G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(4):e34
https://doi.org/10.5808/GI.2018.16.4.e34

ORIGINAL ARTICLE

RNA-Seq *De Novo* Assembly and Differential Transcriptome Analysis of Korean Medicinal Herb *Cirsium japonicum* var. *spinossimum*

Neha Samir Roy[1,2], Jung-A Kim[3], Ah-Young Choi[4], Yong-Wook Ban[5], Nam-Il Park[6], Kyong-Cheul Park[1], Hee-sun Yang[3], Ik-Young Choi[1,2]*, Soonok Kim[3]**

[1]Department of Agriculture and Life Industry, Kangwon National University, Chuncheon 24341, Korea, [2]Agriculture and Life Sciences Research Institute, Kangwon National University, Chuncheon 24341, Korea, [3]Biological Resources Assessment Division, National Institute of Biological Resources, Incheon 22689, Korea, [4]DNACare Genomics Institute, Seongnam 13558, Korea, [5]Department of Forest Environmental System, Kangwon National University, Chuncheon 24341, Korea, [6]Department of Plant Science, Gangneung Wonju National University, Gangneung 25457, Korea

Cirsium japonicum belongs to the Asteraceae or Compositae family and is a medicinal plant in Asia that has a variety of effects, including tumour inhibition, improved immunity with flavones, and antidiabetic and hepatoprotective effects. Silymarin is synthesized by 4-coumaroyl-CoA via both the flavonoid and phenylpropanoid pathways to produce the immediate precursors taxifolin and coniferyl alcohol. Then, the oxidative radicalization of taxifolin and coniferyl alcohol produces silymarin. We identified the expression of genes related to the synthesis of silymarin in *C. japonicum* in three different tissues, namely, flowers, leaves, and roots, through RNA sequencing. We obtained 51,133 unigenes from transcriptome sequencing by *de novo* assembly using Trinity v2.1.1, TransDecoder v2.0.1, and CD-HIT v4.6 software. The differentially expressed gene analysis revealed that the expression of genes related to the flavonoid pathway was higher in the flowers, whereas the phenylpropanoid pathway was more highly expressed in the roots. In this study, we established a global transcriptome dataset for *C. japonicum*. The data shall not only be useful to focus more deeply on the genes related to product medicinal metabolite including flavolignan but also to study the functional genomics for genetic engineering of *C. japonicum*.

Keywords: *Cirsium japonicum*, differentially expressed gene, RNA-sequencing, silymarin, transcriptome

Introduction

Cirsium japonicum is a wild perennial herb found in many areas of Korea, Japan, and China. *C. japonicum* is used as an anti-haemorrhagic, anti-hypertensive and uretic agent in traditional Chinese medicine [1]. In traditional medicine, *C. japonicum* is sometimes used for the management of different types of cancer, including liver and uterine cancer and leukaemia [2]. To date, many studies have explored the effects of *C. japonicum* on various diseases. However, no reports have performed molecular biology studies or investigated comprehensive genomic and transcriptomic data in *C. japonicum*.

In Europe, silymarin, which is synthesized in *Silybum marianum*, belongs to the same family as *C. japonicum* and is prescribed for the treatment of chronic liver disease and the prevention of recurrent hepatitis C in liver transplant recipients [3, 4]. Silymarin is the key component of *S. marianum* and includes flavonolignans (silibinin, isosilibinin, silychristin, isosilychristin, and silydianin) and a flavonoid (taxifolin). Studies suggest that silymarin exerts a protective effect on the liver in alcoholic-induced liver disease, non-alcoholic fatty liver disease and carbon tetrachloride-induced oxidative liver [1, 5]. This major bioactivity has also been reported in *C. japonicum* DC, which has properties

Received December 10, 2018; Revised December 17, 2018; Accepted December 17, 2018; Published online December 28, 2018

***Corresponding author:** Tel: +82-33-250-7768, Fax: +82-33-251-9556, E-mail: choii@kangwon.ac.kr

****Corresponding author:** Tel: +82-32-590-7110, Fax: +82-32-590-7069, E-mail: sokim90@korea.kr

similar to those of *S. marianum* [6]. *C. japonicum* var. *spinossimum* is often compared to the basic species thistle (*C. japonicum* Fisch ex DC) [7]. Silymarin is synthesized through the phenylpropanoid pathway and transforms phenylalanine into 4-coumaroyl-CoA, which acts via both the flavonoid and monolignol pathways [8-10].

Significant progress has been achieved in the field of next generation sequencing, which is driven by genomic/transcriptomic-based inquiries in biology. High-quality sequences are generated in a high-throughput manner at a low cost and with little labour. The Illumina platform is a highly utilized platform employed for transcriptome analyses of various model and non-model organisms and medicinal plants due to its potential for a high sequence yield [11-13]. The RNA sequencing (RNA-Seq) method is superior to whole genome sequencing because RNA-Seq only studies the transcribed regions and provides a comprehensive and integrated view of the transcriptome with the precise locations of the transcriptome boundaries [14]. RNA-Seq is currently a very popular method used to examine both coding and non-coding gene annotations. Remarkable progress has been achieved in the exploration of medicinal plants at the genomic and transcriptomic levels, and the ultimate goal is to identify genes that are involved in biologically active phytocompounds and related pathways [15-18].

In this study, we established transcript databases for *C. japonicum* var. spinossimum and provided additional genetic information for further genome-wide research and analyses. We also aimed to investigate the transcripts that contribute to the production of silymarin in this herb.

Methods

Plant material, RNA extraction, and library preparation

C. japonicum was obtained from the National Institute of Biological Resources, Korea. The flower, leaf, and root tissues were immediately dissected and grinded in liquid nitrogen for the RNA extraction. The total RNA was extracted using Hybrid-R (Geneall, PN: 3033522) according to the manufacturer's instructions. The total RNA was further treated with RNase-free DNase I (TaKaRa, Tokyo, Japan) and purified on an RNA-purification column (Qiagen, Valencia, CA, USA) to eliminate possible gnomic contamination. The RNA quality was evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and RNA samples with RNA integrity number values above eight were used for the subsequent cDNA synthesis. The DNA was sheared with an average of 500 bp fragment sizes. The TruSeq Library Preparation Kit (Illumina Inc., San Diego, CA, USA) was used to construct the DNA library according to the manufacturer's protocol. The DNA libraries were sequenced with 150-bp paired-end sequencing using an Illumina Hiseq2500. The quality of the constructed libraries was confirmed by a LabChip GX system (PerkinElmer, Waltham, MA, USA).

De novo assembly, functional annotation, Gene Ontology, and Kyoto Encyclopedia of Genes and Genomes pathway analysis

The raw data were quality-filtered using trimmomatic with the following options: minimum quality of base (3); sliding window (4); average quality phred score (20); and minimum read size (50 bp). The primer and adapter sequences incorporated during cDNA synthesis were removed. The *de novo* assembly of the reads was performed using Trinity assembler with the default options to form the contigs. Based on the final transcriptome isoform sequences, the candidate-coding regions were identified using TransDecoder software (The Broad Institute, Cambridge, MA, USA). A BLAST analysis was carried out using these candidate-coding regions against the UniProt and NCBI non-redundant (nr) protein databases to determine the sequence similarity with genes from other species at an E-value cut-off of 1×10^{-6}. The functional categories of these sequences were matched to the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) algorithms using blast. The GO analysis, including the biological function, cell component and molecular function categories, and the KEGG pathway analysis were performed using the GOstats program (Roswell Park Center Institute, Buffalo, NY, USA) as implemented in the sequence annotation tool Blast2GO (BioBam Bioinformatics SL, Valencia, Spain). The gene annotation and GO analysis were performed at NICEM, Seoul National University (Seoul, Korea). The assembled data were arranged according to the read length, GenBank number, E-value and species, and the specific composition of the GO terms was calculated and presented in a bar chart according to the percentage.

Differential gene expression analysis

The filtered raw reads were mapped to the *C. japonicum* contigs that were determined by the *de novo* assembly and annotation using HISAT2. The expression values in RNA-Seq were calculated based on the read count, and the expression levels were analysed using the counts (number of fragments) and FPKM (Fragments per Kilobase of transcript per Million fragments mapped reads). The differentially expressed gene (DEG) analysis among the leaf, flower and root tissues was carried out using the edgeR package [18]. The significantly DEGs were screened at a threshold false

discovery rate < 0.05 and an absolute log2-fold change value > 1. Subsequently, GO functional enrichment and KEGG pathway analyses of the DEGs were performed.

