Journal of Applied Botany and Food Quality 90, 18 - 24 (2017), DOI:10.5073/JABFQ.2017.090.004

[1]Department of Industrial Plant Science and Technology, College of Agricultural, Life and Environmental Sciences, Chungbuk National University, Cheongju, Republic of Korea
[2]Nature Environment Research Park of Gangwon Province, Hongcheon, Republic of Korea

De novo transcriptomic analysis to reveal functional genes involved in triterpenoid saponin biosynthesis in *Oplopanax elatus* NAKAI

Seung Hee Eom[1], Jae Kook Lee[1], Heekyu Kim[2], Tae Kyung Hyun[1,*]

(Received September 24, 2016)

Summary

Oplopanax elatus is a valuable medicinal plant in the family Araliaceae that recommends itself as source of herbal preparations. Although, as in ginseng, triterpenoid saponins make up the major bioactive component of *O. elatus*, nothing is known about the genes that are involved in the biosynthesis of these complex compounds, as reflected also by a lack of genomic information in public databases. Using Illumina paired-end sequencing technology, we have therefore generated a transcriptome library of *O. elatus* from a pooled RNA sample from different organs. 208,959 unigenes were assembled from approximately 77 million high-quality reads, and 110,202 unigenes (52.7 % of the unigenes) were annotated. In addition, 47,273 cDNA-derived SSRs in 38,446 unigenes were identified as potential molecular markers. Furthermore, 122 unigenes encoding 47 putative enzymes related to the biosynthesis of the backbone of triterpenoid saponins were identified by analyzing our library. The organ-specific expression of selected genes suggests that the leaves of *O. elatus* are the main site of triterpenoid saponin biosynthesis. The transcriptome data reported here provides valuable and comprehensive information for further research into the metabolic pathways of *O. elatus* as well as into genetic variation in the *Oplopanax* genus.

Introduction

The *Oplopanax* genus, a small genus in the family Araliaceae, consists of three species of deciduous shrubs, *Oplopanax horridus* (SMITH) Miq., which is called devil's club, *O. japonicas* (NAKAI), which is exclusively found in Japan, and *O. elatus* (NAKAI), which ranges from northeastern Asia to western North America (Calway et al., 2012; Huang et al., 2014). *Oplopanax* has traditionally been used as a medicine for ailments such as diabetes, rheumatoid arthritis and as an adaptogen to treat chronic fatigue syndrome in America, Russia, China and Korea (Semalty et al., 2013). In Asia, *O. elatus* has been regarded as an adaptogenic herb in the same vein as *Panax ginseng* and has been used to treat arthritis, diabetes mellitus, rheumatism, neurasthenia, and cardiovascular diseases (Lee et al., 2002). Bioactive compounds from the root, rhizome, and stem of *O. elatus*, including essential oils, saponins, flavonoids, anthraquinones, polyynes, and other metabolites, have been isolated and characterized (Shikov et al., 2014). In addition, a beneficial effect of *O. elatus* on reproductive function and other pharmacological activities such as adaptogenic, anticonvulsant, anti-diabetic, anti-fungal, anti-arthritis, anti-psoriasis and anti-oxidant effects with low toxicity in human and animals have been reported on the basis of different types of experimental studies and clinical data (Dou et al., 2009; Shikov et al., 2014), suggesting that *O. elatus* has potential as a crude drug and dietary health supplement.

Triterpenoid saponins are naturally occurring triterpene glycosides and form a class of common secondary metabolites in the plant kingdom (Moses et al., 2014; Seki et al., 2015). Although the physiological functions of triterpenoid saponins in higher plants are not fully understood, they have been shown to have a broad range of pharmacological properties, including immunogenic, anti-cholesterolemic, anti-diabetic, and anti-cancer activities (Man et al., 2010; Sun et al., 2010; Elekofehinti, 2015). In the case of *O. elatus*, it has been shown that isolated triterpenoid saponins reduce blood pressure in animals (Shikov et al., 2014). To date, twenty-one cirenshenosides, including the lupene-type triterpenoid saponins known as cirenshenosides A, E, F, G and H, the oleanene-type triterpenoid saponins known as cirenshenosides B, C, D, I, J, K and L, and cirenshenosides S, T, U and V, have been isolated from *O. elatus* (Shikov et al., 2014). While there are various studies of the occurrence, chemical structure, and biological activities of saponins in the *Oplopanax* genus, the enzymes and genes involved in the biosynthesis of these complex molecules are largely unknown.

In non-model plants, the genes of the biosynthetic pathways of triterpenoid saponins have been identified by a next generation sequencing (NGS) approach (Kalra et al., 2013; Zheng et al., 2014; Hwang et al., 2015; Zhang et al., 2015; Rai et al., 2016). In addition, RNA sequencing (RNA-seq) by means of NGS has shown great potential for transcriptome analysis, including transcript discovery, transcriptional structure of genes, quantification of expression levels, and functional gene mining (Schliesky et al., 2012). RNA-seq has also been used to identify genetic variation via single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) (Mutz et al., 2013). In this way, as no reference genome information is available for non-model plants, RNA-seq analysis has facilitated transcriptome characterization in hundreds of non-model plant species.

To analyze the triterpenoid saponin biosynthetic pathway in *O. elatus*, we firstly employed an RNA-seq approach to generate 77 million high-quality reads on an Illumina HiSeq^{TM} 2000 that were assembled into 208,959 unigenes. Comprehensive analysis indicated that most of the known genes encoding enzymes involved in the biosynthesis of the backbone of triterpenoid saponin, including the methyl-erythritol phosphate (MEP) pathway and mevalonate (MVA) pathway, were represented in the *O. elatus* transcriptome library. In addition, cDNA-derived SSRs (cSSRs) were identified to facilitate the analysis of genetic variations by marker-assisted selection. This novel dataset will be an important resource for further functional gene discovery and the analysis of transcriptional regulation and molecular markers in the *Oplopanax* genus, as well as in the family Araliaceae.

