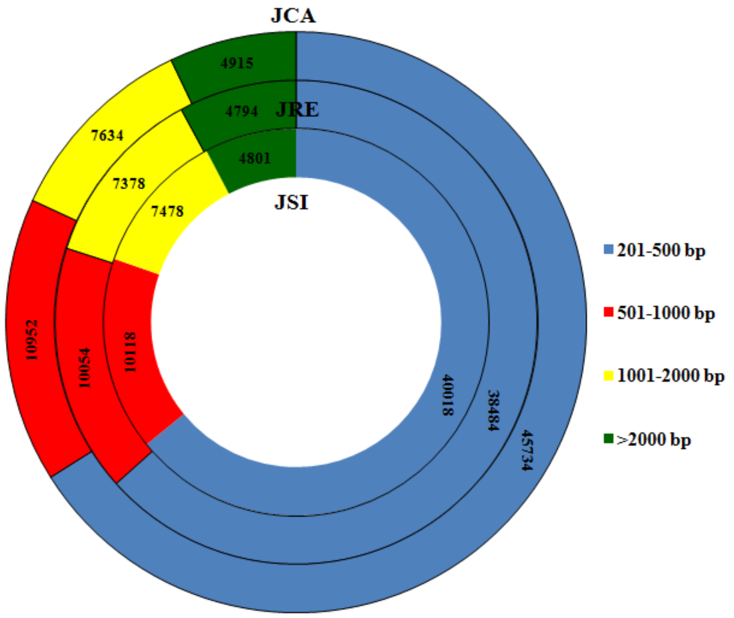

Fig 1. Length distribution of assembled unigenes of JRE, JSI and JCA. The outer, middle and inner cycles represent the distribution of JCA, JRE and JSI unigenes, respectively. The data in each colours region indicates the number of unigenes in this range of nucleotide length. Blue region indicates unigene size range from 201 to 500 bp, red 501-1000 bp, yellow 1001-2000 bp, green above 2000 bp.

(7.48%) and membrane (5.35%) (Fig. 4). The GO results were similar to herbaceous plants *Eleusine indica* (Shu et al., 2015), *Daucus carota* var. *sativus* (Iorizzo et al., 2011) and woody plants *Camellia sinensis* (Shi et al., 2011), *Hevea brasiliensis* (Xia et al., 2011), suggesting that our unigenes are broadly representative of the flesh plant leaf transcriptome. The concordance in the overall distributions trend of three species suggests that our library widely sampled across GO categories and provides a good representation of the *Juglans* species leaf transcriptome. These KOG and GO annotations provide a valuable clue for investigating the specific processes and molecular function of *Juglans* species.

It can be seen from Fig 4 that the number of classified unigenes of *J. regia* is robustly smaller than other two species. One possible explanation for this is that a unigene or its product can be annotated to more than one term of each ontology, at any level within each ontology. Another possible reason is that many of the assembled sequences may represent distinct non-overlapping regions of the same genomic locus. Finally, it is likely due to the sequence data of *J. regia* far outnumber *J. sigillata* and *J. cathayensis* in GO database, many of short unigenes have been assembled together into full-length unigenes, reducing the total number of unigenes in the transcriptome.

When the samples were collected for sequencing, the walnut trees were actively growing. For example *J. regia* under the fourth level GO terms, vigorous cell growth and metabolic processes were reflected by the assignation of 2152 transcripts to ‘cellular biosynthetic process’, 2270 to ‘organic substance biosynthetic process’, 2192 to ‘organic cyclic compound metabolic process’, 2133 to ‘protein metabolic process’, 2121 to ‘cellular nitrogen compound metabolic process’, 2071 to ‘cellular aromatic compound metabolic process’ and 1343 to ‘organic acid metabolic process’. The plant has a strong self-regulation ability to respond to stimuli, and this was supported by the assignation of 2685 transcripts to ‘response to stimulus’, 3339 to ‘biological regulation’ and 3112 to ‘regulation of biological process’.

Comparative analysis of differential expressed genes (DEGs)

The coefficient of determination (R^2) of gene expression profile reflects the similarity between samples. The higher the R^2, the

Table 2. Summary of the unigene annotations of the three *Juglans* species.

Annotation database	*J. regia*	*J. sigillata*	*J. cathayensis*
NR	25970 (42.77%)	29417 (47.13%)	31095 (44.91%)
GO	11398 (18.77%)	22471 (36%)	23923 (34.55%)
SwissProt	17883 (29.45%)	20179 (32.33%)	21006 (30.34%)
PFAM	8816 (14.52%)	19074 (30.55%)	20222 (29.2%)
NT	15100 (24.87%)	16323 (26.15%)	16835 (24.31%)
KOG	8964 (14.76%)	10374 (16.62%)	10600 (15.31%)
KO	7703 (12.68%)	8847 (14.17%)	9101 (13.14%)
Annotated in all Databases	3118 (5.14%)	4151 (6.65%)	4319 (6.23%)
Annotated in at least one Database	26704 (43.98%)	32055 (51.35%)	34099 (49.25%)