Results

De novo assembly

To study the transcriptome of *C. japonicum*, three tissues, viz., flowers, leaves, and roots, were extracted, and the RNA was isolated. The total RNA from each sample was used for the mRNA preparation, fragmentation, cDNA synthesis and library preparation. *De novo* assembly was selected since a reference genome for *C. japonicum* is lacking. Each library was sequenced using the Illumina Hiseq2500 platform. The platform generated 74,566,546 raw reads for *C. japonicum* for the whole transcriptome, which accounted for 18Gbases of sequence data (Supplementary Table 1). These data were further subjected to a quality check and adaptor trimming (using the trimmomatic 0.30 tool), generating 67,153,874 high-quality paired end reads (Tables 1 and 2). Then, an error correction of the data was performed, and 66,078,302 reads were obtained. The paired end sequencing yielded on average a 64× sequencing depth. The mean Phred value was above 20, and the short reads (>50) were removed. The whole transcriptome was assembled using the Trinity program, resulting in 51,133 unigenes, accounting for 37 Mbases, with a mean size of 648.36 bp. The length distribution of all unigenes is shown in Fig. 1. The GC content was distributed within 42%–45%. The minimum size of the transcripts was 147 bp, whereas the maximum size was 15,402 bp (Table 1).

Table 1. The total transcripts and length by *de novo* assembly of RNA sequence data in *Cirsium japonicum*

Assembly data	No.
Transcripts	51,133
Total (bp)	37,533,942
Mean size (bp)	648
Maximum size	15,402

Functional annotation and classification

The transcripts derived from *C. japonicum* were subjected to annotation using the BLASTX program based on a homology search against the NCBI non-redundant (nr) protein database. The highest homologous subject in each contig was selected for gene annotation. The identified contigs were contigs that hit to variable species with a high similarity homologous BLAST search. In total, 33,525 (65%) unigenes were annotated using the BLASTX program. The similarity distribution demonstrated that *C. japonicum* was highly similar to *Cynara cardunculus* var. *scolymus* (72%) (Fig. 2). GO broadly categorizes genes into one of three classes (biological process, cellular component, and molecular function) to describe their involvement in plant biological phenomena. The transcript functions were predicted with the help of the GO terms using Blast2GO based on their similarities to transcripts available in the nr database. All assembled transcripts were surveyed at an E-value of 1×10^{-6}. The numbers of genes assigned to the biological process, molecular function, and cellular component categories were 9,336, 2,371, and 1,163, respectively (Fig. 3). Among the biological processes transcripts, the highest number of transcripts was assigned to cellular processes (13,538),

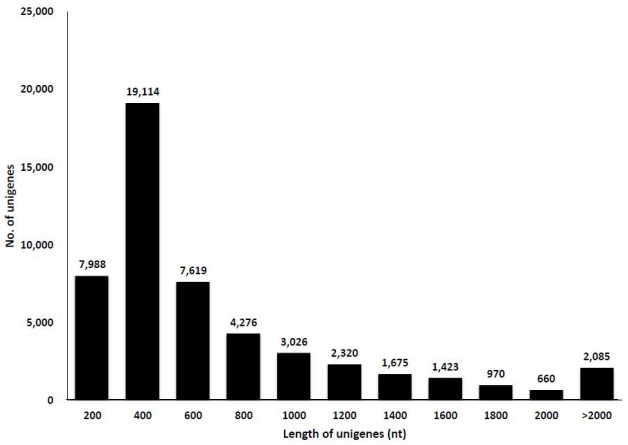


Fig. 1. Distribution of the length of the unigenes obtained by the *de novo* assembly of the transcriptome sequencing in *Cirsium japonicum*.

Table 2. The mapped reads to the transcripts using trimmed RNA sequence data

	Mapping reads to transcripts for DEG analysis				
	Trimmed reads	Mapped reads, n (%)	No. of mapped transcripts	GC (%)	Q30 (%)
Flower	20,728,854	9,751,931 (47.05)	24,663	44.88	96.29
Leaf	30,141,768	16,009,899 (53.12)	34,031	42.9	96.46
Root	16,283,252	8,121,204 (49.87)	19,545	44.88	96.34

followed by metabolic processes (12,322) and single organism processes (11,785). Several other processes, including biological regulation and response to stimulus, also comprised a significant number of transcripts (Fig, 3, biological process). The most prominent molecular functions included binding activity (12,037) and catalytic activity (10,141) (Fig. 3, molecular function). In the cellular component analysis, most transcripts were assigned to cell (12,808), cell part (12,251), organelle (8,806), and membrane (8,272). Macromolecular complex (5,632), membrane part (5,265), and organelle part (4,717) also had a substantial number of transcripts (Fig. 3, cellular component).

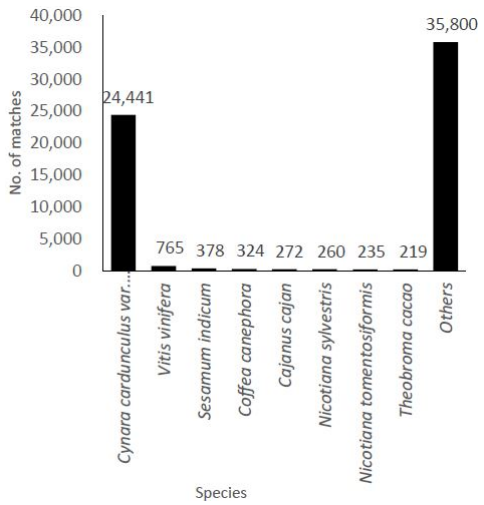


Fig. 2. Distribution of the species hits based on the gene annotation of the transcriptome by the BlastX search in *Cirsium japonicum* var. *spinossimum*.

Functional classification of KEGG pathways

For the KEGG analysis, 17,583 annotated transcripts were mapped to identify the active pathways in *C. japonicum* at a cut-off E-value < 0.00001. Of these transcripts, 9,418 transcripts had significant matches in the database and were assigned to 137 KEGG pathways. Metabolism had the most unigenes; amino acid metabolism had the highest number of annotated unigenes, followed by carbohydrate metabolism, xenobiotic biodegradation and metabolism, lipid metabolism, biosynthesis of antibiotics, etc. (Table 3). These annotations provide insight into the transcriptome by enhancing our understanding of the specific functions and pathways in *C. japonicum*.

Screening and identification of DEGs

To identify the differentially ex pressed genes, the transcriptome data from *C. japonicum* leaf, root and flower tissues were analysed using the EdgeR package in R software [19]. The DEGs were visualized as an MA plot (log ratio vs. abundance plot) of flower vs. leaf, flower vs. root and leaf vs. root (Supplementary Fig. 1). The red dots represent transcripts with positive and negative log2fold change values, indicating the up-regulation and down-regulation of the DEGs in each comparison. Using these criteria, most transcripts were mapped in the leaf tissue (34,031), followed by the flower (24,663) and root (19,545) tissues (Table 2). The expression