Material and methods

Sample collection and RNA isolation

Two year-old plantlets (*Oplopanax elatus* NAKAI) regenerated via somatic embryogenesis (Moon et al., 2013) were obtained from the Korea Forest Research Institute, Republic of Korea. Plants were grown under greenhouse conditions at 22 °C to 24 °C and 16 h light/8 h night cycle (light intensity: 180 µmol m^{-2} s^{-1}). Leaves, stems, and

* Corresponding author

roots were collected separately, immediately frozen in liquid nitrogen, and kept at -80 °C.
Total RNA was isolated from each sample using an RNEasy Plant Mini kit (Qiagen, USA) according to the manufacturer's instructions. Quantity of total RNA were verified with a Nanodrop 2000C spectrophotometer (Thermo Scientific, USA) and equal amounts of total RNA from each organ were mixed together for cDNA preparation. Quality (purity and integrity) was assessed by Agilent Bioanalyzer 2100 System (Agilent Technologies, California, USA).

Illumina sequencing and sequence assembly
For transcriptome analysis, mRNA from 20 µg total pooled RNA was isolated by means of oligo(dT) magnetic beads and fragmented into short pieces. Then, a cDNA library was generated, as previously described (Hyun et al., 2014), and sequenced on the Illumina HiSeq^{TM}2000 sequencing platform. After removing adapter sequences, empty reads, low-quality reads (i.e., those with ambiguous sequences 'N'), and reads with more than 10 % Q < 20 bases (i.e., with a base quality of less than 20) in the standard Illumina pipeline including the CASSAVA program, 78,646,554 clean paired-end reads were used for *de novo* transcriptome assembly. *De novo* assembly was carried out with short reads assembling program, Trinity (assembly parameter, k-mer value = 25, CPU = 25) (Grabherr et al., 2011). The Illumina reads generated in this study are available from the National Agricultural Biotechnology Information Center (NABIC, http://nabic.rda.go.kr) with accession number NN-2578.
Unigenes were identified by sequence comparisons with the NCBI non-redundant protein (NR) database (http://www.ncbi.nlm.nih.gov) and the Cluster of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG) by BLAST searches with a cut-off E-value of 10^{-5}.

Mining of simple sequence repeats (SSRs)
All unique sequences were analyzed with the MISA (Microsatellite) Perl script (http://pgrc.ipk-gatersleben.de/misa) to determine the composition, frequency, and distribution of simple sequence repeats (SSRs) (Wei et al., 2014). The search criteria of the MISA script included mono-nucleotides that were repeated more than 10 times, di-nucleotides that were repeated more than 6 times, and tri-, tetra-, penta- and hexa- nucleotides that were repeated more than 5 times.

Expression analysis of genes involved in the biosynthesis of the backbone of triterpenoid saponin
Total RNA isolated from leaves, stem, and roots was reverse-transcribed into cDNA using the QuantiTect® Reverse Transcription Kit (Qiagen, USA) in accordance with the manufacturer's recommendations. qReal-time PCR was performed with a CFX Connect^{TM} Real-Time PCR Detection System (Bio-Rad, USA) with SYBR Green Real-time PCR Master Mix (Toyobo, Japan). The expression levels of different genes were normalized to the constitutive expression level of elongation factor 1-alpha. The specific primer pairs used for real-time PCR are shown in Tab. S1. Three biological replicates were used for all samples.

Results

***De novo* assembly and functional annotation of *O. elatus* transcriptome**
To generate a transcriptome library of *O. elatus*, total RNA was extracted from different organs, including leaves, stem, and root. A pooled RNA sample was prepared by combining equivalent RNA from different organs. Using pair-end read with the Illumina platform, a total of 78,646,554 raw sequencing reads were generated. After data cleaning, approximately 77 million high-quality reads were obtained with 98 % G20 (base quality of greater than 20). The high-quality reads were assembled into 208,959 unigenes with an average length of 1,073 bp and an N50 of 1,768 bp (Tab. 1). While the majority of the unigenes were between 200 to 400 bp in length, we were able to obtain 132,214 unigenes (63.27 % of unigenes) which were longer than 500 bp (Fig. S1). The distribution of GC content frequency was from 23 % to 67 %, with an optimum of 44 %.
To validate and annotate the unigenes, Blastx was used to compare all assembled unigenes with the NCBI non-redundant (NR) database and InterProScan, and 110,202 unigenes (52.7 % of the unigenes) with significant similarity (cut-off E-value of 10^{-5}) were annotated (Tab. S2). In order to predict physiological functions, the assembled unigenes were analyzed according to Cluster of Orthologous Groups (COG). Of these 25 COG categories, "General function only" represented the largest cluster (24,466 unigenes), followed by "Posttranslational modification, protein turnover, chaperones" (12,301 unigenes) and "Signal transduction mechanisms" (12,027 unigenes), whereas "Cell motility" (47 unigenes) represented the smallest groups (Fig. 1). In addition, our transcriptome library yielded 3968 unigenes that, when annotated with COG database, represented the "Secondary metabolites biosynthesis, transport and catabolism" category, indicating that our transcriptome library provides a valuable resource for investigating biological processes and functional pathways in *O. elatus*.

Tab. 1: Overview of the sequencing and assembly.

