DENDROBIOLOGY

2019, vol. 81, 31–39

http://dx.doi.org/10.12657/denbio.081.004

*Huaizhi Mu, Yuting Gao, Fushan Cheng, Lin Lin, Gerong Wang, Fucai Xia**

Transcriptomic analysis of different tissues in Korean arborvitae

Received: 13 May 2017; Accepted: 9 April 2019

Abstract: Korean arborvitae (*Thuja koraiensis*) is an evergreen shrub or small tree native to Korea and the extreme Changbai Mountain of China threatened by habitat loss. Due to the limited genomic sources of Korean arborvitae, it is important to explore transcriptome to understand this economically important plant. We used RNA-seq technology to characterize the transcriptome of root, stem and leaf in Korean arborvitae, respectively. Based on the bioinformatics analysis of assembled transcriptome data, transcriptional pathways and differentially expressed genes (DEGs) were identified.
There were 152.26 million reads generated, and 446,568 unigenes with a mean size of 423.51 nt obtained via *de novo* assembly. Of these, 204,091 unigenes (45.70%) were further annotated by comparison to public protein databases. A total of 63,495 unigenes (14.22%) were associated into 130 pathways by searching against the KEGG database. DEGs analysis identified 11,890, 5,900 and 10,136 DEGs from the comparison of root vs. stem, root vs. leaf and stem vs. leaf, respectively. Besides, photosynthesis, plant hormone signal transduction and metabolism and biosynthesis of sugar and amino acids were changed in different tissues. We obtained 446,568 unigenes in Korean arborvitae and 11,890, 5,900 and 10,136 DEGs from the comparison of root vs. stem, root vs. leaf and stem vs. leaf, respectively. These results will aid in understanding and carrying out future studies on the molecular basis of Korean arborvitae and contribute to future artificial production and applications.

Keywords: *Thuja koraiensis*, transcriptome, tissue specific

Address: H. Z. Mu, Key Laboratory of State Forestry Administration on Conservation and Efficient Utilization of Precious and Rare Forest Resource in Changbai Mountain, Forestry College, Beihua University, 132013, Jilin, China, e-mail: huaizhimu@126.com
Y. T. Gao, Key Laboratory of State Forestry Administration on Conservation and Efficient Utilization of Precious and Rare Forest Resource in Changbai Mountain, Forestry College, Beihua University, 132013, Jilin, China, e-mail: 1758101445@qq.com
F. S. Cheng, Key Laboratory of State Forestry Administration on Conservation and Efficient Utilization of Precious and Rare Forest Resource in Changbai Mountain, Forestry College, Beihua University, 132013, Jilin, China, e-mail: 120019304@qq.com
L. Lin, Key Laboratory of State Forestry Administration on Conservation and Efficient Utilization of Precious and Rare Forest Resource in Changbai Mountain, Forestry College, Beihua University, 132013, Jilin, China, e-mail: linlin198212@126.com
G. R. Wang, Key Laboratory of State Forestry Administration on Conservation and Efficient Utilization of Precious and Rare Forest Resource in Changbai Mountain, Forestry College, Beihua University, 132013, Jilin, China, e-mail: 313003488@qq.com
F. C. Xia, Key Laboratory of State Forestry Administration on Conservation and Efficient Utilization of Precious and Rare Forest Resource in Changbai Mountain, Forestry College, Beihua University, 132013, Jilin, China, e-mail: xfc0707@163.com
*corresponding author

Introduction

Korean arborvitae (*Thuja koraiensis*) is a species of Cupressaceae, an evergreen shrub or small tree native to Korea and the extreme the Changbai Mountain of China (Wang et al., 2017). It is a valued medicinal tree species that is rich in terpenoids, flavonoids and essential oil, and as a source of lumber for sculpture (Wang et al., 2007; Zhang et al., 2014). In addition, it is occasionally grown as an ornamental tree for the contrast between the green upper and bright white lower sides of the foliage (Chung et al., 2011). The small population in China is protected in the Changbai Mountain Nature Reserve, as is the small population in Soraksan Nature Reserve in northern South Korea, but most of the species range in North Korea is unprotected and threatened by habitat loss (Yin et al., 2016). Studies of Korean arborvitae have mainly focused on propagation, chemical composition and biogeography, with only few reports relating to molecular biology (Yang et al., 1994; Li & Xiang, 2005; Peng & Wang, 2008; Yin et al., 2013). Due to the limited genomic sources of Korean arborvitae, it is important to explore transcriptome to understand this economically important plant.

In the past, large-scale analyses of transcript enrichment largely depend on hybridization, such as cDNA and oligo-nucleotide arrays (Alves-Ferreira et al., 2007; Yuan et al., 2012). However, array analyses and other hybridization-based approaches have several limitations, including knowledge of genes for probe design, non-specific hybridization, and difficulty in detecting low level expression (Marioni et al., 2008). On the other hand, more recently developed RNA sequencing (RNA-seq) technologies can overcome such limitations of hybridization-based approaches and other conventional large-scale gene expression analysis methods (Garber et al., 2011). It also has a highest great sensitivity, allowing the detection of transcripts with lower expression levels, such as those of many transcription factors (Mu et al., 2013; Zhang et al., 2015). In the last few years, RNA-seq has been extensively applied in the characterization of transcriptomes regarding developmental stage, organ, even specific cell types or single cell level, from yeast to human, including several plant species, such as *de novo* transcriptome sequencing of gymnosperms (Hao et al., 2011; Li et al., 2014; Zhang et al., 2016).