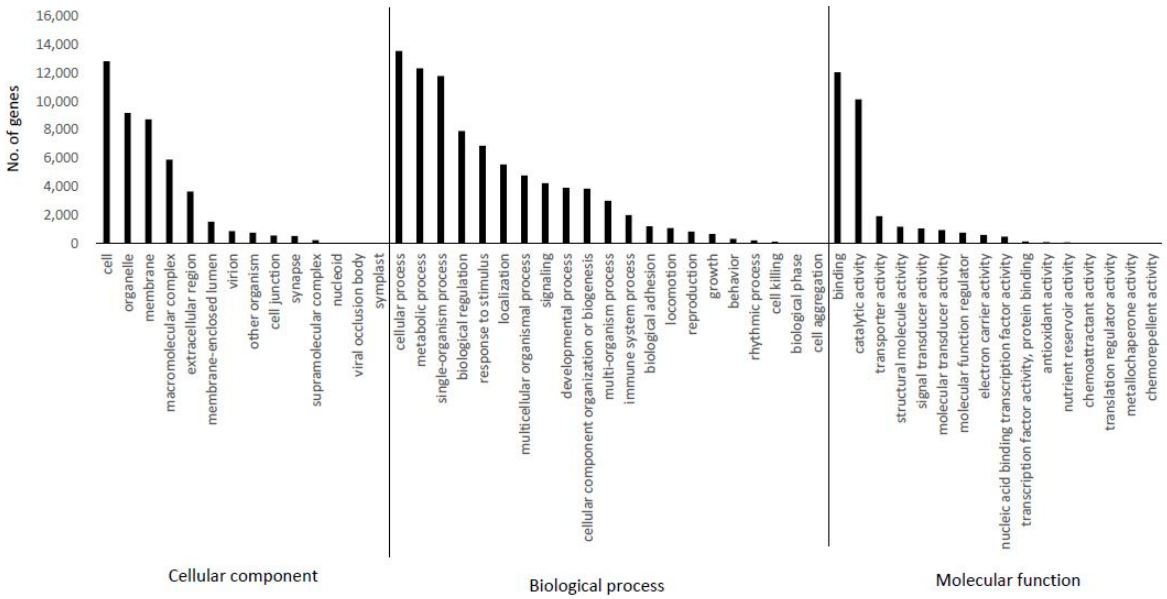


Fig. 3. Gene Ontology (GO) annotations of non-redundant consensus sequences. The best hits were aligned to the GO database, and the most consensus sequences were grouped into three major functional categories, namely, biological process, molecular function, and cellular component.

Table 3. KEGG analysis pathway distribution among the transcriptome of *Cirsium japonicum* var. *spinossimum*

KEGG pathways classification	Sub classification	Count of pathway	Sum of unigenes in pathway
Metabolism	Amino acid metabolism	14	1597
	Carbohydrate metabolism	15	1354
	Xenobiotics biodegradation and metabolism	18	1312
	Lipid metabolism	15	1258
	Biosynthesis of antibiotics	1	957
	Energy metabolism	7	876
	Nucleotide metabolism	2	519
	Metabolism of cofactors and vitamins	12	509
	Biosynthesis of other secondary metabolites	17	316
	Metabolism of other amino acids	9	230
	Glycan biosynthesis and metabolism	11	185
	Metabolism of terpenoids and polyketides	9	62
Environmental information processing	Signal transduction	3	117
Organismal systems	Immune system	2	66
Genetic information processing	Translation	1	44
Human disease	Drug resistance: antimicrobial	1	16
Grand total		137	9,418

KEGG, Kyoto Encyclopedia of Genes and Genomes.

Table 4. Number of expressed read counts from 51,133 transcripts in *Cirsium japonicum*

	Expressed transcripts	Percentage
Flower	37,665	73.66
Leaf	46,991	91.90
Root	32,257	63.08

values corresponding to specific genes were arranged in a matrix form, and two tables of the counts and FPKM were generated to represent the genes (data not shown). The count matrix represented the number of reads mapped to the reference, and the FPKM matrix represented the normalized value using the FPKM-based count value. In total, 51,133 expressed reference sequences were obtained. Overall, the leaf (Cjapleaf) samples had the most expressed transcripts (46,991, 91.90%), followed by the flower (Cjapflower) (37,665, 73.66%) and root (Cjaproot) (32,257, 63.08%), as shown in Table 4. Based on the expression differences shown in the Venn diagram in Fig. 4, GO and KEGG analyses were performed again based on the nr database annotations. The unigenes were assigned to 61 GO categories as follows: Cjapleaf vs. Cjaproot (29,491 up-regulated and 27,898 down-regulated unigenes) (Supplementary Fig. 2); Cjaproot vs. Cjapflower (16,365 up-regulated and 32,596 down-regulated unigenes) (Supplementary Fig. 3) and Cjapflower vs Cjapleaf (19,207 up-regulated and 33,472 down-regulated unigenes) (Supplementary Fig. 4).

In the KEGG analysis, 2,615 unigenes were annotated in

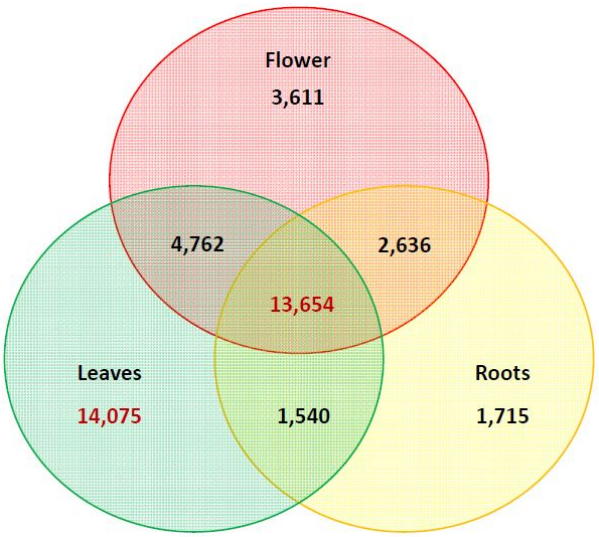


Fig. 4. Venn diagram representing the significantly differentially expressed transcripts among flower, leaf, and root tissues from *Cirsium japonicum*. The number includes both the up- and down-regulated genes. Genes that are differentially expressed in more than one comparison are depicted in the overlapping area.

the Cjapleaf vs. Cjaproot comparison; of these unigenes, 2,115 up-regulated unigenes were assigned to 109 pathways, and 500 down-regulated unigenes were assigned to 113 pathways. In the comparison Cjaproot vs. Cjapflower, 877 unigenes were annotated in the KEGG classification; of these unigenes, 348 up-regulated unigenes were assigned to 103 pathways, and 529 down-regulated unigenes were assigned to 116 pathways. In the comparison Cjapflower vs. Cjapleaf,

934 unigenes were annotated; of these unigenes, 383 up-regulated unigenes were assigned to 97 pathways, and 551 down-regulated unigenes were assigned to 114 pathways. The differences in the expression levels were two-fold or greater, indicating a significant difference between the compared samples. The comparison of the up-regulated and down-regulated transcripts was performed by comparison among the tissues.

Identification of putative unigenes involved in silymarin biosynthesis

Silybin A and B, iso-silybins A and B, silychristin A, isosilychristin, and silydianin are all collectively called silymarins. Silymarins are the major flavolignans produced in *C. japonicum*. Taxifolin and coniferyl alcohol are the two major precursors responsible for the production of silymarin [20] and are produced by the flavonoid and phenylpropanoid pathways, respectively, as shown in Fig. 5 (the highlighted boxes represent the enzymes annotated in our data). The unigenes identified for the silymarin production in *C. japonicum* are listed in Supplementary Table 2. Most enzymes responsible for the production of silymarin were found in all three tissues (flower, leaf, and root) of *C. japonicum*. Subsequently, silymarin formation occurs via the oxidative

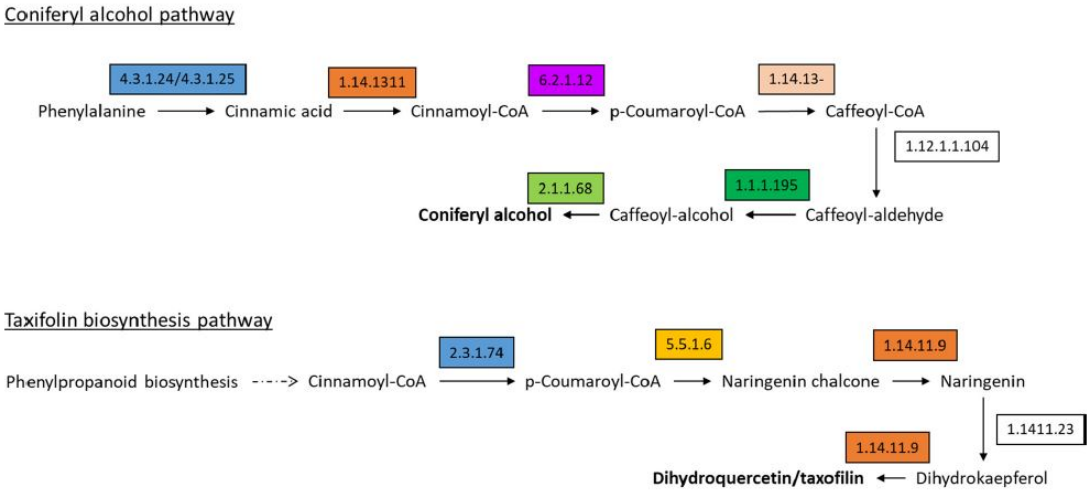


Fig. 5. Silymarin is produced via the oxidative coupling of two major precursors, namely, taxifolin and coniferyl alcohol. The figure represents the formation of coniferyl alcohol and taxifolin via the phenylpropanoid and flavonoid pathways, respectively. The number in the boxes represents the enzyme as code per the Kyoto Encyclopedia of Genes and Genomes (KEGG) classification, and the details are provided in Supplementary Table 2. The coloured boxes represent the enzymes that were found in our transcriptome data.