	Oplopanax elatus Nakai
Sequences before filtering	
Total nucleotides (nt)	7,864,655,400
Number of reads	78,646,554
Sequences after filtering	
Total nucleotides (nt)	7,657,276,643
Number of reads	77,023,196
Assembly statistics	
Length of all unigenes (nt)	224,284,436
Total number of unigenes	208,959
Average sequence size of unigenes (nt)	1,073

Identification of cDNA-derived SSRs
Since microsatellite markers established by improved detection of combined simple sequence repeats (SSRs) have been widely used in plant genetic analysis, large amount of transcriptomic data generated from NGS approaches present an opportunity for high-throughput identification of SSRs (Zalapa et al., 2012). To develop new molecular markers, the 208,959 unigenes generated in this study were mined for potential microsatellites, which were defined as mono- to hexa-nucleotide SSR with a minimum of five repetitions for all motifs. A total of 47,273 potential cSSRs were identified in 38,446 unigenes by MISA (http://pgrc.ipk-gatersleben.de/misa/). 6,956 sequences contained more than 1 cSSR, and 2,665 cSSRs were present in compound formation (Tab. 2). Among the 47,273 cSSRs, the mono-nucleotide repeat motif (40.79 %) was most abundant, followed by di-nucleotide (37.87 %), tri-nucleotide (18.59 %), tetra-nucleotide (2.49 %), penta-nucleotide (0.17 %) and hexa-nucleotide (0.08 %) repeat motifs. Di- to hexa-nucleotide repeat motifs were further analyzed for the number of repeat units. As shown in Tab. S3, the most common repeat unit was 6, accounting for 29.58 % of the total potential cSSRs, followed by 5 (22.64 %), 7 (17.31 %), and 8 (11.91 %) repeat units. Furthermore, cSSR length was mostly distributed from

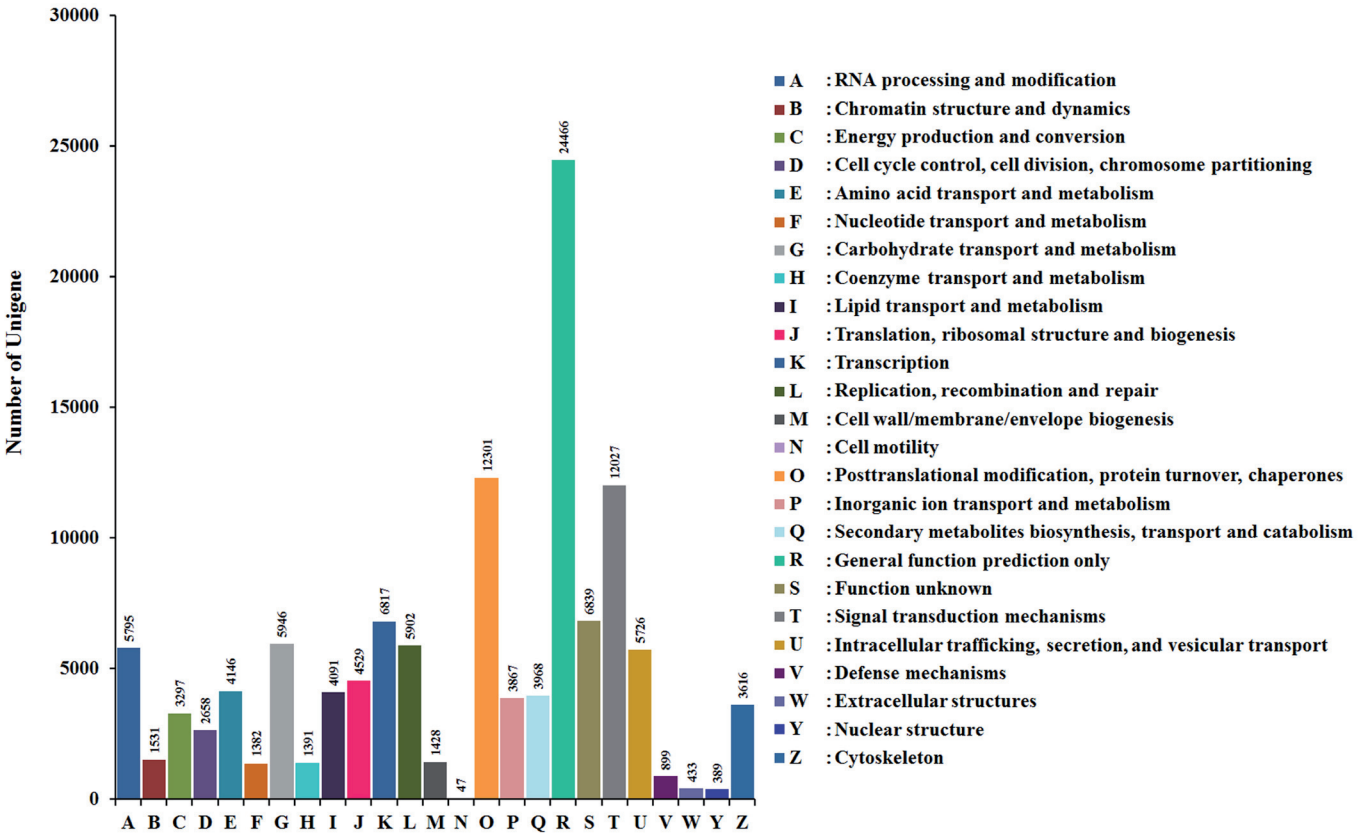

Fig. 1: Histogram presentation of the clusters of orthologous groups (COG) classification. All unigenes were aligned to the COG database to predict and classify possible functions.

Tab. 2: Statistics of SSRs identified in *Oplopanax elatus* transcriptome assembly.

Searching Item	Numbers
SSR mining	
Total number of sequences examined	208,959
Total size of examined sequences (bp)	224,284,436
Total number of identified SSRs	47,273
Number of SSR containing sequences	38,446
Number of sequences containing more than 1 SSR	6,956
Number of SSRs present in compound formation	2,665
Distribution of SSRs in different repeat types	
Mono-nucleotide (≥10 repeats)	19,285 (40.79%)
Di-nucleotide (≥6 repeats)	17,904 (37.87%)
Tri-nucleotide (≥5 repeats)	8,788 (18.59%)
Tetra-nucleotide (≥5 repeats)	1,177 (2.49%)
Penta-nucleotide (≥5 repeats)	80 (0.17%)
Hexa-nucleotide (≥5 repeats)	39 (0.08%)

12 to 20 bp, and this range accounted for 98.03% of the total potential cSSRs. Finally, di-nucleotide and tri-nucleotide repeats were present in 4 and 10 types, respectively (Fig. 2). The AG/CT di-nucleotide repeat motif was most abundant motif (31.13%), followed by AT/AT (22.86%), AC/GT (9.69%), AAG/CTT (7.11%) and AAT/ATT (5.36%). Taken together, these findings demonstrate that the transcriptome library of *O. elatus* facilitates the discovery of molecular markers that will provide a valuable genetic resource for this species as well as for the Araliaceae family.