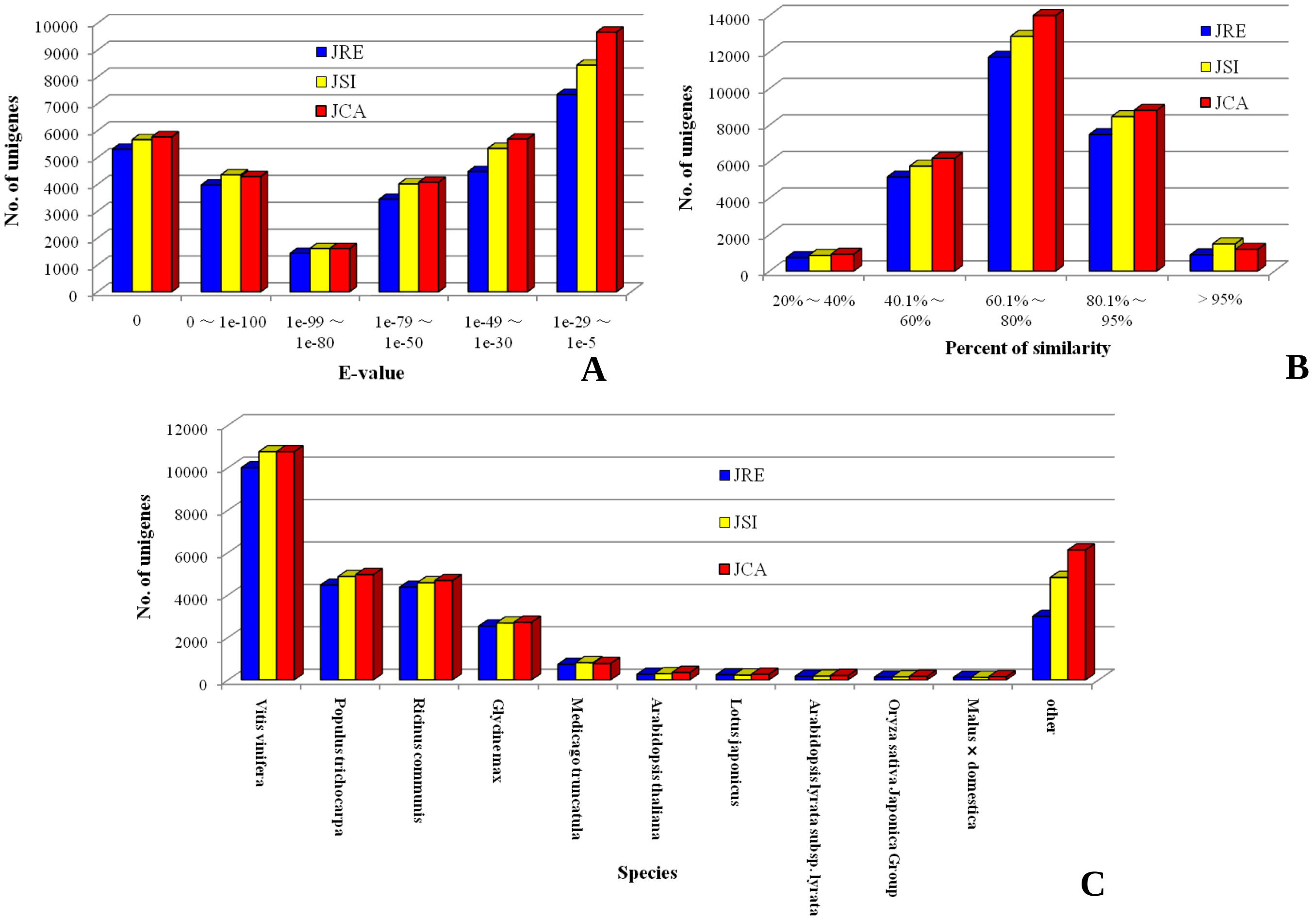

Fig 2. Homology alignments of unigenes of JRE, JSI and JCA against NR database. E-value distribution (A), similarity distribution (B) and species distribution.

stronger the similarity. The R^2 was higher between JSI and JRE ($R^2 = 0.708$) than between JRE and JCA ($R^2 = 0.533$) or JSI and JCA ($R^2 = 0.507$). The result of hierarchical clustering analysis within three species showed also the gene expression pattern is closer between JSI and JRE than between JSI and JCA or JRE and JCA (Fig 5). The results about gene expression of three species, to sum up, agreed with traditional botany taxonomy, in which JRE and JSI belong to same section *Juglans* but JCA belongs to another section *Cardiocaryon* (Manning, 1978; Wang et al., 2015).

To study the conservation and divergence of global expression patterns, we performed comparative analysis for all unigenes from three species. As shown in Volcano plot (Fig 6) and Venn diagram (Fig 7), a total of 684 genes were differentially expressed between JRE *Vs* JSI analysis, of which 587 and 97 genes showed the common and exclusive differential expression with other analyzed groups, respectively. Totally, 2215 DEGs were identified in JCA *vs* JSI analysis, of which 1841 and 374 genes showed the common and exclusive differential expression with rest of the analyzed groups, respectively. Highest number of genes (2784) differentially expressed in JRE *vs* JCA analysis, of which 1846 and 938 genes showed the common and exclusive differential expression with rest of the analyzed groups, respectively. 211 genes were commonly expressed between both JRE and JCA samples compared with JSI one. About 1470 genes were commonly expressed between JCA *vs* JSI and JRE *vs* JCA samples, while 216 genes showed the common expression in between JRE *vs* JCA and JRE *vs* JSI samples. Of all, only 160 genes were commonly expressed in all three species, among that,

Table 3. The differential expressed genes (DEGs) identified from transcriptome data within JRE, JSI and JCA.