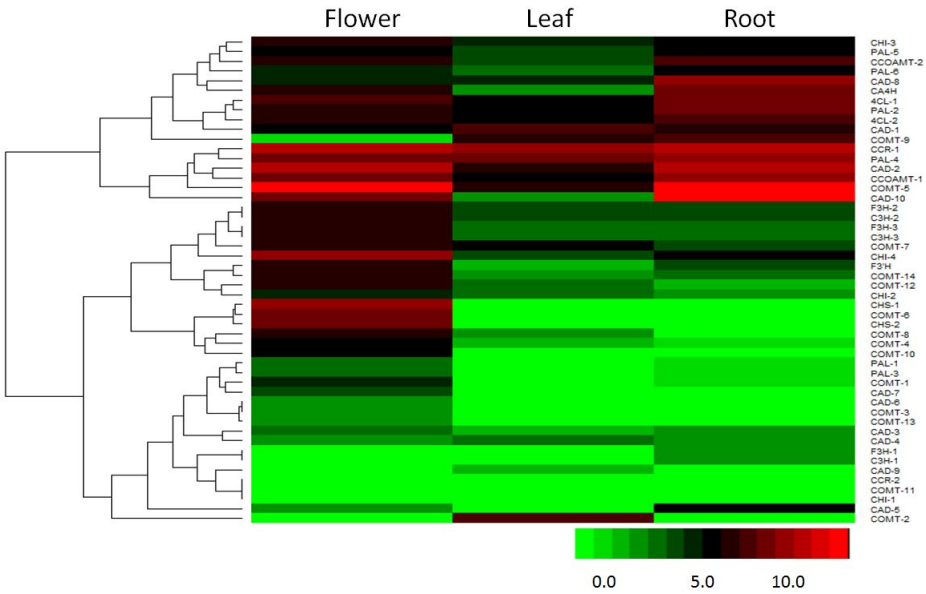


Fig. 6. Heat map of the differently expressed genes involved in silymarin biosynthesis. Red and green colours indicate up-and down-regulated gene expression, respectively.

radicalization of taxifolin and coniferyl alcohol, followed by a combinational radical coupling [21]. Genes corresponding to the silymarin biosynthetic pathway (flavonoid and phenylpropanoid pathway-related genes) were selected from the *C. japonicum* transcripts, and the expression patterns were compared among the flower, leaf and root tissues (Supplementary Table 2). The organ-specific gene expression patterns in *C. japonicum* were confirmed by the heat map technique (Fig. 6). The maximum number of genes was found to be downregulated in leaf tissues.

Discussion

Cirsium japonicum var. *spinossimum* is a wild perennial herb that grows in the mountains and fields of Jeju Island, South Korea, and is also found in Japan. This species is often compared to a similar herb called *C. japonicum* DC [7]. In Korea, this herb is known for its medicinal properties, including its antihaemorrhagic, antihypertensive, antihepatitic, and uretic medicinal properties [2]. Since the advent of next generation technology, this research tool has become important for the generation of vast information regarding genomic and transcriptomic data. These data eventually help infer various basic biological, molecular and cellular processes, especially in non-model organisms and non-sequenced genomes [22-24]. Thus far, no studies have performed a transcriptome analysis of this herb. We characterized the transcriptome of *C. japonicum* var. *spinossimum* using an Illumina Hiseq2500. This study represents the first report of the whole transcriptome and differential gene expression among the flower, leaf, and root tissues in *C. japonicum* var. *spinossimum* and provides insight into fundamental molecular data.

We obtained 66,078,874 reads after trimming and performing error corrections, and these reads assembled into 51,133 unigenes with an average length of 648 bp (Table 1). The number of mapped transcripts in each tissue was 24,663, 34,031, and 19,545 in the flower, leaf, and root tissues, respectively (Table 2). We obtained unigenes that ranged from 200 to 15,402 bp (Fig. 1), which is comparable to similar RNA-Seq reports [25-27]. This number also represents the high quality of our generated data. Although prior genome information for *C. japonicum* is lacking, the BLASTX program successfully annotated 33,525 (65.59%) transcripts. The highest similarity obtained with *Cynara cardunculus* var. *scolymus*, is a medicinal plant belonging to same family (Asteraceae) [28, 29]. Thus, the *C. japonicum* transcriptome can prove useful for functional gene studies or molecular biology studies investigating *C. japonicum*. Of the 17,583 mapped transcripts, the KEGG classification annotated 9,418 (53.56%) transcripts into several pathways (Table 3), but no functional annotation was found for the remaining 46.44% of the assembled unigenes. This might be due to one of the following reasons: either the unigenes matched proteins with unknown functions or they did not have any homologous sequence matches in the database. Thus, these unknown unigenes may be highly important for further research since they may be considered novel transcripts.

The DEG patterns were investigated to further profile the global gene expression differences among the leaf, flower and root in *C. japonicum*. The most abundant genes were expressed in the leaves, followed by the flowers and roots (Fig. 4). We obtained 26,289 genes that were expressed in all three tissues, including 14,075 genes that were specifically expressed in the leaf tissues, 3,611 genes that were specifically expressed in the flower tissues and 1,715 genes that were specifically expressed in the root tissues. The molecular function category showed that more than 40% of the unigenes were expressed in all three tissues (Supplementary Figs. 2–4). The major categories in which differential gene expression was observed were biological process and cellular components.

Silymarin is a main component in *C. japonicum* that contributes to its medicinal value. Therefore, we aimed to expand our knowledge by identifying putative unigenes that contribute to the product. Silymarin is reportedly produced via the oxidative coupling of two major precursors, namely, taxifolin and coniferyl alcohol [21]. Most major enzymes responsible to produce both taxifolin and coniferyl alcohol were present in our transcriptome data (Supplementary Table 2), which is consistent with other medicinal plants observed in the same family [6, 20]. In the *C. japonicum* transcriptome assembly, we identified silymarin biosynthesis-related unigenes (Fig. 5), and their differential expression was assessed (Fig. 6). Compared to their expression in the flower and root tissues, these genes were highly expressed in the leaf tissues. Silymarin biosynthesis has been reported to be lacking in the flower of *S. marianum* [6].

In addition to these unigenes, unigenes encoding enzymes crucial for the biosynthesis of antibiotics and the biosynthesis of secondary metabolites were identified. Medicinal plants represent a rich source of secondary metabolites [30, 31], and their data could be useful for molecular biology research and the mass production of significant metabolites.

This report represents the first *de novo* assembly of the transcriptome of *C. japonicum* var. *spinossimum* obtained from the Korean Peninsula. This study provides resources for comparative transcriptomics, especially in the field of the biochemical and molecular biosynthesis pathways of silymarin. In total, 51,133 unigenes were obtained with a mean length of 648.36 bp. In total, 33,525 annotated

sequences were assigned to 64 GO functional groups, and 10,535 unigenes were assigned to 107 KEGG pathways. The DEG analysis revealed that the highest numbers of genes were expressed in the leaf tissues. Candidate genes that might be involved in silymarin biosynthesis may help further functional genomic and transcriptomic analyses in *C. japonicum*. Putative unigenes that facilitate the production of taxifolin and coniferyl alcohol, which are the major precursors in silymarin biosynthesis, were identified. This work will contribute to the comprehensive knowledge of new traditional medicinal plants for growers and consumers and provides additional characteristics and information regarding the pharmaceutical benefits associated with silymarin biosynthesis.