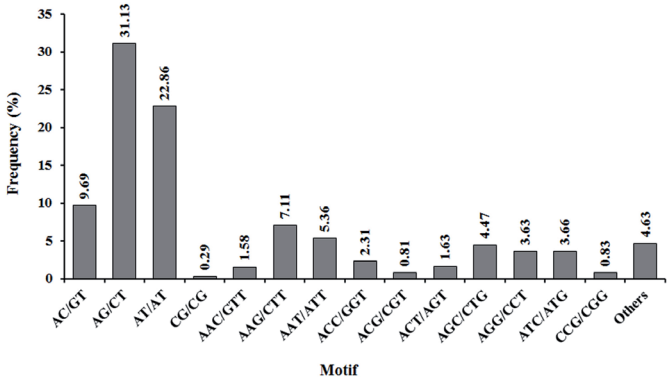

Fig. 2: Frequency distribution of cSSRs according to sequence motifs. A total 81 sequence motifs were identified. Di-, tri- and other (including tetra-, penta- and hexa-) nucleotide repeats were present in 4, 10, and 67 types, respectively.

Candidate enzymes involved in the biosynthesis of the backbone of triterpenoid saponin

The saponins constitute a large class of secondary metabolites in *O. elatus* and are the major source of its medicinal properties (Shikov et al., 2014). As shown in Fig. 3A, triterpenoid saponins in plants are derived from the condensation of isopentenyl diphosphate (IPP) and dimethylallyl pyrophosphate (DMAPP), which is produced by two independent pathways, the methyl-eryhritol phosphate (MEP) pathway and mevalonate (MVA) pathway (Sawai and Saito, 2011). Farnesyl pyrophosphate (FPP) is the common precursor of the vast array of sesquiterpenes, and squalene synthase (SQS, EC 2.5.1.21)