Gene ID	Gene Name	Gene description	Size (bp)	Fold change / log2		
				JRE Vs JSI	JRE Vs JCA	JCA Vs JSI
comp115501_c0	FAD8	Omega-3 fatty acid desaturase (delta-15 desaturase)	1636	-2.9833	-4.2523	1.3058
comp125017_c2	FAD2	Omega-6 fatty acid desaturase (delta-12 desaturase)	1859	-1.0242	-2.3015	1.3141
comp118997_c0	KCS	3-ketoacyl-CoA synthase 6	913	-2.1354	-3.7428	1.6442
comp116221_c0	KCS	3-ketoacyl-CoA synthase 12	1749	-3.2335	-4.9476	1.7509
comp120968_c0	KCS	3-ketoacyl-CoA synthase 21	2265	-1.6482	-3.563	1.9516
comp114842_c0	CHS	Naringenin-chalcone synthase	1656	1.0658	-3.3973	4.4998
comp115430_c0	CHS	Chalcone synthase	1590	-1.2052	-2.7808	1.6123
comp125700_c0	WSD1	Wax ester synthase-like Acyl-CoA acyltransferase	1971	-2.2993	-3.7916	1.5291
comp117542_c0	GDE1	Glycerophosphoryl diester phosphodiesterase	1822	-1.7483	-3.0505	1.339
comp116141_c0	CICLE	GDSL-like Lipase/Acylhydrolase	1631	1.3438	3.7606	-2.3800
comp114457_c0	AUX28	Auxin-induced protein	1281	-1.1681	-2.2551	1.1238
comp119129_c0	TEM1	AP2/ERF and B3 domain-containing transcription represso	1563	-1.0104	1.2174	-2.1910
comp121653_c0	HSFF	Heat stress transcription factor A-2b	2841	-2.5351	-1.1395	-1.3589
comp123626_c0	HSF30	Heat shock factor protein	2684	-1.6521	1.3662	-2.9815
comp95810_c0	NAC29	NAC transcription factor 29	1318	1.6982	4.2963	-2.5613
comp123093_c0	TAG	WRKY transcription factor 40	4696	1.4162	2.7957	-1.3427
comp114659_c0	HSP20	Heat shock protein	1052	-1.7824	5.7675	-7.5131
comp101719_c0	ARG2	Indole-3-acetic acid-induced protein	1866	1.1289	3.1685	-2.0029

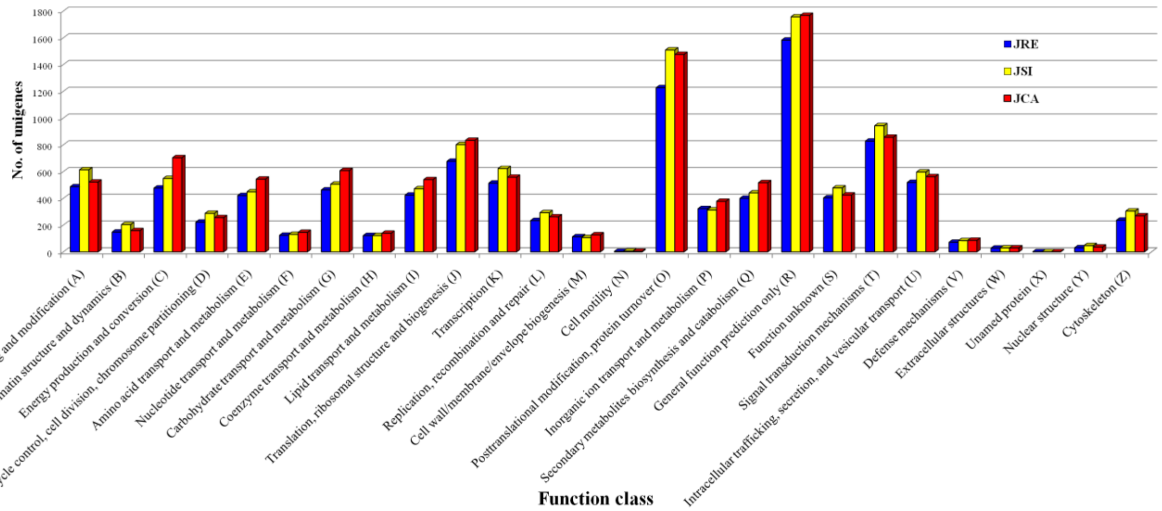

Fig 3. Functional classifications of KOG terms of unigenes from JRE, JSI and JCA.

there were 20 involved in oxidation-reduction process, 11 in regulation of transcription, 19 ones were hypothetical in nature and have not yet been annotated for coding specific protein. The DEGs having least p-values and higher fold changes within JRE, JSI and JCA species were further sorted out based on the functional involvement in fatty acid metabolism, regulation of transcription and response to stress, important genes are listed in Table 3.

KEGG pathway enrichment analysis of differential expressed genes (DEGs)

KEGG is a database resource for understanding high-level functions from molecular-level information, especially large-scale molecular datasets generated by high-throughput experimental technologies (Kanehisa et al., 2008). DEGs were subjected to KEGG pathway enrichment analysis, and 18.99% (1079/5683) of the DEGs could be annotated, which were associated with 417 KEGG pathways (Supplementary Table S2, Table S3, Table S4,). The 20 top KEGG pathways with the highest representation of the DEGs are shown in Table 4. The Biosynthesis of amino acids (ko01230), Starch and sucrose metabolism (ko00500), Phenylpropanoid biosynthesis (ko00940), Cysteine and methionine metabolism (ko00270), Phenylalanine metabolism (ko00360), Photosynthesis (ko00195), Phenylalanine, tyrosine and tryptophan biosynthesis (ko00400), Flavonoid biosynthesis (ko000941) and Fatty acid elongation (ko00062) pathways are significantly enriched both JRE *vs* JCA and JSI *vs* JCA. Nevertheless, Protein processing in endoplasmic reticulum (ko04141), Phagosome (ko04145), Gap junction (ko04540) and Phenylpropanoid biosynthesis (ko00940) pathways are significantly enriched in JRE *vs* JSI, and the number of DEGs was less than other comparisons.