ORCID: Neha Samir Roy: https://orcid.org/0000-0002-4529-4861; Jung-A Kim: https://orcid.org/0000-0002-0136-5410; Ah-Young Choi: https://orcid.org/0000-0002-2013-6533; Yong-Wook Ban: https://orcid.org/0000-0002-1912-0374; Nam-Il Park: https://orcid.org/0000-0001-6725-0758; Kyong-Cheul Park: https://orcid.org/0000-0002-3737-815X; Hee-sun Yang: https://orcid.org/0000-0002-3289-6345; Ik-Young Choi: https://orcid.org/0000-0003-4168-0471; Soo nok Kim: https://orcid.org/0000-0003-1654-3643

Authors' contribution

Conceptualization: SK, IYC
Data curation: NSR, AYC, YWB, KCP
Formal analysis: NSR, YWB
Funding acquisition: SK, HSY
Methodology: IYC, NSR, NIP, JAK
Writing – original draft: NSR, YWB
Writing – review & editing: NSR, YWB, SK, IYC

Conflicts of Interest

No potential conflicts of interest relevant to this article was reported.

Acknowledgments

This work was supported by the grants from 2017 Research Grant, Kangwon National University and the National Institute of Biological resources (NIBR), funded by the Minister of Environment (MOE) of the Republic of Korea (NIBR201626103, NIBR201830101).

Supplementary material

Supplementary data including two tables and four figures can be found with this article online at https://doi.org/10.5808/GI.2018.16.4.e34.

References

1. Ge H, Turhong M, Abudkrem M, Tang Y. Fingerprint analysis of *Cirsium japonicum* DC. using high performance liquid chromatography. J Pharm Anal 2013;3:278-284
2. Liu S, Luo X, Li D, Zhang J, Qiu D, Liu W, *et al*. Tumor inhibition and improved immunity in mice treated with flavone from *Cirsium japonicum* DC. *Int Immunopharmacol* 2006;6:1387-1393.
3. Federico A, Dallio M, Loguercio C. Silymarin/silybin and chronic liver disease: a marriage of many years. *Molecules* 2017;22:E191.
4. Loguercio C, Festi D. Silybin and the liver: from basic research to clinical practice. *World J Gastroenterol* 2011;17:2288-2301.
5. Kim NC, Graf TN, Sparacino CM, Wani MC, Wall ME. Complete isolation and characterization of silybins and isosilybins from milk thistle (*Silybum marianum*). *Org Biomol Chem* 2003;1:1684-1689.
6. Ma Q, Wang LH, Jiang JG. Hepatoprotective effect of flavonoids from *Cirsium japonicum* DC on hepatotoxicity in comparison with silymarin. *Food Funct* 2016;7:2179-2184.
7. WoRMS. *Cirsium japonicum* var. *spinosissimum* Kitam. WoRMS, 2008. Accessed 2018 Nov 1. Available from: http://www.marinespecies.org/aphia.php?p=taxdetails&id=1157330.
8. Boerjan W, Ralph J, Baucher M. Lignin biosynthesis. *Annu Rev Plant Biol* 2003;54:519-546.
9. Morazzoni P, Bombardelli E. *Silybum marianum* (*Carduus marianus*). *Fitoterapia* 1995;66:3-42.
10. Petrussa E, Braidot E, Zancani M, Peresson C, Bertolini A, Patui S, *et al*. Plant flavonoids: biosynthesis, transport and involvement in stress responses. *Int J Mol Sci* 2013;14:14950-14973.
11. Nakasugi K, Crowhurst RN, Bally J, Wood CC, Hellens RP, Waterhouse PM. *De novo* transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*. *PLoS One* 2013;8:e59534.
12. Lehnert EM, Walbot V. Sequencing and *de novo* assembly of a Dahlia hybrid cultivar transcriptome. *Front Plant Sci* 2014;5:340.
13. Rastogi S, Meena S, Bhattacharya A, Ghosh S, Shukla RK, Sangwan NS, *et al*. *De novo* sequencing and comparative analysis of holy and sweet basil transcriptomes. *BMC Genomics* 2014;15:588.
14. Wilhelm BT, Marguerat S, Goodhead I, Bähler J. Defining transcribed regions using RNA-seq. *Nat Protoc* 2010;5:255-266.
15. Martínez-García PJ, Crepeau MW, Puiu D, Gonzalez-Ibeas D, Whalen J, Stevens KA, *et al*. The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. *Plant J* 2016;87:507-532.
16. Liu MH, Yang BR, Cheung WF, Yang KY, Zhou HF, Kwok JS, *et al*. Transcriptome analysis of leaves, roots and flowers of *Panax notoginseng* identifies genes involved in ginsenoside and alkaloid biosynthesis. *BMC Genomics* 2015;16:265.
17. Lulin H, Xiao Y, Pei S, Wen T, Shangqin H. The first Illumina-based *de novo* transcriptome sequencing and analysis

of safflower flowers. *PLoS One* 2012;7:e38653.
18. Upadhyay AK, Chacko AR, Gandhimathi A, Ghosh P, Harini K, Joseph AP, *et al*. Genome sequencing of herb Tulsi (*Ocimum tenuiflorum*) unravels key genes behind its strong medicinal properties. *BMC Plant Biol* 2015;15:212.
19. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139-140.
20. AbouZid SF, Ahmed HS, Moawad AS, Owis AI, Chen SN, Nachtergael A, *et al*. Chemotaxonomic and biosynthetic relationships between flavonolignans produced by *Silybum marianum* populations. *Fitoterapia* 2017;119:175-184.
21. Lv Y, Gao S, Xu S, Du G, Zhou J, Chen J. Spatial organization of silybin biosynthesis in milk thistle [*Silybum marianum* (L.) Gaertn]. *Plant J* 2017;92:995-1004.
22. Mudalkar S, Golla R, Ghatty S, Reddy AR. *De novo* transcriptome analysis of an imminent biofuel crop, *Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers. *Plant Mol Biol* 2014;84:159-171.
23. Yang Y, Xu M, Luo Q, Wang J, Li H. *De novo* transcriptome analysis of *Liriodendron* chinense petals and leaves by Illumina sequencing. *Gene* 2014;534:155-162.
24. Zhu L, Zhang Y, Guo W, Xu XJ, Wang Q. *De novo* assembly and characterization of *Sophora japonica* transcriptome using RNA-seq. *Biomed Res Int* 2014;2014:750961.
25. Zhang X, Allan AC, Li C, Wang Y, Yao Q. *De novo* assembly and characterization of the transcriptome of the Chinese medicinal herb, *Gentiana rigescens*. *Int J Mol Sci* 2015;16:11550-11573.
26. Zhang X, Liu T, Duan M, Song J, Li X. *De novo* transcriptome analysis of *Sinapis alba* in revealing the glucosinolate and phytochelatin pathways. *Front Plant Sci* 2016;7:259.
27. Sun H, Li F, Xu Z, Sun M, Cong H, Qiao F, *et al*. *De novo* leaf and root transcriptome analysis to identify putative genes involved in triterpenoid saponins biosynthesis in *Hedera helix* L. *PLoS One* 2017;12:e0182243.
28. Acquadro A, Barchi L, Portis E, Mangino G, Valentino D, Mauromicale G, *et al*. Genome reconstruction in *Cynara cardunculus* taxa gains access to chromosome-scale DNA variation. *Sci Rep* 2017;7:5617.
29. Scaglione D, Lanteri S, Acquadro A, Lai Z, Knapp SJ, Rieseberg L, *et al*. Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. *Plant Biotechnol J* 2012;10:956-969.
30. Hartmann T. Plant-derived secondary metabolites as defensive chemicals in herbivorous insects: a case study in chemical ecology. *Planta* 2004;219:1-4.
31. Goyal S, Lambert C, Cluzet S, Merillon JM, Ramawat KG. Secondary metabolites and plant defence. In: *Plant Defence*: *Biological Control* (Merillon JM, Ramawat KG, eds.). Dordrecht: Springer, 2012. pp. 109-138.