In the present work, we used RNA-seq technology to characterize the transcriptome of root, stem and leaf in Korean arborvitae. The RNA samples from these three different tissues were sequenced with the high-throughput Illumina deep sequencing technique. Based on the bioinformatics analysis of assembled transcriptome data, we characterized transcriptional pathways on the different tissues of Korean arborvitae. Furthermore, we identified the differentially expressed genes (DEGs) subject to regulation during different tissues. The transcriptome sequencing may help to discover new genes and pathways and may provide helpful insights into Korean arborvitae.

Materials and Methods

Plant material

Cuttings from Korean arborvitae for cuttage was collected from Changbai County Forest Farm (41.75°N, 127.93°E), Jilin Province. Cuttings were cut in black soil for seedlings regeneration and hundreds of seedlings were transplanted into plastic pots in the greenhouse. Root, stem and leaf from three different individuals were collected and mixed into three samples, and then the three samples were immediately frozen in liquid nitrogen respectively and stored at −80°C for future use.

RNA extraction, library construction and RNA-Seq

Total RNA was extracted separately from each sample using the R6827-01 Plant RNA Kit (Guduo, Shanghai, China). The concentration of RNA was analyzed using a spectrophotometer (UV-Vis Spectrophotometer, Quawell Q5000; Quawell, San Jose, CA, USA), and the integrity of RNA was evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Equal quantities of high-quality RNA of each sample from different tissues were combined into a single large pool for cDNA synthesis.

The mRNA-seq library was constructed using TruSeq RNA Sample Preparation Kit (Illumina Inc, San Diego, CA, USA). The poly-A mRNA was enriched using oligo (dT) magnetic beads, and the mRNA was broken into fragments by fragmentation buffer. The cleaved RNA fragments were transcribed into first-strand cDNA using random hexamer primers, followed by second strand cDNA synthesis using DNA polymerase I and RNase H. The short fragments were purified with the QiaQuick PCR Purification Kit (Qiagen) and eluted in EB buffer for end-repaired by addition of poly(A) to 3'. Then the suitable fragments were separated by an agarose gel electrophoresis and selected for PCR amplification as sequencing templates. The constructed mRNA-seq library was sequenced on the Illumina HiSeq™ 2500 sequencing platform.

De novo assembly and functional annotation

To obtain high-quality clean data for *de novo* assembly, the raw reads were filtered by removing the adapter sequences, low quality sequences (reads with ambiguous nucleotides "N"), and reads in which more than 20% of nucleotides had a *Q*-value <30. Reads were *de novo* assembled using the Bowtie and RSEM packages (Langmead et al., 2009; Li & Colin, 2011). The clean reads were assembled into contigs using Trinity (Grabherr et al., 2011). After Trinity *de novo* assembly and correction, the contigs without any gaps were linked into transcripts according to the paired-end information of the sequences. Related contigs were clustered into transcripts based on nucleotide sequence identity. The longest transcript was regarded as unigene redundancy was removed. Finally, the unigenes were combined to produce the final assembly used for annotation.

To determine the functional annotation of the unigenes, the assembled sequences were compared against the NCBI Nr (Deng et al., 2006), SwissProt (Apweiler et al., 2004), COG (Tatusov et al., 2000), KOG (Koonin et al., 2004), EggNOG (Huerta-Cepas et al., 2015), Pfam (Finn et al., 2014), GO (Ashburner et al., 2000) and KEGG (Kanehisa et al., 2004) with an *E*-value ≤ 10^{-5}. Gene names were assigned based on the best Blast hit (Altschul et al., 1997). Open reading frames (ORFs) were predicted using the GetORF program from the EMBOSS suite (Rice et al., 2000). The longest ORF extracted from each unigene was defined as coding sequence (CDS), and the CDSs were translated into amino acid sequences using the standard codon table. The Blast2GO program (Conesa et al., 2005) was applied to obtain GO annotation of unigenes including molecular function, biological process, and cellular component categories. The unigenes sequences were aligned to EggNOG database to classify and predict possible functions. Annotations of Korean arborvitae unigenes were used to predict biochemical pathways using the pathways tools. The KEGG database was used to analyze gene products related to metabolism and gene function in cellular processes.

Simple sequence repeats (SSRs) markers detection

The assembled sequences longer than 1 kb were used for SSRs markers detection. Potential SSRs markers were detected among the 29,240 unigenes using the MISA software (Beier et al., 2017). The parameters were set for the identification of perfect mono-nucleotide motifs with a minimum of ten repeats, di-nucleotide motifs with a minimum of six repeats, and tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum of five repeats (Zeng et al., 2010; Wei et al., 2011).