SSR and SNP discovery

Transcriptomes are an important resource for the rapid and cost-effective development of genetic markers (Liao et al., 2014). The molecular markers derived from the transcribed regions are more conservative, providing a greatest potential for identifying functional genes. Among the various molecular markers, simple sequence repeats (SSRs) and single nucleotide

Table 4. 10 top KEGG pathways with high representation of the DEGs.

Rank	JRE vs JCA			JSI vs JCA			JRE vs JSI		
	Pathway	Pathway ID	No. of DEGs	pathway	Pathway ID	No. of DEGs	pathway	Pathway ID	No. of DEGs
1	Biosynthesis of amino acids	ko01230	44 (8.66%)	Biosynthesis of amino acids	ko01230	39 (9.18%)	Protein processing in endoplasmic reticulum	ko04141	22 (15.17%)
2	Phenylpropanoid biosynthesis	ko00940	25 (4.91%)	Starch and sucrose metabolism	ko00500	21 (4.94%)	Phagosome	ko04145	9 (6.21%)
3	Plant hormone signal transduction	ko04075	25 (4.91%)	Cysteine and methionine metabolism	ko00270	18 (4.24%)	Gap junction	ko04540	7 (4.83%)
4	Starch and sucrose metabolism	ko00500	25 (4.91%)	Phenylpropanoid biosynthesis	ko00940	18 (4.24%)	Phenylpropanoid biosynthesis	ko00940	7 (4.83%)
5	Cysteine and methionine metabolism	ko00270	19 (3.73%)	Phenylalanine metabolism	ko00360	14 (3.29%)	MAPK signaling pathway	ko04010	6 (4.14%)
6	Phenylalanine metabolism	ko00360	15 (2.95%)	Flavonoid biosynthesis	ko00941	12 (2.82%)	alpha-Linolenic acid metabolism	ko00592	6 (4.14%)
7	Photosynthesis	ko00195	14 (2.75%)	Phenylalanine, tyrosine and tryptophan biosynthesis	ko00400	12 (2.82%)	Endocytosis	ko04144	6 (4.14%)
8	Flavonoid biosynthesis	ko00941	13 (2.55%)	Photosynthesis	ko00195	11 (2.59%)	Peroxisome	ko04146	6 (4.14%)
9	Phenylalanine, tyrosine and tryptophan biosynthesis	ko00400	13 (2.55%)	Sulfur metabolism	ko00920	8 (1.88%)	Flavonoid biosynthesis	ko00941	5 (3.45%)
10	Fatty acid elongation	ko00062	12 (2.36%)	Fatty acid elongation	ko00062	8 (1.88%)	Cysteine and methionine metabolism	ko00270	5 (3.45%)

Table 5. Statistics of SNPs generated among three *Juglans* species.

Types	JRE	JSI	JCA
Transitions	36408 (59.28%)	54458 (59.78%)	31973 (59.39%)
A-G	18167 (29.58%)	27476 (30.16%)	16197 (30.09%)
C-T	18241 (29.70)	26982 (29.62%)	15776 (29.31%)
Transversions	25014 (40.72%)	36643 (40.22%)	21859 (40.61%)
A-C	6104 (10.00%)	9131 (10.02%)	5248 (9.75%)
A-T	7688 (12.52%)	10996 (12.07%)	6923 (12.86%)
C-G	4783 (7.79%)	7036 (7.72%)	4229 (7.86%)
T-G	6439 (10.48%)	9480 (10.41%)	5459 (10.14%)
non coding SNP	44459 (72.38%)	65190 (71.56%)	36683 (68.14%)
coding SNP	16963 (27.62%)	25911 (28.44%)	17149 (31.86%)
synonymous	16883 (27.49%)	25806 (28.33%)	17060 (31.69%)
nonsynonymous	80 (0.13%)	105 (0.12%)	89 (0.17%)
Total	61422 (100%)	91101 (100%)	53832 (100%)