SUPPLEMENTARY INFORMATION

RNA-Seq *De Novo* Assembly and Differential Transcriptome Analysis of Korean Medicinal Herb *Cirsium japonicum* var. *spinossimum*

Neha Samir Roy[1,2], Jung-A Kim[3], Ah-Young Choi[4], Yong-Wook Ban[5], Nam-Il Park[6], Kyong-Cheul Park[1], Hee-sun Yang[3], Ik-Young Choi[1,2]*, Soonok Kim[3]**

[1]Department of Agriculture and Life Industry, Kangwon National University, Chuncheon 24341, Korea, [2]Agriculture and Life Sciences Research Institute, Kangwon National University, Chuncheon 24341, Korea, [3]Biological Resources Assessment Division, National Institute of Biological Resources, Incheon 22689, Korea, [4]DNACare Genomics Institute, Seongnam 13558, Korea, [5]Department of Forest Environmental System, Kangwon National University, Chuncheon 24341, Korea, [6]Department of Plant Science, Gangneung Wonju National University, Gangneung 25457, Korea

https://doi.org/10.5808/GI.2018.16.4.e34

Supplementary Table 1. Summary of data output from Illumina HiSeq2500 for *Cirsium japonicum* var. *spinossimum*

	Sample: *Cjap* reference	Reads	Read bases	Average	Largest length
Reference sequencing data summary	Raw data	74,566,546	18,716,203,046	251	251
	Trimmed data	67,153,874	12,346,643,676	183.86	251
	Error correction data	66,078,302	11,841,793,951	179.21	251
Reference assembly summary	Trinity (RNA-seq *de novo* assembly) Transdecoder	51,133	37,533,942	648.36	15,402

Supplementary table 2. Log2Fold change values obtained for the enzymes involved in Coniferyl alcohol and Taxifolin pathways for Heat map Enzymes for Coniferyl alcohol pathway

1. phenylalanine ammonia-lyase [EC:4.3.1.24]_PAL

				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	pleaf/Cjaproot
Cjap_01115	184	CAL91171.1	phenylalanine ammonia-lyase 2	2.9978367	-5.141469	-2.577308	2.5641612
Cjap_02759	484	CAL91169.1	phenylalanine ammonia-lyase 3	5.6998326	-0.986101	1.2824969	2.2685978
Cjap_03456	169	CAL91171.1	phenylalanine ammonia-lyase 2	2.601145	-5.622095	-2.057934	3.5641612
Cjap_08760	707	AAL55242.1	AF299330_1 phenylalanine ammonia-lyase	6.79	0.10	0.65	0.55
Cjap_15637	129	CAL91170.1	phenylalanine ammonia-lyase 3	5.07	-1.87	-0.20	1.67
Cjap_21084	140	CAL91170.1	phenylalanine ammonia-lyase 3	4.1946695	-1.658621	0.8680656	2.5266865

2. trans-cinnamate 4-monooxygenase [EC:1.14.13.11] (=cinnamate 4-hydroxylase)_CA4H

				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	pleaf/Cjaproot
Cjap_11124	99	CAM84301.1	putative cinnamate 4-hydroxylase	5.6849458	-3.542055	1.3255965	4.867652

3. p-coumarate 3-hydroxylase [EC: 1.14.13.-] (p-coumaric acid --> caffeic acid)_C3H

Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	pleaf/Cjaproot
Cjap_00996	81	AAM48133.1	AF509338_1 putative flavanone 3-hydroxylase	-	-	-	-
Cjap_31912	190	BAJ17667.1	flavanone 3-hydroxylase	5.2658979	-1.580754	-1.960637	-0.379882
Cjap_53742	252	AKN79608.1	flavanone 3-hydroxylase	5.4928934	-2.61451	-2.526083	0.0884277

4. caffeic acid 3-O-methyltransferase [EC:2.1.1.68] _ caffeic acid --> ferulic acid _COMT

Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	pleaf/Cjaproot
Cjap_03410	131	KVI04224.1	Caffeate O-methyltransferase (COMT) family	3.8165984	-5.470092	-4.490893	0.9791987
Cjap_03611	376	KVG04787.1	Caffeate O-methyltransferase (COMT) family	-2.463287	8.3071634	-0.736006	-9.043169
Cjap_20580	87	KVH94228.1	Caffeate O-methyltransferase (COMT) family	1.8221154	-	-	-
Cjap_34331	269	KVI04221.1	Caffeate O-methyltransferase (COMT) family	4.7048188	-4.079777	-4.422506	-0.342729
Cjap_35980	394	KVH88410.1	Caffeate O-methyltransferase (COMT) family	9.156582	-3.541274	0.1843988	3.7256733
Cjap_36967	355	KVI04793.1	Caffeate O-methyltransferase (COMT) family	7.050898	-	-	!
Cjap_40202	403	KVI08129.1	Caffeate O-methyltransferase (COMT) family	5.46	-0.96	-1.84	-0.88
Cjap_43214	255	KVH53040.1	Caffeate O-methyltransferase (COMT) family	5.5437239	-3.76583	-	-
Cjap_57357	365	KVI04575.1	Caffeate O-methyltransferase (COMT) family	0.3867889	4.9951101	5.8053675	0.8102574
Cjap_00320	135	XP_011468886.1	PREDICTED: LOW QUALITY PROTEIN: caffeic acid 3-O-	4.7890307	-	-	-
Cjap_10005	73	XP_007015117.1	Caffeic acid 3-O-methyltransferase 1 isoform 1	-	-	-	-
Cjap_22372	197	XP_006352624.1	PREDICTED: caffeic acid 3-O-methyltransferase 1-like	5.214549	-2.623781	-4.307548	-1.683766
Cjap_42425	82	XP_007015118.1	Caffeic acid 3-O-methyltransferase 1 isoform 2	1.5449368	-	-	-
Cjap_52886	184	XP_007015118.1	Caffeic acid 3-O-methyltransferase 1 isoform 2	5.3981204	-3.734398	-2.608358	1.1260401

5. 4-coumarate--CoA ligase [EC:6.2.1.12] _ caffeic acid --> caffeoyl-coA_ 4CL

				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	pleaf/Cjaproot

Cjap_07087	539	AFL93685.1	4-coumarate: coenzyme A ligase	6.05	-0.98	0.62	1.60
Cjap_27264	531	XP_009603383.1	PREDICTED: 4-coumarate--CoA ligase-like 7	5.6059055	-0.496624	0.4001129	0.8967365

6. cinnamoyl-CoA reductase [EC:1.2.1.44] _ caffeoyl-coA --> caffeyl-aldehyde_ CCR

Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjappleaf/Cjaproot	
Cjap_19620	324	XP_002275195.1	PREDICTED: cinnamoyl-CoA reductase 1	8.0614507	-0.606384	-0.240069	0.3663147
Cjap_32839	344	XP_002300619.2	cinnamoyl-CoA reductase family protein				

7. cinnamyl-alcohol dehydrogenase [EC:1.1.1.195]_ caffeyl-aldehyde --> caffeyl alcohol_ CAD

Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjappleaf/Cjaproot	
Cjap_03029	360	XP_012073738.1	PREDICTED: probable cinnamyl alcohol dehydrogenase	4.875403	1.221775	0.287078	-0.9347
Cjap_04466	386	BAN91671.1	cinnamyl alcohol dehydrogenase 2	8.136022	-2.45746	-0.34242	2.115048
Cjap_04954	165	BAN91672.1	cinnamyl alcohol dehydrogenase 3	2.429251	-1.16772	-0.18852	0.979199
Cjap_20491	206	BAN91672.1	cinnamyl alcohol dehydrogenase 3	1.578558	0.787296	0.00096	-0.78634
Cjap_26535	88	BAN91672.1	cinnamyl alcohol dehydrogenase 3	1.941329	-	2.98646	-
Cjap_27993	77	BAN91670.1	cinnamyl alcohol dehydrogenase 1	1.828347	-	-	-
Cjap_33336	89	BAN91670.1	cinnamyl alcohol dehydrogenase 1	3.142962	-	-	-
Cjap_41375	102	BAN91672.1	cinnamyl alcohol dehydrogenase 3	4.189993	-0.26775	3.10102	3.368765
Cjap_50558	91	BAN91672.1	cinnamyl alcohol dehydrogenase 3	-	-	-	-
Cjap_51684	368	BAN91670.1	cinnamyl alcohol dehydrogenase 1	6.999154	-5.01742	2.733929	7.751348