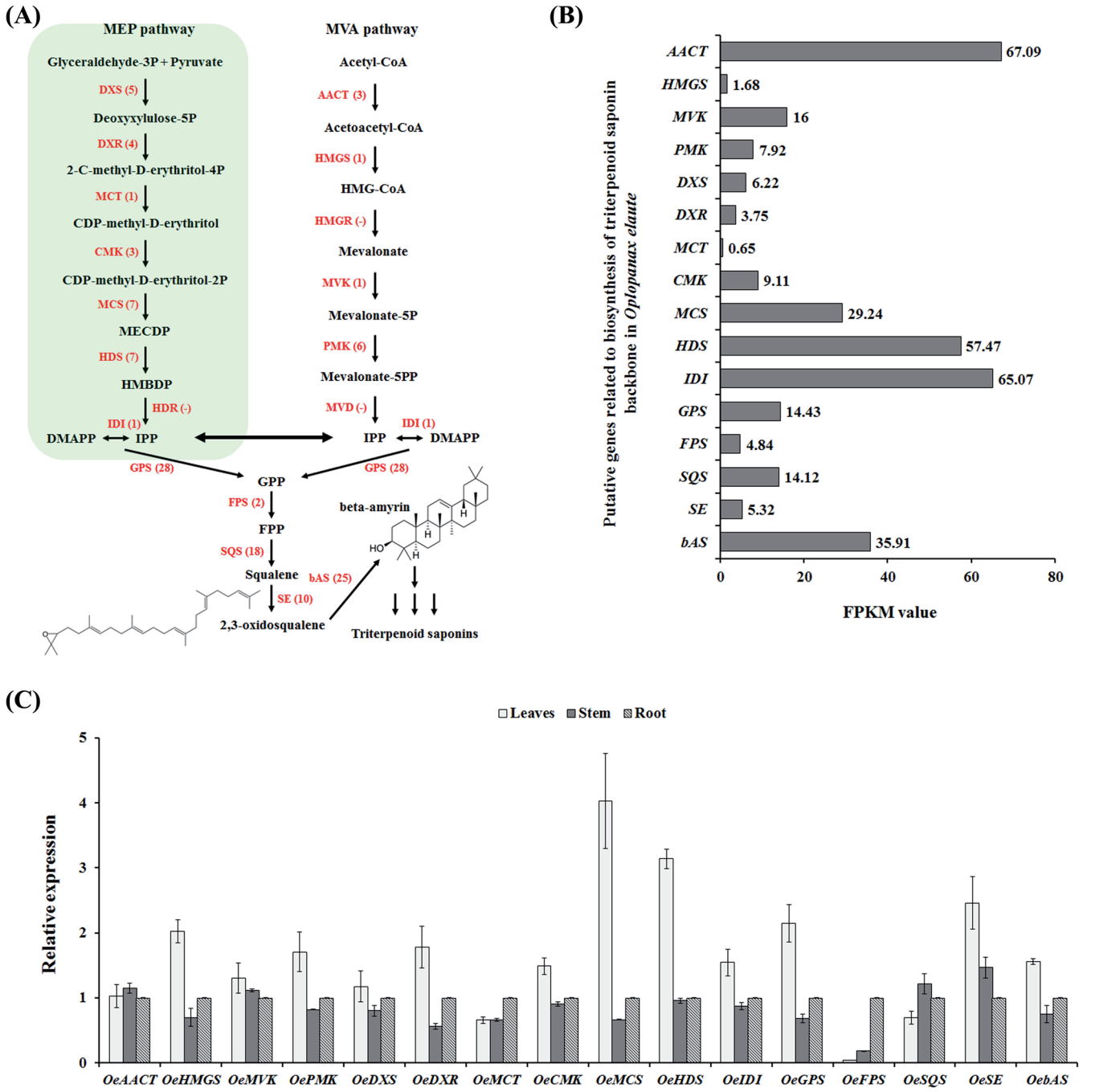

Fig. 3: Relative expression of putative genes involved in the biosynthesis of the triterpenoid saponin backbone. **(A)** *Oplopanax elatus* unigenes involved in the biosynthesis of the backbone of triterpenoid saponins. Enzyme names are abbreviated as follows: Acetyl-CoA C-acetyltransferase (AACT), Hydroxymethylglutaryl-CoA synthase (HMGS), Hydroxymethylglutaryl-CoA reductase (HMGR), Mevalonate kinase (MVK), Phosphomevalonate kinase (PMK), Diphosphomevalonate decarboxylase (MVD), 1-deoxy-D-xylulose-5-phosphate synthase (DXS), 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR), 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase (MCT), 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (CMK), 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MCS), (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase (HDS), 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (HDR), Isopentenyl-diphosphate delta-isomerase (IDI), Geranyl pyrophosphate synthase (GPS), Farnesyl diphosphate synthase (FPS), Squalene synthase (SQS), squalene monooxygenase (SE), and beta-amyrin synthase (bAS). The red number in bracket following each gene name indicates the number of corresponding unigenes. **(B)** FPKM values of 16 selected genes involved in the biosynthesis of the triterpenoid saponin backbone. **(C)** Expression of selected genes in different organs. The expression of each gene in different organs was normalized to its expression in root. Data represent the means ± SE of three biological replicates.

converts two FPP (C15) units into a squalene (C30). Subsequently, squalene monooxygenase (SE, EC 1.14.12.132) catalyzes the conversion of squalene to 2,3-oxidosqualene. In addition, beta-amyrin synthase (bAS, EC 5.4.99.39) catalyzes the cyclization of 2,3-oxidosqualene into beta-amyrin, a major configuration of pentacyclic triterpenoids (Sun et al., 2006). Consistent with our prediction, which was based on the KEGG pathway assignment, we found 122 unigenes encoding 47 putative enzymes involved in the biosynthesis of the backbone of triterpenoid saponins in the *O. elatus* transcriptome library (Tab. 3 and Fig. 3A). For almost all of these, more than

Tab. 3: *Oplopanax elatus* unigenes potentially involved in triterpenoid saponin biosynthesis.

	Gene	EC number	Gene Name[1]	NU	MNPU	NGSN
Mevalonate pathway	Acetyl-CoA C-acetyltransferase	2.3.1.9	AACT	3	2	-
	Hydroxymethylglutaryl-CoA synthase	2.3.3.10	HMGS	1	1	-
	Hydroxymethylglutaryl-CoA reductase	1.1.1.34	HMGR	-	-	-
	Mevalonate kinase	2.7.1.36	MVK	1	1	-
	Phosphomevalonate kinase	2.7.4.2	PMK	6	1	-
	Diphosphomevalonate decarboxylase	4.1.1.33	MVD	-	-	-
MEP pathway	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	DXS	5	3	-
	1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	DXR	4	1	-
	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	2.7.7.60	MCT	1	1	-
	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	2.7.1.148	CMK	3	2	-
	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	4.6.1.12	MCS	7	5	-
	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase	1.17.7.1	HDS	7	2	-
	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	1.17.1.2	HDR	-	-	-
	Isopentenyl-diphosphate delta-isomerase	5.3.3.2	IDI	1	1	-
Sesquiterpenoid and triterpenoid biosynthesis	Geranyl pyrophosphate synthase	2.5.1.29	GPS	28	7	-
	Farnesyl diphosphate synthase	2.5.1.10	FPS	2	2	-
	Squalene synthase	2.5.1.21	SQS	18	4	-
	squalene monooxygenase	1.14.13.132	SE	10	8	-
	beta-amyrin synthase	5.4.99.39	bAS	25	6	-
Total				122	47	-

[1] Following the *Arabidopsis thaliana* genes nomenclature.
NU, number of unigenes; **MNPU**, minimum number of proteins encoded by unigenes**; NGSN**, number of genes stored in NCBI

one unigene sequence was annotated as same enzyme, indicating that such unigene sequences might either represent different fragments of a single transcript or different members of a gene or both (Hyun et al., 2012). Based on the FPKM (Fragments Per Kilobase of exon per Million fragments mapped) values, we selected unigenes as the major genes in each enzymatic step. The read coverage of theses unigenes was average of 21.18 FPKM (Fig. 3B), which showed high coverage of *O. elatus* transcripts (average of 3.97 FPKM). To investigate the organ distribution of transcripts for 16 major genes, organ-specific expression was analyzed by quantitative real-time PCR with unigene-specific primer pairs. As shown in Fig. 3C, most of the selected genes were highly expressed in the leaves, relative to root, whereas the transcription level of *OeMCT* and *OeFPS* were higher in root than in the leaves. *OeSQS* was expressed at almost equal levels in the leaves, stem, and root. This suggests that the leaves may be main site of triterpenoid saponin biosynthesis via the isoprenoid pathway.

Discussion

O. elatus, commonly called Asian devil's club, contains a variety of bioactive compounds, including triterpenoid saponins and polyacetylenes, and has been used for treating hypopiesis, cardiovascular, diabetes mellitus, neurasthenic and rheumatism (Dou et al., 2009; Shikov et al., 2014). Although this indicates that *O. elatus* is a valuable medicinal plant that, like ginseng, recommends itself as source of herbal preparations, little molecular information on this plant is available. In NCBI, only 22 nucleotide sequences are available from *O. elatus*. Since NGS technology has become available, transcriptome analysis based on RNA-seq and *de-novo* assembly has proven to be an invaluable tool for the identification of putative genes involved in various metabolic processes in non-model plants (Xiao et al., 2013; Unamba et al., 2015). In view of the pharmacological importance of the triterpenoid saponins that are found in the roots and rhizomes of *O. elatus*, a transcriptome library was generated on an Illumina HiSeq^{TM}2000 sequencing platform. Based on *de novo* assembly, we obtained approximately 77 million high quality reads and identified a total of 208,959 unigenes (Tab. 1). In addition, 110,202 unigenes were annotated by BLAST analysis against public databases with a cut-off E-value of 10^{-5} (Tab. S2). Although this is the first attempt of *de novo* sequencing and assembly of *O. elatus* without a reference genome, the gene catalogs, functional annotation and classification suggest that the *O. elatus* transcriptome library provides important new insights and will facilitate further studies of functional genes in the family Araliaceae.