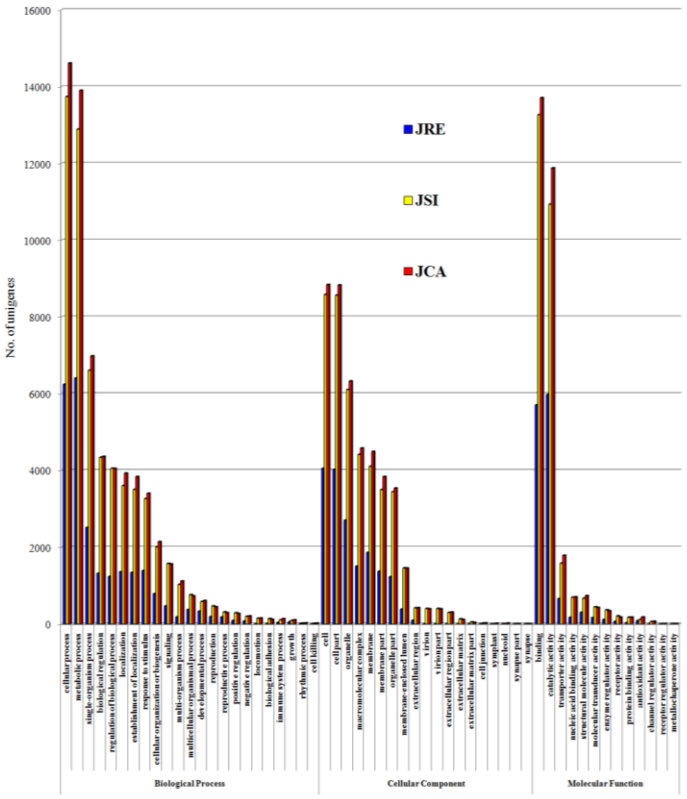

Fig 4. Functional classification of GO terms of unigenes from JRE, JSI and JCA.

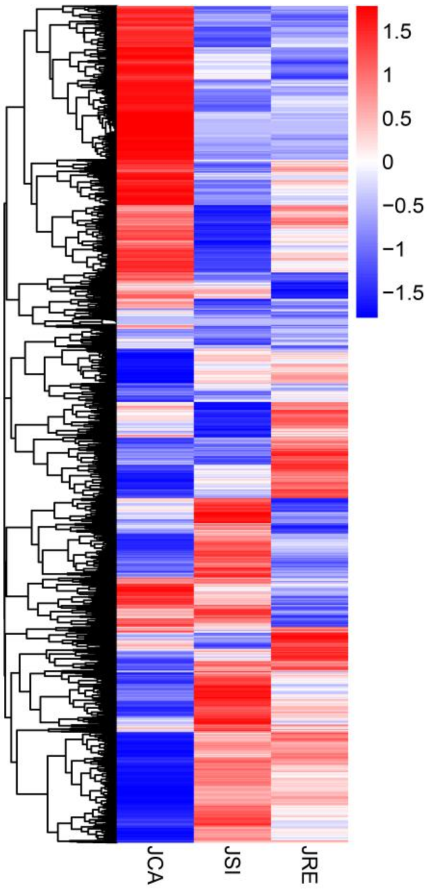

Fig 5. Hierarchical cluster analysis of differential expressed unigenes (DEGs) within three species. Expression differences are shown in different colours. The colour from blue to red means that transcript abundance of unigenes was from relatively low to relatively high. The tags and counts of DEGs were listed in supplementary file.

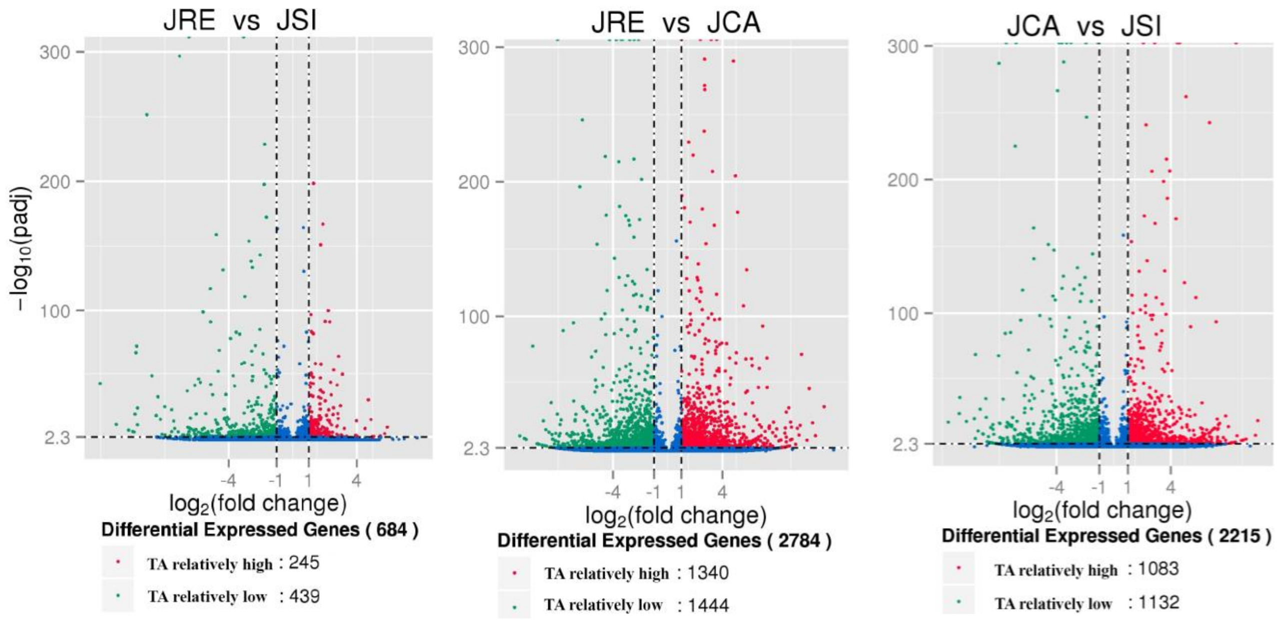

Fig 6. Volcano plot of differential expressed unigenes (DEGs) between two species. X axis represents the fold change (log transformed), and Y axis represents the p-value (log transformed). If the intensity ratio of a unigene between two species is more than 2 fold (**T**ranscript **A**bundance of unigenes was relatively high, red dot) or less than 2 fold (**TA** of unigenes was relatively low, green dot) having p-value less than 0.05, considered as a DEG. Those dots shown in blue are unigenes that did not show obviouschanges and did not identified as DEGs.