8. caffeoyl-CoA O-methyltransferase [EC: 2.1.1.104]_

Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjappleaf/Cjaproot	
Cjap_26162	247	BAG71891.1	caffeoyl CoA O-methyltransferase	6.511625	-1.566256	1.2530564	2.819312
Cjap_48779	172	BAG71892.1	caffeoyl CoA O-methyltransferase	5.8153237	-2.460632	0.2436025	2.7042341

Enzymes for Taxifolin pathway

1. phenylalanine ammonia-lyase [EC:4.3.1.24]

				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjappleaf/Cjaproot	
Cjap_01115	184	CAL91171.1	phenylalanine ammonia-lyase 2	2.9978367	-5.141469	-2.577308	2.5641612
Cjap_02759	484	CAL91169.1	phenylalanine ammonia-lyase 3	5.6998326	-0.986101	1.2824969	2.2685978
Cjap_03456	169	CAL91171.1	phenylalanine ammonia-lyase 2	2.601145	-5.622095	-2.057934	3.5641612
Cjap_08760	707	AAL55242.1	AF299330_1 phenylalanine ammonia-lyase	6.79	0.10	0.65	0.55
Cjap_15637	129	CAL91170.1	phenylalanine ammonia-lyase 3	5.07	-1.87	-0.20	1.67
Cjap_21084	140	CAL91170.1	phenylalanine ammonia-lyase 3	4.1946695	-1.658621	0.8680656	2.5266865

2. trans-cinnamate 4-monooxygenase [EC:1.14.13.11] (=cinnamate 4-hydroxylase)

				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjappleaf/Cjaproot	

Cjap_11124	99	CAM84301.1	putative cinnamate 4-hydroxylase	5.6849458	-3.542055	1.3255965	4.867652

3. 4-coumarate--CoA ligase [EC:6.2.1.12]

				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	leaf/Cjaproot
Cjap_07087	539	AFL93685.1	4-coumarate: coenzyme A ligase	6.05	-0.98	0.62	1.60
Cjap_27264	531	XP_009603383.1	PREDICTED: 4-coumarate--CoA ligase-like 7	5.6059055	-0.496624	0.4001129	0.8967365

4. chalcone synthase [EC:2.3.1.74]

				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	leaf/Cjaproot
Cjap_00191	405	ABC71308.1	chalcone synthase	7.18	-8.89	-	-
Cjap_17518	401	KVI02820.1	Chalcone/stilbene synthase, active site-containing prote	6.6437545	-8.914877	-	-

5. chalcone isomerase [EC:5.5.1.6]

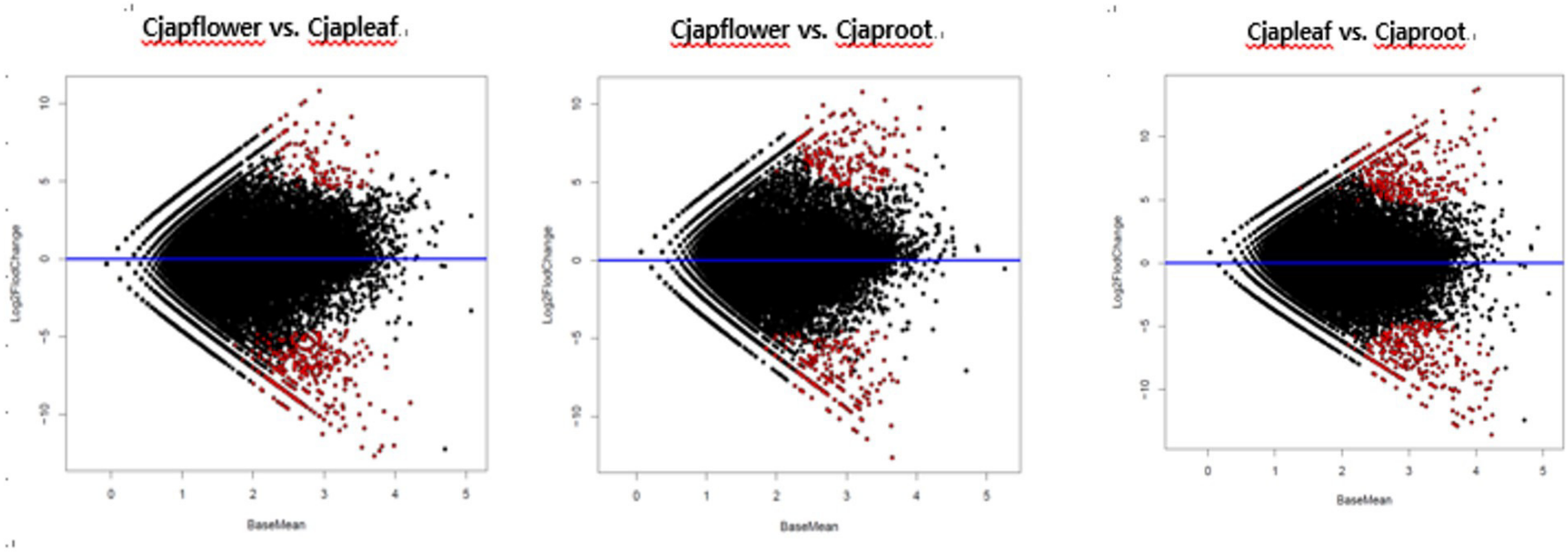
				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	leaf/Cjaproot
Cjap_01214	54	KVI06946.1	Chalcone isomerase	-	-	-	-
Cjap_20743	294	KVI03948.1	Chalcone isomerase, partial	4.08	-1.07	-2.40	-1.33
Cjap_21018	410	KVG40999.1	Chalcone isomerase	5.442806	-1.26206	-0.50458	0.757478
Cjap_22891	243	Q8LKP9.1	CFI_SAUME RecName: Full=Chalcone--flavonone isome	7.287225	-3.65812	-2.34313	1.314989

6. flavanone 3-hydroxylase [EC:1.14.11.9]

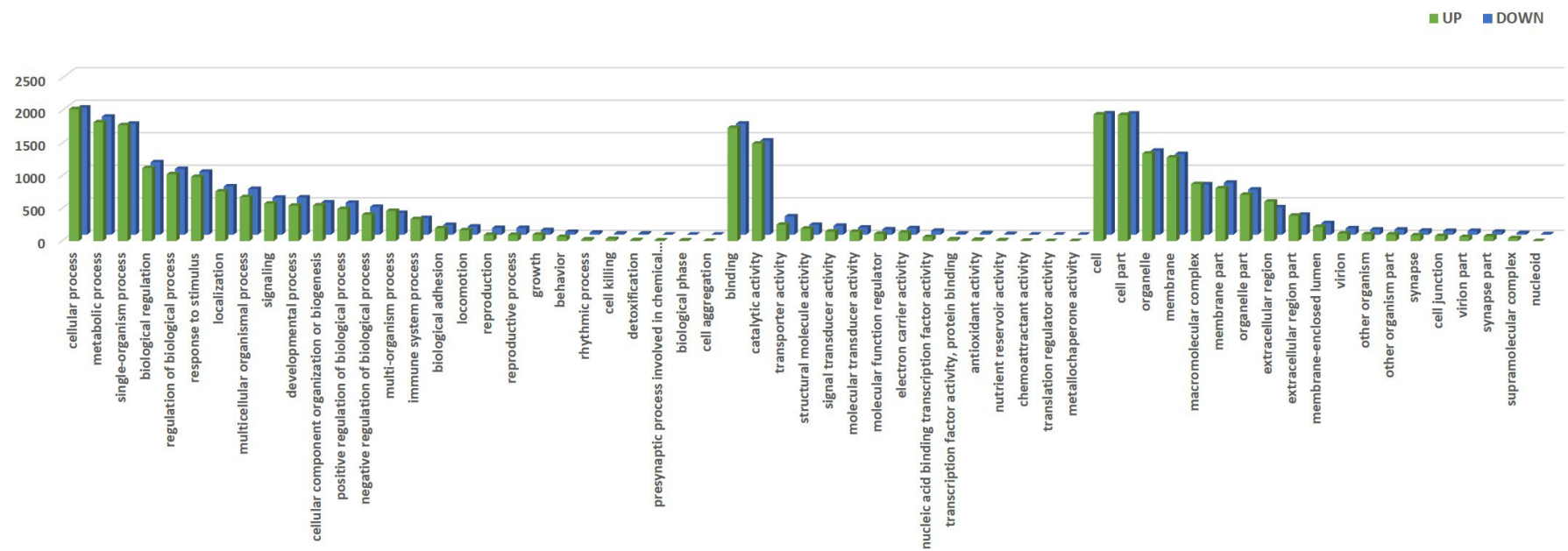
				Log2FoldChange			
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	leaf/Cjaproot
Cjap_00996	81	AAM48133.1	AF509338_1 putative flavanone 3-hydroxylase	-	-	-	-
Cjap_31912	190	BAJ17667.1	flavanone 3-hydroxylase	5.2658979	-1.580754	-1.960637	-0.379882
Cjap_53742	252	AKN79608.1	flavanone 3-hydroxylase	5.4928934	-2.61451	-2.526083	0.0884277