Triterpenoid saponins are common in the Araliaceae (Shikov et al., 2014; Yang et al., 2014). Although the biological role of saponins in plants is not fully understood, they are considered to be part of the defense systems in virtue of their antimicrobial, antifungal, and anti-insect activities (Sparg et al., 2004). In addition, they exhibit several pharmacological activities, including anti-inflammatory, antidiabetic and antitumor properties (Augustin et al., 2011; Yang et al., 2014), and thus make attractive candidates for novel drug development (Augustin et al., 2011). Triterpenoid saponins derive from condensation of five-carbon building block designated IPP and DMAPP, which are synthesized by the MEP and MVA pathways (Sawai and Saito, 2011). In our *O. elatus* transcriptome library, we identified 122 unigenes encoding 47 putative enzymes in pathways for terpenoid backbone (MEP and MVA pathways) and triterpenoid saponin biosynthesis (Tab. 3). These unigenes covered almost all genes involved in triterpenoid saponin biosynthesis, indicating that the *O. elatus* transcriptome library has facilitated the discovery of novel genes involved in secondary metabolite synthesis. It has been shown that genes in the MEP pathway are highly expressed in leaf, relative to other organs (Devi et al., 2016; Dorn et al., 2013), and a similar organ expression pattern of *O. elatus* MEP pathway genes was found

in this study (Fig. 3C). In addition, most of the selected genes involved in triterpenoid saponin biosynthesis were highly expressed in *O. elatus* leaves (Fig. 3C). In *Achyranthus bidentate*, saponins are synthesized in the leaves and then transported to the roots via the phloem (Li and Hu, 2009). *SE* is highly expressed in leaves of *Chlorophytum borivilianum*, relative to the roots, whereas the expression level of selected genes encoding cytochrome P450 and glucosyltransferases was higher in the roots than in the leaves (Kumar et al., 2012). Taken together, this evidence suggests that triterpenoid saponins are mainly synthesized in leaves and modified in roots.

In conclusion, using the Illumina HiSeq^{TM}2000 sequencing platform, we have generated large-scale transcriptome library in *O. elatus*, and shown that the *O. elatus* transcriptome harbors rich information about the genes that are involved in various metabolic pathways, including in the biosynthesis of the backbone of triterpenoid saponins. In addition, a large number of presumably highly polymorphic cDNA-derived SSRs were identified. These will provide a wealth of markers for further genetic study. Furthermore, the organ-specific expression of selected genes indicates that triterpenoid saponins are mainly synthesized in the leaves of *O. elatus*. Further investigation on the transporter of triterpenoid saponins will be an exciting and important endeavor that will continue to disclose the mechanisms of triterpenoid saponin from leaves to roots. Taken together, the transcriptome dataset generated in this study will serve as a valuable resource for accelerating genomic and functional genomic research in *O. elatus* and in the family Araliaceae.

Acknowledgments

This work was supported by the research grant from the National Research Foundation of Korea (NRF-2015R1C1A1A01053448) funded by the Ministry of Science.

References

Augustin, J.M., Kuzina, V., Andersen, S.B., Bak, S., 2011: Molecular activities, biosynthesis and evolution of triterpenoid saponins. Phytochemistry 72, 435-457.

Calway, T., Du, G.J., Wang, C.Z., Huang, W.H., Zhao, J., Li, S.P., Yuan, C.S., 2012: Chemical and pharmacological studies of *Oplopanax horridus*, a North American botanical. J. Nat. Med. 66, 249-256.

Devi, K., Mishra, S.K., Sahu, J., Panda, D., Modi, M.K., Sen, P., 2016: Genome wide transcriptome profiling reveals differential gene expression in secondary metabolite pathway of *Cymbopogon winterianus*. Sci. Rep. 15, 21026.

Dorn, K.M., Fankhauser, J.D., Wyse, D.L., Marks, M.D., 2013: *De novo* assembly of the pennycress (*Thlaspi arvense*) transcriptome provides tools for the development of a winter cover crop and biodiesel feedstock. Plant J. 75, 1028-1038.

Dou, D.Q., Hu, X.Y., Zhao, Y.R., Kang, T.G., Liu, F.Y., Kuang, H.X., Smith, D.C., 2009: Studies on the anti-psoriasis constituents of *Oplopanax elatus* Nakai. Nat. Prod. Res. 23, 334-342.

Elekofehinti, O.O., 2015: Saponins: Anti-diabetic principles from medicinal plants – A review. Pathophysiology 22, 95-103.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., 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.

Huang, W.H., Zhang, Q.W., Yuan, C.S., Wang, C.Z., Li, S.P., Zhou, H.H., 2014: Chemical constituents of the plants from the genus *Oplopanax*. *Chem*. Biodivers. 11, 181-196.

Hwang, H.S., Lee, H., Choi, Y.E., 2014: Transcriptomic analysis of Siberian ginseng (*Eleutherococcus senticosus*) to discover genes involved in saponin biosynthesis. BMC Genomics 16, 180.

Hyun, T.K., Rim, Y., Jang, H.J., Kim, C.H., Park, J., Kumar, R., Lee, S., Kim, B.C., Bhak, J., Nguyen-Quoc, B., Kim, S.W., Lee, S.Y., Kim, J.Y., 2012: *De novo* transcriptome sequencing of *Momordica cochinchinensis* to identify genes involved in the carotenoid biosynthesis. Plant Mol. Biol. 79, 413-427.

Hyun, T.K., Lee, S., Rim, Y., Kumar, R., Han, X., Lee, S.Y., Lee, C.H., Kim, J.Y., 2014: *De-novo* RNA sequencing and metabolite profiling to identify genes involved in anthocyanin biosynthesis in Korean black raspberry (*Rubus coreanus* Miquel). PLoS One 9, e88292.