polymorphisms (SNPs) are highly polymorphic, easier to develop, and serve as a rich resource of diversity (Parchman et al., 2010; Ruperao and Edwards, 2015). To detect new molecular makers, all of the unigenes were used to mine potential SSRs motifs using the MISA software. In total, 34011 unigenes from 185912 sequences examined (116625163 bp) contained 41141 SSRs, from which 5835 unigenes had more than one SSR marker signature (Detailed statistics of SSRs available in Supplementary Table S5). The mono-nucleotide (19356, 47.05%), di-nucleotide (16103, 39.14%) and tri-nucleotide (5120, 12.45%) repeat motifs had the highest frequencies. All SSRs were further counted based on the number of repeat units (Fig 8). After designing and filtering primers, 19784 SSR markers were found to have at less one primer (6674 for JRE, 6077 for JSI and 7033 for JCA) (Supplementary Table S6, Table S7, Table S8). This data could lay a platform for better understanding the polymorphisms of *Juglance* species.

The GATK2 software was used to perform SNP calling (McKenna et al., 2010). In total, 61422, 91101 and 53832 candidate SNPs were identified in *J. regia*, *J. sigillata* and *J. cathayensis*, respectively (Table 5). The average SNP frequency was one SNP per 266 base pairs. The majority of SNPs (76.56%) were detected in transcripts ranging from 100 bp to 1100 bp. Howerer, the number of SNPs per transcripts increased with transcripts size, indicating that larger datasets with greater transcripts size could be used to identify more SNPs. The distribution of substitution types was shown in Table 5. A greater number of transitions (122839) than transversions (83516) were identified, and the ratio between transitions and transversions was 1.47. Among the transition, the number of C/T transitions was a little greater than that of G/A in JRE and JSI; however, the opposite has happened in JCA. Taken as a whole, A/T transversions were more infrequent than other three types among all species. Similar results were found in *J. regia* (Liao et al., 2014) and *Citrus clementina* (Terol et al., 2008). The availability of large numbers of SNPs should facilitate population genetics and gene-based association studies in *Juglans* species.

qPCR validation

The experimental validation has been done on 9 randomly selected transcripts (Supplementary Table S9) in order to confirm DEGs analysis of RNA-seq data by qPCR. Analysis results in leaves of three *Juglans* species confirmed the relative amounts of these transcripts observed with RNA-seq, with a high correlation (R^2 = 0.89) of fold change between RNA-seq and qPCR data (Fig. 9).

Materials and Methods

Plant materials and RNA extraction

In July 2014, fresh leaves of *J. regia*, *J. sigillata* and *J. cathayensis* were collected from the walnut germplasm resources garden of Yunnan Forestry Academy. Tissue samples were frozen in liquid nitrogen and stored at -80 ºC until RNA extraction.

Total RNA of plant materials was extracted using a RNeasy Mini kit (QIAGEN, Shanghai, China). The concentration and integrity of the RNA was measured and assessed using the Qubit 2.0 Flurometer (Life Technologies, CA, USA) and the Agilent Bioanaylzer 2100 system (Agilent Technologies, CA, USA). Finally, one of the best RNA sample (OD260/280 ≧ 1.8, concentration ≧ 100ng/μl, RIN (RNA Integrity Number) ≧ 8.0) for each material was used for the subsequent experiments.

Library construction and sequencing

A total amount of 3 μg total RNA per sample was used to construct the cDNA library using NEBNext Ultra™ RNA Library Prep Kit for Illumina (NEB, USA) following manufac-

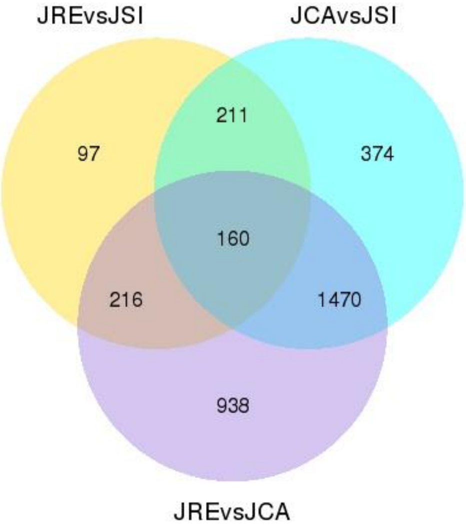

Fig7. Venn analysis of differentially expressed unigenes (DEGs) within JRE, JSI and JCA.

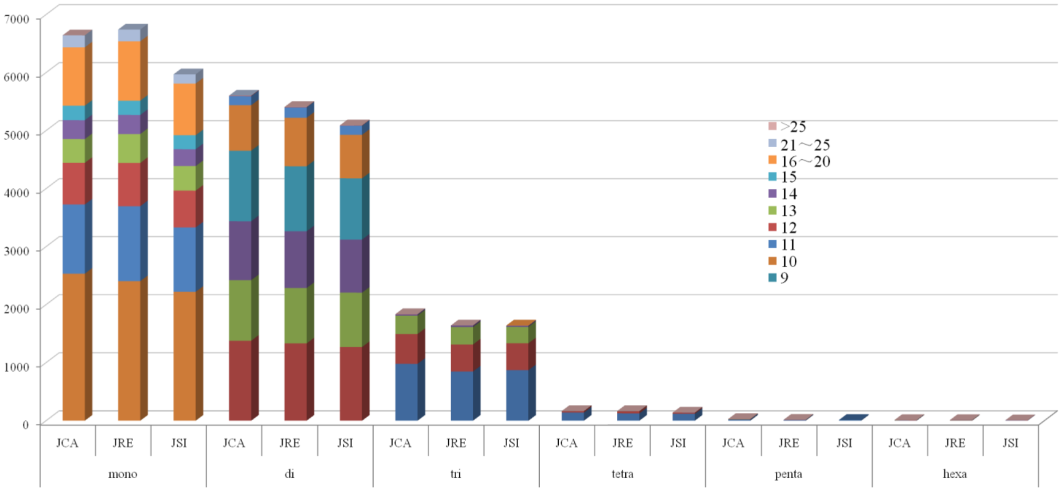

Fig 8. SSRs counts based on the number of motif repeat units.