Differentially expressed genes (DEGs) identification

The gene expression abundance was represented in fragments per kilobase of exon model per million mapped fragments (FPKM). The FPKM measure of fragment density reflects the molar concentration of a transcript for RNA length and for the total fragment number in the measurement (Trapnell et al., 2010). The EBSeq package (Leng et al., 2013) was used to analyze differential gene expression. The FDR (false discovery rate) method (Benjamini et al., 2001) was used to determine the threshold of *P*-values in multiple tests. The $\log_2$ratio was the FPKM ratios of a gene between two samples returned the logarithm based on 2. In this study, FDR ≤ 0.01 and $|\log_2 \text{ratio}| \geq 2$ were the thresholds employed to judge the significance of differentiated gene expression (Zhang et al., 2016).

Pathway enrichment analysis of DEGs

The Blastall program in Blast package (Johnson et al., 2008) was used to annotate the pathways of DEGs against the KEGG database. The hypergeometric test was then used to identify significantly enriched pathways in DEGs compared to the transcriptome background. The formula used to calculate the *P*-value was as follows:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where *N* is the number of genes with KEGG annotation, *n* is the number of DEGs in *N*, *M* is the number of genes annotated to specific pathways and *m* is number of DEGs in *M*. The calculated *P*-value was subjected to Bonferroni correction, taking the corrected *P*-value ≤ 0.05 as a threshold. Pathways fulfilling this condition were defined as significantly enriched pathways in the DEGs.

Results and Discussion

RNA-Seq and *de novo* assembly

To obtain a global overview of the Korean arborvitae transcriptome of different tissues, three RNA samples from root, stem and leaf were sequenced with Illumina HiSeq™ 2500. The raw sequencing

Table 1. Quality of RNA-Seq for Korean arborvitae

Sample	Nucleotide number	Read number	Q30 percentage (%)	GC percentage (%)
Root	13,414,397,440	45,288,733	92.44	44.54
Stem	16,131,778,134	54,437,219	92.21	44.11
Leaf	15,637,002,320	52,536,454	92.43	43.64

Q30 percentage indicates the proportion of nucleotides with quality values greater than 30.

Table 2. Length distribution of *de novo* assembled transcripts and unigenes

Nucleotide length (nt)	Transcripts		Unigenes	
	Number	Percentage (%)	Number	Percentage (%)
200–300	308,801	52.52	279,779	62.65
300–500	119,899	20.39	93,935	21.03
500–1000	73,520	12.51	43,613	9.77
1000–2000	48,246	8.21	19,081	4.27
2000+	37,452	6.37	10,159	2.27
Total	587,920	100.00	446,568	100.00
N50 length (nt)	1,122		440	
Mean length (nt)	609.18		423.51	
Total length (nt)	358,150,061		189,126,994	

N50 indicates that half of the assembled nucleotides were incorporated into unigenes with a length at least.

data of root, stem and leaf were submitted to the SRA database in NCBI with the accession number of SRR8791012, SRR8791011 and SRR8791010, respectively. After stringent quality assessment and data filtering, a total of 152.26 million reads and 45.18 giga nucleotides were generated, and the Q30 percentages were all over 92% (Table 1).

Reads were mapped against the reference transcriptome sequences from the three samples of Korean arborvitae. After the removal of adaptor sequences and exclusion of contaminated or short reads, next-generation short-read sequences were assembled into 587,920 transcripts with mean length of 609.18 nt and N50 length of 1,122 nt using the Trinity *de novo* assembly program. The transcripts were subjected to cluster and assembly analyses. Finally 446,568 unigenes with a mean size of 423.51 nt were obtained, these included 29,240 unigenes (6.54%) with length greater than 1 kb. An overview of the transcripts and unigenes is shown in Table 2.

Functional annotation and classification

The reads of Korean arborvitae in root, stem and leaf were assembled, and several complementary approaches were utilized to annotate the assembled sequences (Fig. 1). Approximately 45.70% of the unigenes (204,091) were able to be annotated based on aligning with sequences deposited in diverse protein databases, including Nr (186,777 unigenes; 41.82%), SwissProt (186,777 unigenes; 41.82%), COG (67,243 unigenes; 15.06%), KOG (99,723 unigenes; 22.33%), EggNOG (177,785 unigenes; 39.81%), Pfam (112,323 unigenes; 25.15%), GO (110,728 unigenes; 24.80%) and KEGG (63,495 unigenes; 14.22%). We found that for 5.76% of the unigenes the most similar proteins sequence was from *Picea sitchensis*, whereas 2.44% were most similar to sequences from *Physcomitrella patens*, and 1.78% to *Amborella trichopoda*. Among the unigenes 5.76% appeared to be most closely related to genes from *P. sitchensis*. Korean arborvitae and *P. sitchensis* are gymnosperms that belong to the Coniferopsida Coniferae. Thus, there is a close relationship between Korean arborvitae and *P. sitchensis* based on both systematic botany and molecular analysis. Compared with those from other conifer trees, our results using samples from different tissues including root, stem and leaf respectively, identified a much larger number of unigenes (Hao et al., 2011; Li et al., 2014; Zhang et al., 2016). The results obtained in this research demonstrated that our final assembly quality was satisfactory and it therefore provides sequence resources and facilitates further gene cloning and functional analyses.