Kalra, S., Puniya, B.L., Kulshreshtha, D., Kumar, S., Kaur, J., Ramachandran, S., Singh, K., 2013: *De novo* transcriptome sequencing reveals important molecular networks and metabolic pathways of the plant, *Chlorophytum borivilianum*. PLoS One 8, e83336.

Kumar, S., Kalra, S., Kumar, S., Kaur, J., Singh, K., 2012: Differentially expressed transcripts from leaf and root tissue of *Chlorophytum borivilianum*: a plant with high medicinal value. Gene 511, 79-87.

Lee, S.-W., Kim, Y.-M., Kim, W.-W., Chung, J.-M., 2002: Genetic variation of I-SSR markers in the natural populations of a rare and endangered tree species, *Oplopanax elatus* in Korea. J. Korean For. Soc. 91, 565-573.

Li, J., Hu, Z., 2009: Accumulation and dynamic trends of triterpenoid saponin in vegetative organs of *Achyranthus bidentata*. J. Integr. Plant Biol. 51, 122-129.

Man, S., Gao, W., Zhang, Y., Huang, L., Liu, C., 2010: Chemical study and medical application of saponins as anti-cancer agents. Fitoterapia 81, 703-714.

Moon, H.K., Kim, Y.W., Hong, Y.P., Park, S.Y., 2013: Improvement of somatic embryogenesis and plantlet conversion in *Oplopanax elatus*, an endangered medicinal woody plant. Springerplus 2, 428.

Moses, T., Papadopoulou, K.K., Osbourn, A., 2014: Metabolic and functional diversity of saponins, biosynthetic intermediates and semi-synthetic derivatives. Crit. Rev. Biochem. Mol. Biol. 49, 439-462.

Mutz, K.O., Heilkenbrinker, A., Lonne, M., Walter, J.G., Stahl, F., 2013: Transcriptome analysis using next-generation sequencing. Curr. Opin. Biotechnol. 24, 22-30.

Rai, A., Yamazaki, M., Takahashi, H., Nakamura, M., Kojoma, M., Suzuki, H., Saito, K., 2016: RNA-seq transcriptome analysis of *Panax japonicus*, and its comparison with other *Panax* species to identify potential genes involved in the saponins biosynthesis. Front. Plant Sci. 12, 481.

Sawai, S., Saito, K., 2011: Triterpenoid biosynthesis and engineering in plants. Front. Plant Sci. 30, 25.

Schliesky, S., Gowik, U., Weber, A.P., Brautigam, A., 2012: RNA-seq assembly-Are we there yet? Front. Plant Sci. 3, 220.

Seki, H., Tamura, K., Muranaka, T., 2015: P450s and UGTs: Key players in the structural diversity of triterpenoid saponins. Plant Cell Physiol. 56, 1463-1471.

Semalty, A., Semalty, M., Panda, V.S., Asrani, K.H., Ashar, H.D., 2012: Herbal drugs in chronic fatigue syndrome: An overview. Schweiz Z. Ganzheitsmed. 24, 155-168.

Shikov, A.N., Pozharitskaya, O.N., Makarov, V.G., Yang, W.Z., Guo, D.A., 2014: *Oplopanax elatus* (Nakai) Nakai: chemistry, traditional use and pharmacology. Chin. J. Nat. Med. 12, 721-729.

Sparg, S.G., Light, M.E., van Staden, J., 2004: Biological activities and distribution of plant saponins. J. Ethnopharmacol. 94, 219-243.

Sun, H., Fang, W.-S., Wang, W.-Z., Hu, C., 2006: Structure-activity relationships of oleanane- and ursane-type triterpenoids. Bot. Stud. 47, 339-368.

Sun, S.X., Li, Y.M., Fang, W.R., Cheng, P., Liu, L., Li, F., 2010: Effect and mechanism of AR-6 in experimental rheumatoid arthritis. Clin. Exp. Med. 10, 113-121.

Unamba, C.I.N., Nag, A., Sharma, R.K., 2015: Next generation sequencing Technologies: the doorway to the unexplored genomics of non-model plants. Front. Plant Sci. 16, e01074.

Wei, L., Li, S., Liu, S., He, A., Wang, D., Wang, J., Tang, Y., Wu, X.,

2014: Transcriptome analysis of *Houttuynia cordata* Thunb. by Illumina paired-end RNA sequencing and SSR marker discovery. PLoS One 9, e84105.

XIAO, M., ZHANG, Y., CHEN, X., LEE, E.J., BARBER, C.J., CHAKRABARTY, R., DESGAGNÉ-PENIX, I., HASLAM, T.M., KIM, Y.B., LIU, E., MACNEVIN, G., MASADA-ATSUMI, S., REED, D.W., STOUT, J.M., ZERBE, P., ZHANG, Y., BOHLMANN, J., COVELLO, P.S., DE LUCA, V., PAGE, J.E., RO, D.K., MARTIN, V.J., FACCHINI, P.J., SENSEN, C.W., 2013: Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. J. Biotechnol. 166, 122-134.

YANG, W.Z., HU, Y., WU, W.Y., YE, M., GUO, D.A., 2014: Saponins in the genus *Panax* L. (Araliaceae): a systematic review of their chemical diversity. Phytochemistry 106, 7-24.

ZALAPA, J.E., CUEVAS, H., ZHU, H., STEFFAN, S., SENALIK, D., ZELDIN, E., MCCOWN, B., HARBUT, R., SIMON, P., 2012: Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am. J. Bot. 99, 193-208.

ZHANG, S., WU, Y., JIN, J., HU, B., ZENG, W., ZHU, W., ZHENG, Y., CHEN, P., 2015: *De novo* characterization of *Panax japonicus* C. A. Mey transcriptome and genes related to triterpenoid saponin biosynthesis. Biochem. Biophys. Res. Commun. 466, 450-455.

ZHENG, X., XU, H., MA, X., ZHAN, R., CHEN, W., 2014: Triterpenoid saponin biosynthetic pathway profiling and candidate gene mining of the *Ilex asprella* root using RNA-Seq. Int. J. Mol. Sci. 15, 5970-5987.