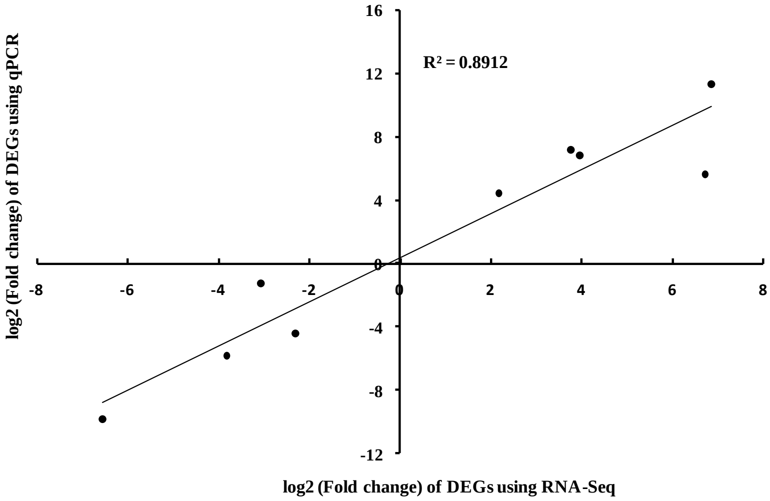

Fig 9. qPCR validation. DEGs detected through RNA-Seq in our study were validated using qPCR. Nine DEGs were randomly selected for validation. The correlation between RNA-Seq and qPCR was shown.

turer's recommendations and index codes were added to attribute sequences. Three cDNA libraries were constructed and sequenced on an Illumina Hiseq™ 2000 platform and paired-end reads were generated at Novogene Bioinformatics Technology Co. Ltd, Beijing, China (www.novogene.cn).

De novo assembly and gene annotation

Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. After filtering the raw reads, *de novo* assembly of the transcriptome was carried out with a short reads assembling program – Trinity (Grabherr et al., 2011). Trinity connects the transcripts and obtains sequences for as long as possible. Sequences not extended on either end were defined as unigenes.

The generated unigenes of *J. regia* (JRE), *J. sigillata* (JSI) and *J. cathayensis* (JCA) were each searched against the public databases for annotation, including the NCBI non-redundant protein sequences (NR) database, NCBI nucleotide sequences (NT) database, eukaryotic ortholog groups (KOG) database, KEGG ortholog (KO) database, Swiss-Prot protein database, Gene Ontology (GO) database, and protein family (PFAM) database. The CDS of unigenes was predicted according to the Wang's method (Wang et al., 2011).

Identification and enrichment analysis of differentially expressed genes (DEGs)

Prior to differential gene expression analysis, for each sequenced library, the read counts were adjusted by edgeR program package through one scaling normalized factor. Differential expression analysis of two samples was performed using the DEGseq R package (Wang et al., 2010). q value < 0.005 and |log2(fold change)| > 1 were set as the threshold for significantly differential expression. The statistical enrichment of DEGs in KEGG pathways was performed using the KOBAS software (Mao et al., 2005).

SSRs and SNPs marker identification

Using MISA software (http://pgrc.ipk-gatersleben.de/misa/) (Thiel et al., 2003), the potential SSR markers with motifs ranging from mono- to hexa-nucleotides in size were detected among the unigenes. The minimum of repeat units were set as follows: ten for mono-nucleotide, six for di- and five for tri-, tetra-, penta- and hexa-nucelotides. Primer pairs flanking each SSR loci were designed using the Primer3 program (http://primer3.ut.ee/). Sequencing data between samples were compared with the unigene database, using GATK2 software (McKenna et al., 2010), which builds consensus sequence, and then analyzes samples to get SNP loci.

qPCR validation

A set of nine genes, including five from relatively high abundance and four from relatively low abundance in RNA-seq analysis, was selected for real-time reverse transcription-PCR (qPCR) to validate the DEGs analysis results (Supplementary Table S9). Total RNA was extracted from the leaves using Trizol (Invitrogen), and then subjected to RQ1 RNase-free RDNase (Promega) digestion in order to remove any residual genomic DNA contamination. The first strand cDNA was synthesized by RevertAidTM First Strand cDNA Synthesis Kit (Fermentas) using the Oligo(dT)18 Primer performed as protocols. The primers used in qRT-PCR (see Supplementary Table S9) were designed using Primer Express 3.0 (Applied Biosystems) with melting temperatures of 58–60 ˚C and amplicon sizes of 100–200 bp. qPCR was performed with an ABI 7500 instrument (Applied Biosystems). *J. mandshurica* ribosomal protein L32 (*JmaRPL32*, accession number: HM466693) was selected as the reference gene (Bai et al., 2010). Data were collected, and threshold cycle (Ct) values were analyzed using $2^{-\Delta\Delta Ct}$ method.