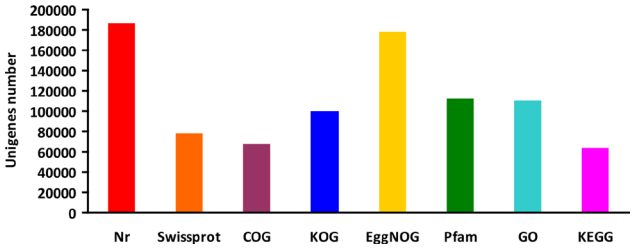

Fig. 1. Functional annotation of Korean arborvitae unigenes

Gene Ontology (GO) analysis was used for functional classification of the assembled transcripts and gene products in terms of their likely associated biological processes, cellular components, and molecular functions. There were 110,728 unigenes were assigned to GO terms (Fig. 2). To better review

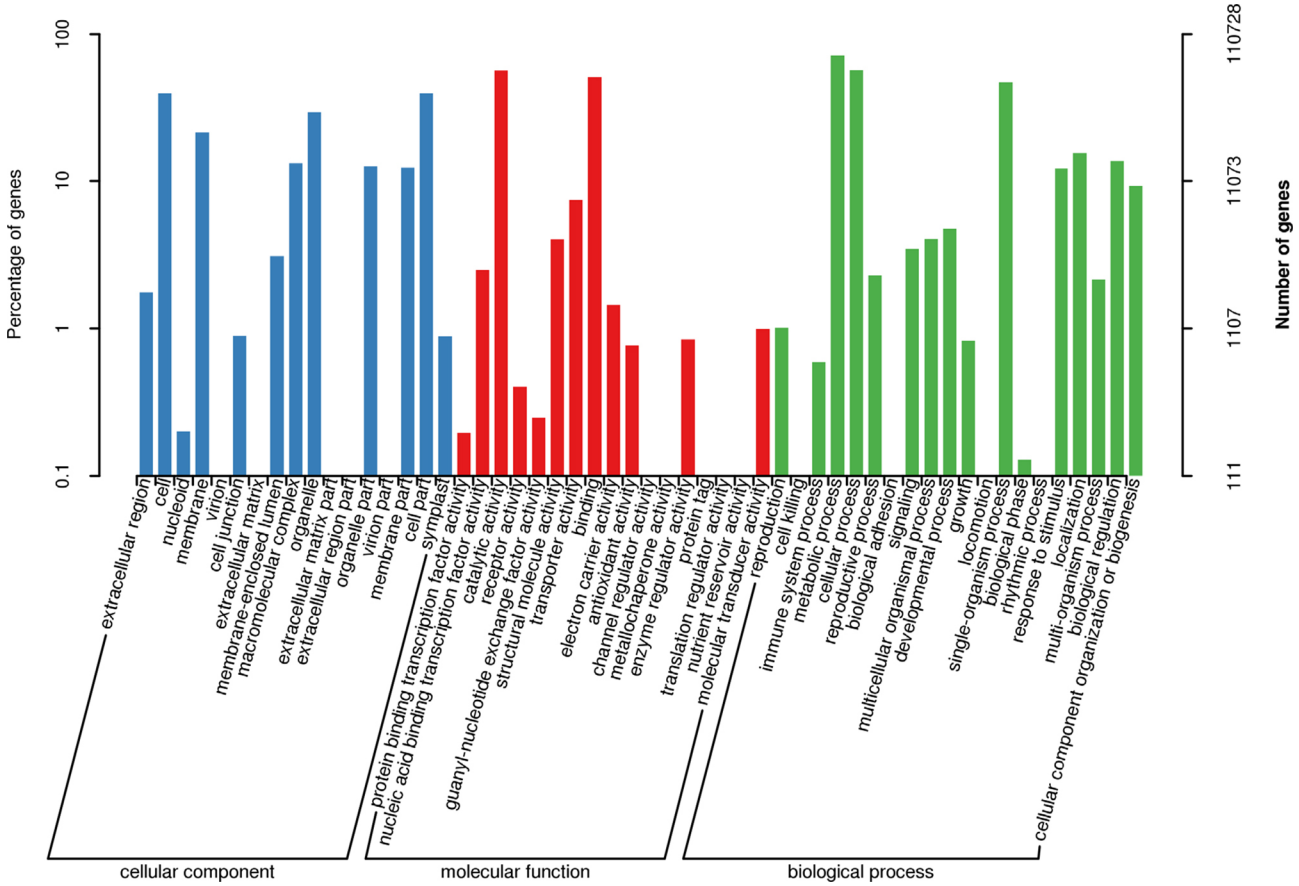

Fig. 2. Functional annotation of assembled sequences based on GO categorization

GO cellular components, the GO terms were further clustered to their parent terms. With regard to biological processes, metabolic processes (79,244 unigenes; 71.56%), cellular processes (63,420 unigenes; 57.28%) and single-organism process (51,993 unigenes; 46.96%) were prominently represented. For molecular function, catalytic activity (62,940 unigenes; 56.84%) and binding (56,740 unigenes; 51.24%) represented the majorities of this category. Cell (43,924 unigenes; 39.67%), membrane (23,890 unigenes; 21.58%), organelle (32,889 unigenes; 29.70%) and cell part (43,924 unigenes; 39.67%) represented a high percentage of the cellular component category.