Address of the corresponding author:
Dr. Tae Kyung Hyun, Department of Industrial Plant Science and Technology, College of Agricultural, Life and Environmental Scien-ces, Chungbuk National University, Cheongju 28644, Republic of Korea
E-mail: taekyung7708@chungbuk.ac.kr

Supplementary material

Tab. S1: Primer sequences for qReal-time PCR analysis.

Primer	Sequences (5'-3')
OeAACT-F	TTCAAGTGCTGGCATAGAGG
OeAACT-Rev	CACCGTGTCTGTTGACCTTC
OeHMGS-F	GTTACTTGAACGCCCTGGAT
OeHMGS-Rev	CTTCTGCACCAGTTTGCAGT
OeMVK-F	GAGCTTCCTCTTGGATCAGG
OeMVK-Rev	AACCTTCGTGGTTGAAGTCC
OePMK-F	AGTTGGAGGCAAGCCATTAG
OePMK-Rev	CCATTGATGGTGTGGATGAT
OeDXS-F	CCCATCTTCTGTGTTCACTTG
OeDXS-Rev	TGTATCAAGCATGGGTGTCA
OeDXR-F	TGGTATCATCACCGTCCCTA
OeDXR-Rev	TTGCAAACTGGCAGCATATT
OeMCT-F	GCCGAAGCATTGATTTGTACT
OeMCT-Rev	CAAGCAATTGTCCAAGCAAA
OeCMK-F	CATTCGAAGTCCTCCCATCT
OeCMK-Rev	TTCCAACTATGGTGCTTCCA
OeMCS-F	TCGGATCCTAAATGGAAAGG
OeMCS-Rev	AAAGTGGCATCCAAGTTTCC
OeHDS-F	GCGGAATACAAAGACGGAAT
OeHDS-Rev	GATTGCGATTGAAACACCAG
OeIDI-F	TCATTGTTCGGGATGTGAAT
OeIDI-Rev	AAACCCTCTTCACCAGCATC
OeGPS-F	TACAAGACTGCGTCGCTGAT
OeGPS-Rev	CATTGACACTTCAGCCGTTT
OeFPS-F	GCAATCGGATTTACCCAAGT
OeFPS-Rev	CCTGTTTAGCTTCCCTCCAG
OeSQS-F	GTGACCCTAACGCTACAGCA
OeSQS-Rev	CTTGGCTTGCCCTTGATTAT
OeSE-F	ACCCATCACCCATCTTGTTT
OeSE-rev	GGAGCAACCACAGTCTTCAA
OebAS-F	TGCGTTTACTTGGAGAAGGA
OebAS-Rev	CCCATGAAGGTATAGCGGTT
OeEF-F	AACGTTGTAGCCGACCTTCT
OeEF-Rev	GTCTGCACGAACAAGATGGA

Tab. S2: Top BLAST hits in the NCBI NR database. BLAST results with respect to the NCBI NR database for all distinct sequences with a cutoff E-value greater than 10^{-5} are shown.

Please note the attached file: 6876-27433-1-SP.xlsx

Tab. 3: Frequencies of the different SSR repeat motifs encountered in transcriptome.

SSR Motif	Number of repeat unit									Total	%
	5	6	7	8	9	10	11	12	13-24		
Di-nucleotide	0	5,793	3,682	3,246	2,698	1,947	534	4	0	17,904	63.97
Tri-nucleotide	5,235	2,307	1,155	83	4	2	0	0	2	8,788	31.40
Tetra-nucleotide	997	170	8	2	0	0	0	0	0	1,177	4.21
Penta-nucleotide	72	6	1	1	0	0	0	0	0	80	0.29
Hexa-nucleotide	32	3	0	3	0	1	0	0	0	39	0.14
Total	6,336	8,279	4,846	3,335	2,702	1,950	534	4	2	27,988	100
%	22.64	29.58	17.31	11.91	9.65	6.97	1.91	0.01	0.01	100	

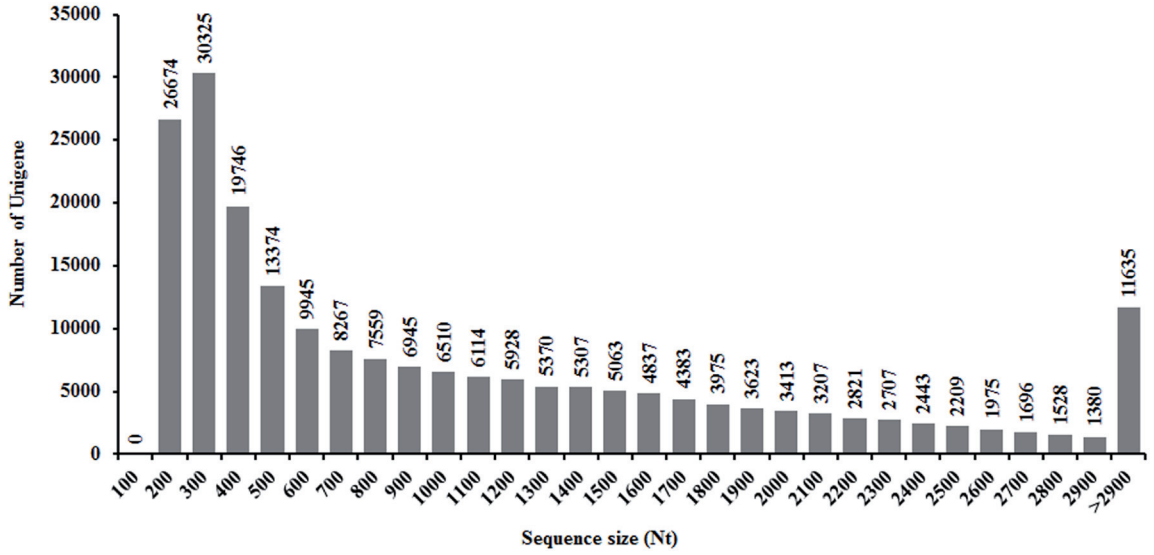

Fig. S1: Length distribution of the assembled unigenes.