Conclusion

This study reports the transcriptomes and their comparative analysis for two economically important nut crops and one potential breeding material: *J. regia* (JRE), *J. sigillata* (JSI) and *J. cathayensis* (JCA). We *de novo* assembled 192360 unigenes (60710 for JRE, 62415 for JSI and 69235 for JCA) with a mean size of 695 bp (712 for JRE, 700 for JSI and 674 for JCA), in which 48.2% (30952) of unigenes showed significant similarities to known sequences in protein databases. A total of 5683 unigenes were found to be differentially expressed genes (DEGs) within three *Juglans* species and the highest representation KEGG pathways of the DEGs was biosynthesis of amino acids (ko01230). We located and predicted 41141 SSRs and 206355 SNPs as potential molecular markers in our assembled and annotated sequences. Overall, these unigenes assembled and markers identified in this study will serve as useful genomic resource for mining and cloning interested genes and understanding the polymorphisms of *Juglance* species.

Acknowledgments

This work was supported by the National Science & Technology Pillar Program (Grant No. 2011BAD46B01), Construction Plan for Science & Technology Innovation Platform of Yunnan Province (2013DH007) and Program for Science & Technology Innovation Talents (2012HC009).

References

Bai WN, Liao WJ, Zhang DY. 2010. Nuclear and chloroplast DNA phylogeography reveal two refuge areas with asymmetrical gene flow in a temperate walnut tree from East Asia. New Phytol. 188 (3), 892-901.

Bennett MD, Leitch IJ (2012) Angiosperm DNA C-values database (release 8.0, Dec. 2012). http://www.kew.org/cvalues/

Britton MT, Leslie CA, Caboni E, Dandekar AM, McGranahan GH (2008) Persian Walnut. In: Chittaranjan K and Timothy CH (ed) Compendium of transgenic crop plants: transgenic temperate fruits and nuts. Wiley-Blackwell, Massachusetts.

Dandekar A, Leslie C, McGranahan G (2005) *Juglans regia* walnut. In: Litz RE (ed). Biotechnology of fruit and nut crops (Biotechnology in Agriculture Series, No. 29). Cromwell, Trowbridge.

Dode LA (1909) Contribution to the study of the genus *Juglans*. Bull Soc Dendrologique de France. 11: 22-90.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29: 644-652.

Iorizzo M, Senalik D, Grzebelus D, Bowman M, Cavagnaro P, Matvienko M, Ashrafi H, Van Deynze A, Simon PW (2011) De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. BMC Genomics. 12(1): 389.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36: D480-484.

Liao Z, Chen Y, Dai X, Li S, Yin T (2014) Genome-wide discovery and analysis of single nucleotide polymorphisms and insertions/ deletions in *Juglans regia* by high-throughput pyrosequencing. Plant Omics. 7: 445-449.

Manning WE (1978) The classification within the Juglandaceae. Ann Missouri Bot Garden. 65:1058-1087.

Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics. 21: 3787-3793.

Martínez ML, Labuckas DO, Lamarque AL, Maestri DM (2010) Walnut (J*uglans regia* L.): genetic resources, chemistry, by-products. J Sci Food Agr. 90:1959-1967.

McGranahan GH, Leslie CA (1990) Walnut (*Juglans* L.). In: Moore JN, Ballington JR (ed) Genetic resources of temperate fruit and nut crops, vol 2. International Society for Horticultural Science, Wageningen.

McGranahan GH, Leslie CA (2009) Breeding walnuts (*Juglans regia*). In: Jain SM, Priyadarshan PM (ed) Breeding plantation tree crops: temperate species. Springer, New York.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297-1303.

Metzker ML (2010) Sequencing technologies-the next generation. Nat Rev Genet. 11: 31-46.

Parchman TL, Geist KS, Grahnen JA Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics. 11:180.

Pehtea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003) TIGR gene indices clustering tools(TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 19: 651-652.

Ruperao P, Edwards D (2015) Bioinformatics: Identification of Markers from Next-Generation Sequence Data. In: Batley J (ed) Plant Genotyping: Methods and Protocols - Methods in Molecular Biology (Vol.1245). Springer, New York.

Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T, Wan XC (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics. 12: 131.

Shu C, McElroy JS, Dane F, Peatman E (2015) Optimizing transcriptome assemblies for *Eleusine indica* leaf and seedling by combining multiple assemblies from three *de novo* assemblers. Plant Genome. 8(1): 1-10.

Terol J, Naranjo MA, Ollitrault P, Talon M (2008) Development of genomic resources for *Citrus clementina*: characterization of three deep-coverage BAC libraries and analysis of 46000 BAC end sequences. BMC Genomics. 9: 423.

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST database for the development and characterization of gen-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 106: 411–422.

Wang H, Pan G, Ma Q, Zhang J, Pei D (2015) The genetic diversity and introgression of *Juglans regia* and *Juglans sigillata* in Tibet as revealed by SSR markers. Tree Genet Genomes. 11: 804.

Wang L, Feng Z, Wang X, Wang X, Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 26: 136-138.

Wang XW, Luan JB, Li JM, Su YL, Xia J, Liu SS (2011) Transcriptome analysis and comparison reveal divergence between two invasive whitefly cryptic species. BMC Genomics. 12: 458.

Xia Z, Xu H, Zhai J, Li D, Luo H, He C, Huang X (2011) RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*. Plant Mol Biol. 77(3): 299-308.