In addition, all unigenes were aligned to the EggNOG database for further functional prediction and classification. Overall, 177,785 of the 446,568 sequences were assigned to 25 EggNOG categories (Fig. 3). The category of function unknown represented the largest group (35,544 unigenes; 19.99%), followed by general function prediction only (31,624 unigenes; 17.79%) and posttranslational modification, protein turnover, chaperones (15,112 unigenes; 8.50%). Only a few unigenes were assigned to cell motility, extracellular structures and nuclear structure (67, 351 and 513 unigenes, respectively). Furthermore, 7,788 unigenes were assigned to carbohydrate transport and metabolism and 2,321 unigenes were assigned to coenzyme transport and metabolism.

KEGG is a public database for networks of molecular interactions in cells and their variants specific to particular organisms. To further examine the usefulness of the Korean arborvitae unigenes generated in the present study, the unigenes were compared with the KEGG database using Blastx and the corresponding pathways were established. Only 63,495 unigenes (14.22%) were assigned to 130 pathways. The pathways with highest unigene representation were ribosome (ko03010; 4327 unigenes; 6.81%), followed by carbon metabolism (ko01200; 3759 unigenes; 5.92%) and biosynthesis of amino acids (ko01230; 3053 unigenes; 4.81%).

SSRs markers discovery

SSRs can be used as powerful molecular markers for genetics, evolution and breeding studies. To explore SSR profiles in the unigenes of Korean arborvitae, the 29,240 unigene sequences were searched for SSRs. In total, 4,424 sequences containing 4,962 SSRs were obtained, with 459 unigene sequences containing more than one SSR. Mono-nucleotide repeat motifs (58.61%) were the most abundant, followed by di-nucleotide repeats (19.97%) and