7. flavonoid 3',5'-hydroxylase [EC:1.14.13.88]

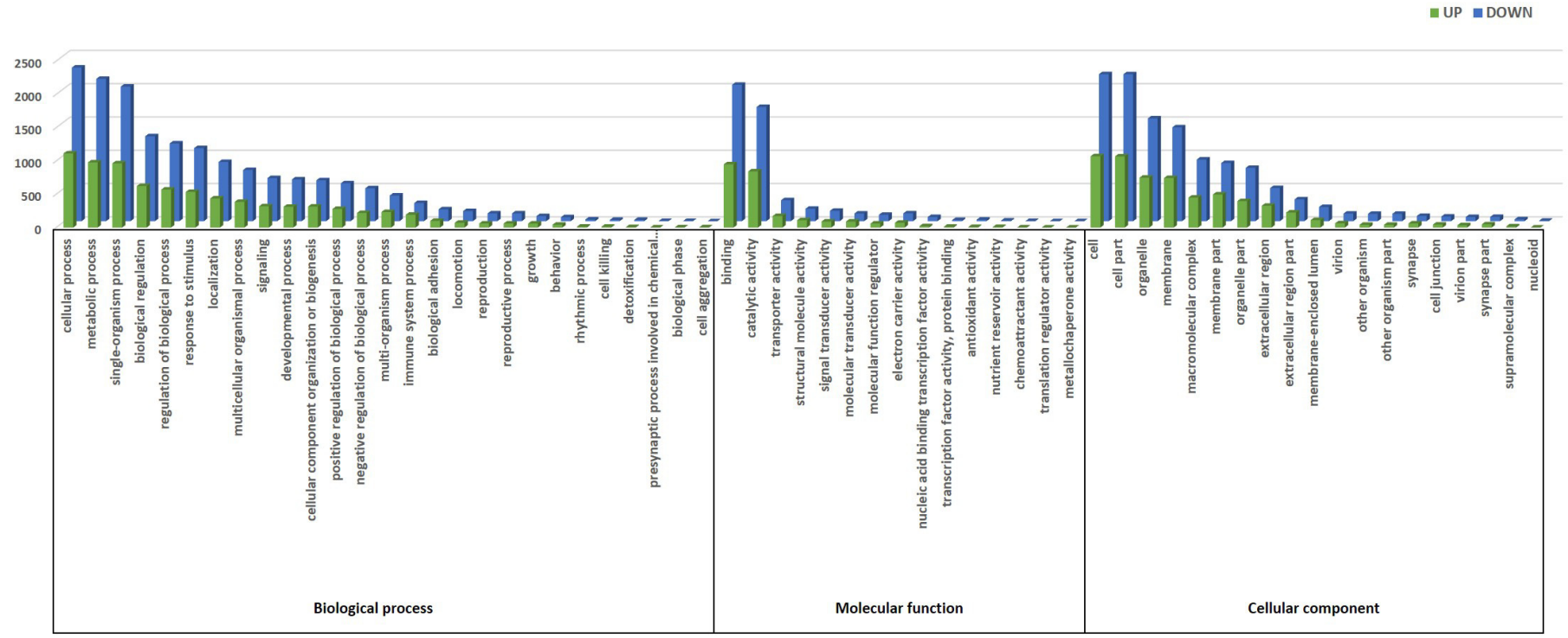
Name	Length	AC	Description	Cjapflower	flower/Cja	flower/Cjap	leaf/Cjaproot
Cjap_08727	517	ADM26615.1	flavonoid 3'-hydroxylase	5.395738	-3.90333	-2.00902	1.89431



Supplementary Fig. 1. MA plot (log ratio vs. abundance plot) for flower vs. leaf, flower vs. root and leaf vs. root in *Cirsium japonicum*.

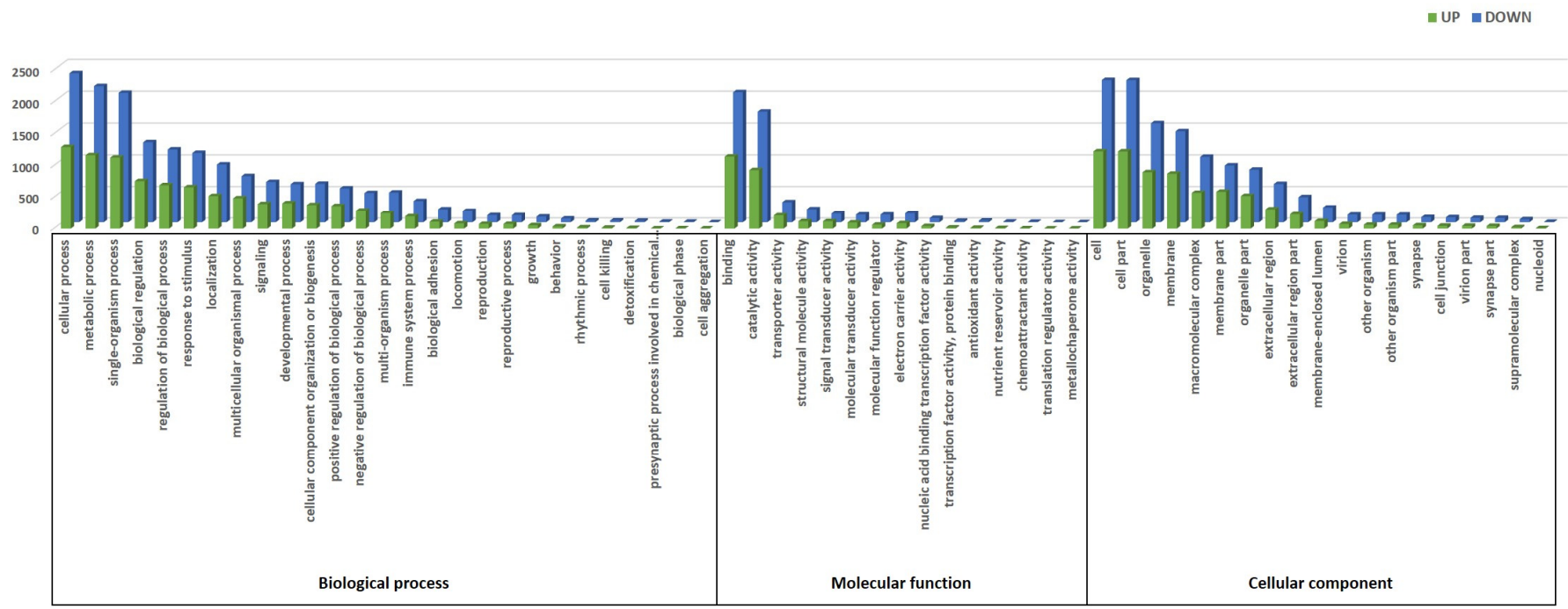


Supplementary Fig. 2. Comparison between up-regulated and down-regulated genes based on functional categories in *Cirsium japonicum* leaf and root tissues.



Cjapflower vs Cjaproot

Supplementary Fig. 3. 4-Comparison between up-regulated and down-regulated genes based on functional categories in *in Cirsium japonicum* flower and root tissues.



Cjapflower vs Cjapleaf

Supplementary Fig. 4. Comparison between up-regulated and down-regulated genes based on functional categories in in Cirsium japonicum flower and leaf. tissues