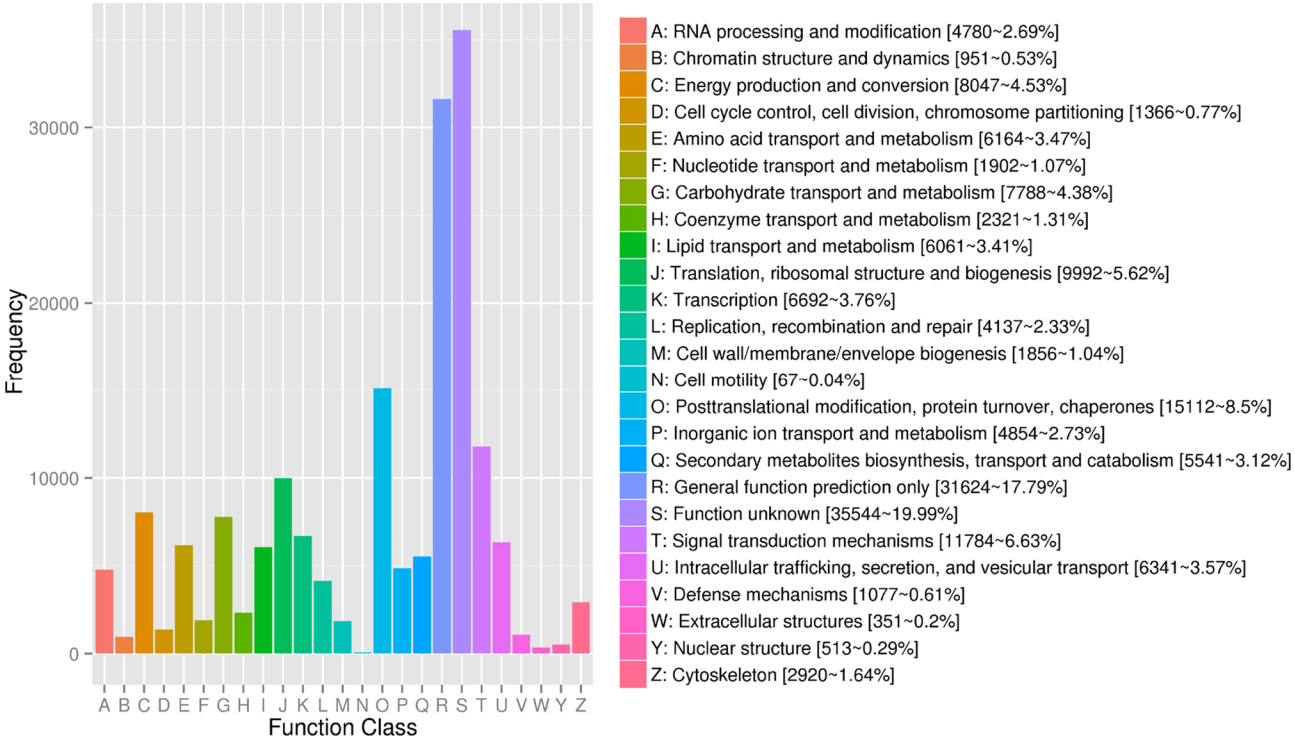

Fig. 3. EggNOG function classification of consensus sequence

Table 3. Frequency of candidate SSRs in Korean arborvitae

Motif		Mono	Di	Tri	Tetra	Penta	Hexa	Total	Percentage (%)
Repeat number	5	–	–	629	53	10	4	696	14.03
	6	–	350	238	6	1	3	598	12.05
	7	–	185	100	0	0	1	286	5.76
	8	–	129	14	0	0	0	143	2.88
	9	–	112	1	0	0	0	113	2.28
	10	1,449	134	1	0	0	1	1,585	31.94
	11	522	78	0	0	0	0	600	12.09
	12	234	3	1	0	0	0	238	4.80
	13	140	0	0	0	0	0	140	2.82
	14	106	0	0	0	0	0	106	2.14
	15	83	0	0	0	0	0	83	1.67
	>15	374	0	0	0	0	0	374	7.54
Total		2,908	991	984	59	11	9	4,962	100.00
Percentage (%)		58.61	19.97	19.83	1.19	0.22	0.18	100.00	

tri-nucleotide repeats (19.83%) (Table 3). The most abundant repeat type was A/T (2,810; 56.63%), followed by GA/TC (214; 4.31%), and AG/CT (193; 3.89%).

Identification and pathway enrichment of DEGs

A total of 446,568 unigenes were detected from the clean reads of all samples as described above. To detect DEGs between the samples harvested from different tissues, DESeq was used with the criteria FDR ≤ 0.01 and $|\log_2 \text{ratio}| \geq 2$. The representative genes of up- and down-regulated DEGs of Korean arborvitae at different tissues are shown in Fig. 4. For root vs. stem comparison, most DEGs were up-regulated. With regard to stem vs. leaf comparison, most DEGs were down-regulated. A roughly similar number of up-regulated DEGs and down-regulated DEGs was produced in root vs. leaf comparison (Table 4).

To provide a global view of the biological pathways that function in Korean arborvitae, KEGG pathway analysis was used to identify the biological pathways of the DEGs. With regard to root vs. stem comparison, 2,990 DEGs were mapped into 119 pathways. For root vs. leaf comparison, 1,519 DEGs were mapped into 112 pathways. 2,565 DEGs mapped into 121 pathways were produced in stem vs. leaf comparison. Differences in the dynamics and

Table 4. Number of up- and down-regulated DEGs of Korean arborvitae in different tissues

Comparison	All DEGs number	Up-regulated DEGs number	Down-regulated DEGs number
Root vs. Stem	11,890	8,067	3,823
Root vs. Leaf	5,900	2,552	3,348
Stem vs. Leaf	10,316	2,148	8,168

Table 5. Pathway enrichment of DEGs of Korean arborvitae

Pathway	Root vs. Stem		Root vs. Leaf		Stem vs. Leaf	
	DEGs number	Corrected *P*-value	DEGs number	Corrected *P*-value	DEGs number	Corrected *P*-value
Photosynthesis	63	0	64	0	44	0
Phenylpropanoid biosynthesis	75	2.03E-09	30	9.26E-02	80	0
Ribosome	434	3.89E-09	256	0	353	0
Photosynthesis – antenna proteins	26	2.39E-08	26	0	22	9.86E-07
Carbon fixation in photosynthetic organisms	96	2.02E-06	40	6.30E-01	85	3.76E-06
Phenylalanine metabolism	63	3.69E-06	18	1	71	0
Glycolysis/Gluconeogenesis	129	2.88E-04	37	1	113	5.11E-04
Carotenoid biosynthesis	21	7.19E-03	11	2.61E-01	13	1
Pentose and glucuronate interconversions	50	1.39E-02	17	1	48	1.26E-03
Plant hormone signal transduction	39	3.03E-01	27	9.56E-03	43	8.15E-04
Protein processing in endoplasmic reticulum	153	1	28	1	153	1.58E-02
Plant-pathogen interaction	42	1	9	1	45	1.92E-02

absolute levels of DEGs expression in different tissues were seen in the following categories: photosynthesis (ko00195), phenylpropanoid biosynthesis (ko00940), ribosome (ko03010), photosynthesis – antenna proteins (ko00196), carbon fixation in photosynthetic organisms (ko00710), phenylalanine metabolism (ko00360), glycolysis/gluconeogenesis (ko00010), carotenoid biosynthesis (ko00906), pentose and glucuronate interconversions (ko00040), plant hormone signal transduction (ko04075), protein processing in endoplasmic reticulum (ko04141) and plant-pathogen interaction (ko04626) (Table 5).

Differences in gene expression profiles can yield insight into mechanisms underlying physiological changes, and DEGs were found among different tissues, treatments, and species (Feng et al., 2012; Lin et al., 2013; Hu et al., 2016; Gaete-Loyola et al., 2017). The root, stem and leaf samples were collected represented the different tissues in Korean arborvitae, and there were fewer DEGs in root vs. leaf. However, many 11,890 DEGs and 10,316 DEGs were detected in root vs. stem and stem vs. leaf, respectively. Furthermore, many DEGs were annotated to specific pathways using KEGG database, including photosynthesis, plant hormone signal transduction and metabolism and biosynthesis of sugar and amino acids. These results indicate considerable changes of gene expression in different tissues. A similar phenomenon was reported for the organ differentiation and formation in which expression key genes expression about photosynthesis, plant hormone and metabolism and biosynthesis pathways (Hao et al., 2011; Niu et al., 2015; Hu et al., 2016). We can conclude that these key genes

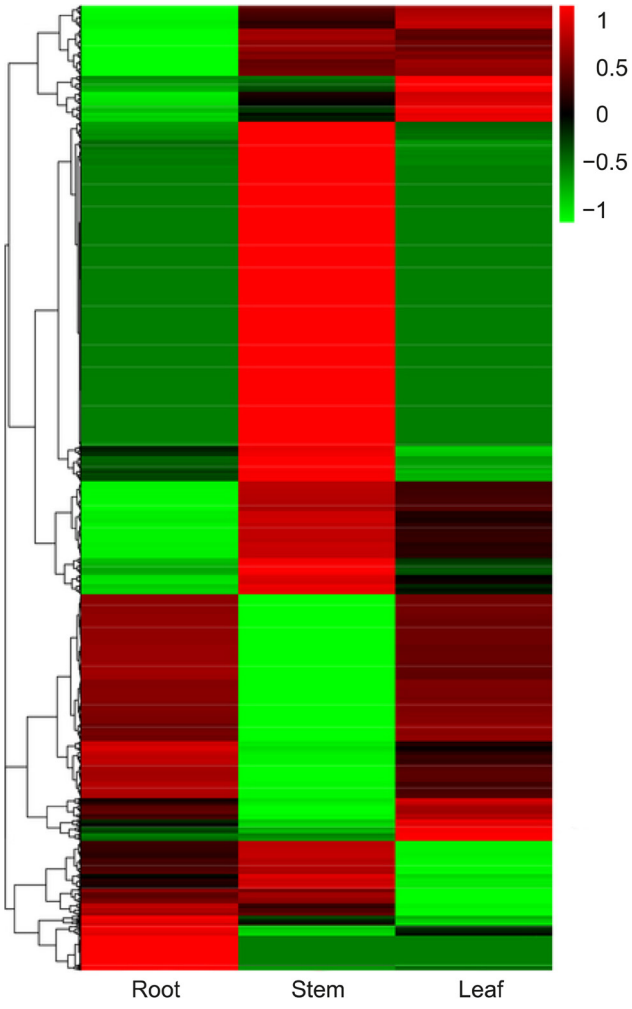

Fig. 4. Heatmap of the relative expression levels of DEGs

expressed preferentially in different tissues of Korean arborvitae may be essential for organ differentiation and formation, and which are involved in a broad range of physiological functions.

Acknowledgement

This work was financially supported by a grant from the National Key Research and Development Program of China (2017YFC0504102) and the Youth Training Foundation of Beihua University (2017QNJJL07).

References

Altschul SF, Madden TL, Schäffer AA, Zhang JH, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.

Alves-Ferreira M, Wellmer F, Banhara A, Kumar V, Riechmann JL & Meyerowitz EM (2007) Global expression profiling applied to the analysis of *Arabidopsis* stamen development. Plant Physiology 145: 747–762.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N & Yeh LSL (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Research 32: D115–D119.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM & Sherlock G (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25–29.

Beier S, Thiel T, Münch T, Scholz U & Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33: 2583–2585.

Benjamini Y, Drai D, Elmer G, Kafkafi N & Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behavioural Brain Research 125: 279–284.

Chung IM, Praveen N & Ahmad A (2011) Composition of the essential oil and antioxidant activity of petroleum ether extract of *Thuja koraiensis*. Asian Journal of Chemistry 23: 3703–3706.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M & Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Deng YY, Li JQ, Wu SF, Zhu YP, Chen YW & He FC (2006) Integrated nr database in protein annotation system and its localization. Computer Engineering 32: 71–74.

Feng C, Chen M, Xu C, Bai L, Yin XR, Li X, Allan AC, Ferguson IB & Chen KS (2012) Transcriptomic analysis of Chinese bayberry (*Myrica rubra*) fruit development and ripening using RNA-Seq. BMC Genomics 13: 19.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J & Punta M (2014) Pfam: the protein families database. Nucleic Acids Research 42: D222–D230.

Gaete-Loyola J, Lagos C, Beltrán MF, Valenzuela S, Emhart V & Fernández M (2017) Transcriptome profiling of *Eucalyptus nitens* reveals deeper insight into the molecular mechanism of cold acclimation and deacclimation process. Tree Genetics & Genomes 13: 37.

Garber M, Grabherr MG, Guttman M & Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods 8: 469–477.

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. Nature Biotechnology 29: 644–652.

Hao DC, Ge G, Xiao P, Zhang Y & Yang L (2011) The first insight into the tissue specific *Taxus* transcriptome via Illumina second generation sequencing. PLoS ONE 6: e21220.

Hu Z, Zhang T, Gao XX, Wang Y, Zhang Q, Zhou HJ, Zhao GF, Wang ML, Woeste KE & Zhao P (2016) *De novo* assembly and characterization of the leaf, bud, and fruit transcriptome from the vulnerable tree *Juglans mandshurica* for the development of 20 new microsatellite markers using Illumina sequencing. Molecular Genetics and Genomics 291: 849–862.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C & Bork P (2015) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research 44: D286–D293.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S & Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Research 36: W5–W9.

Kanehisa M, Goto S, Kawashima S, Okuno Y & Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Research 32: D277–D280.

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ & Natale DA (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5: R7.

Langmead B, Trapnell C, Pop M & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM & Kendziorski C (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29: 1035–1043.

Li B & Colin ND (2011) RSEM: accurate transcript quantification from RNA Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

Li JH & Xiang QP (2005) Phylogeny and biogeography of *Thuja* L. (Cupressaceae), an eastern Asian and North American disjunct genus. Journal of Integrative Plant Biology 47: 651–659.

Li WF, Han SY, Qi LW & Zhang SG (2014) Transcriptome resources and genome-wide marker development for Japanese larch (*Larix kaempferi*). Frontiers of Agricultural Science and Engineering 1: 77–84.

Lin L, Mu HZ, Jiang J & Liu GF (2013) Transcriptomic analysis of purple leaf determination in birch. Gene 526: 251–258.

Marioni JC, Mason CE, Mane SM, Stephens M & Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 18: 1509–1517.

Mu HZ, Lin L, Liu GF & Jiang J (2013) Transcriptomic analysis of incised leaf-shape determination in birch. Gene 531: 263–269.

Niu J, Hou XY, Fang CL, An JY, Ha DL, Qiu L, Ju YX, Zhao HY, Du WZ, Qi J, Zhang ZX, Liu GN & Lin SZ (2015) Transcriptome analysis of distinct *Lindera glauca* tissues revealed the differences in the unigenes related to terpenoid biosynthesis. Gene 559: 22–30.

Peng D & Wang XQ (2008) Reticulate evolution in *Thuja* inferred from multiple gene sequences: Implications for the study of biogeographical disjunction between eastern Asia and North America. Molecular Phylogenetics and Evolution 47: 1190–1202.

Rice P, Longden I & Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends in Genetics 16: 276–277.

Tatusov RL, Galperin MY, Natale DA & Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33–36.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJ, Salzberg SL, Wold BJ & Pachter L (2010) Transcript assembly and quantification by RNA Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511–515.

Wang GR, Xia FC, Liu BD, Sun Y & Mu HZ (2017) Habitat and height growth rhythm of *Thuja koraiensis*. Journal of Beihua University (Natural Science) 18: 312–314.

Wang HS, Deng ZG, Huang ZZ & Li RX (2007) Structural research on secondary xylem of stem of *Thuja koraiensis* Nakai under SEM. Journal of Tonghua Teachers College 28: 6–13.

Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, Lv H & Zhang X (2011) Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics 12: 451.

Yang ZY, Tian ZL, Liu Q & Sun YH (1994) Studies on the chemical constituents of the volatile oil from leaves of *Thuja koraiensis* Nakai Maxim. Journal of Northeast Normal University (Natural Science) 26: 136–140.

Yin H, Zhao Y, Cui KF, Liu LJ, Yu CB, Chen QH, Dai YH & Zhao W (2013) Asexual reproduction technique of *Thuja koraiensis* Nakai. Chinese Wild Plant Resources 32: 68–69.

Yin H, Jin H, Zhao Y, Qin LW, Dai YH, Liu LJ & Zhao W (2016) Present situation and conservation strategy of rate and endangered species *Thuja koraiensis* in Changbai Mountain. Journal of Beihua University (Natural Science) 17: 40–42.

Yuan HM, Chen S, Lin L, Wei R, Li HY, Liu GF & Jiang J (2012) Genome-wide analysis of a *TaLEA*-introduced transgenic *Populus simonii* × *Populus nigra* dwarf mutant. International Journal of Molecular Sciences 13: 2744–2762.

Zhang LS, Wang L, Yang YL, Cui J, Chang F, Wang YX & Ma H (2015) Analysis of *Arabidopsis* floral transcriptome: detection of new florally expressed genes and expansion of Brassicaceae-specific gene families. Frontiers in Plant Science 5: 802.

Zhang XW, Choe YH, Park YJ & Kim BS (2014) Effect of Korean arbor vitae (*Thuja koraiensis*) extract on antimicrobial and antiviral activity. African Journal of Pharmacy and Pharmacology 8: 274–277.

Zhang YX, Han XJ, Sang J, He XL, Liu MY, Qiao GR, Zhuo RY, He GP & Hu JJ (2016) Transcriptome analysis of immature xylem in the Chinese fir at different developmental phases. PeerJ 4: e2097.

Zeng S, Xiao G, Guo J, Fei Z, Xu Y, Roe BA & Wang Y (2010) Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. BMC Genomics 11: 94.